Topic Shift Detection for Mixed Initiative Response

Rachna Konigari **Saurabh Chand** **Vijay Vardhan Alluri** **Manish Shrivastava**
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad
Gachibowli, Hyderabad, Telangana-500032
{konigari.rachna@research., saurabh.ramola@research.
vijayvardhan.a@research., m.shrivastava@}iiit.ac.in

Abstract

Topic diversion occurs frequently with engaging open-domain dialogue systems like virtual assistants. The balance between staying on topic and rectifying the topic drift is important for a good collaborative system. In this paper, we present a model which uses a fine-tuned XLNet-base to classify the utterances pertaining to the major topic of conversation and those which are not, with a precision of *84%*. We propose a preliminary study, classifying utterances into major, minor and off-topics, which further extends into a system initiative for diversion rectification. A case study was conducted where a system initiative is emulated as a response to the user going off-topic, mimicking a common occurrence of mixed initiative present in natural human-human conversation. This task of classifying utterances into those which belong to the major theme or not, would also help us in identification of relevant sentences for tasks like dialogue summarization and information extraction from conversations.

1 Introduction

Conversational systems have become a part and parcel of our everyday life and virtual assistants like Amazon's Alexa[1], Google Home[2] or Apple's Siri [3] are soon becoming conventional household items (Terzopoulos and Satratzemi, 2020). Most of the conversational systems were built with the primary goal of accessing information, completing tasks, or executing transactions. However, recent conversational agents are transitioning towards a novel hybrid of both task-oriented and a non-task-oriented systems (Akasaki and Kaji, 2017) from the earlier models that resembled factual information systems (Leuski et al., 2006). But with this transition, they are failing to engage in complex information seeking tasks and conversations where multiple turns tend to get involved (Trippas et al., 2020). These new-age open-domain dialogue systems also suffer from a different kind of user behaviour called "anomalous state of knowledge" (Belkin and Vickery, 1985) where the user has vague information requirements and is often unable to articulate it with enough precision. This leads to the user deviating from their original path and traversing into a sub-topic without their knowledge (Larsson, 2017). Thus, we need a context-dependent user guidance without presupposing a strict hierarchy of plans and task goals of the user. Such a guidance, without topic information provided beforehand, is a difficult task to achieve in an open-domain system.

In this work, we observe how a human-human open-domain conversation with an initial topic to begin with, handles topic drift and its rectification in a conversation. We work on the Switchboard dataset (Godfrey et al., 1992) and annotate 74 conversations with 'major', 'minor' and 'off-topic' tags (Section 4). A key result of our finding was that most of the topic shift detection models [(Takanobu et al., 2018), (Wang and Goutte, 2018), (Stewart et al., 2006)] have previously defined topic set to assign to utterances. But as we see in Switchboard dataset, modeling such a pre-defined set is not a property of an open-domain non-task-oriented conversational system. We create a novel model which can, with a precision of 84%, predict the utterances that belong to the major topic and those which are deviating from the same, *without a predetermined topic set*. This is a major contribution as it can help in informational retrieval in conversational systems (Bartl and Spanakis, 2017), dialogue summarization (Gurevych and Strube, 2004) and in the case study that we explored viz. introducing a system initiative in a conversation.

[1] https://developer.amazon.com/en-US/alexa
[2] https://assistant.google.com/
[3] https://www.apple.com/siri/

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 161–166
July 29–31, 2021.

2 Task Definition

Mixed Initiative (MI) is an important aspect for effectively solving multi-agent collaboration problems and is generally referred to as a flexible interaction strategy where each agent can contribute to a task that it is best at (Horvitz, 1999). Here, we'll look into an example of topic shift in a conversation, which sheds light on this issue in a conversation that is common in our day-to-day lives.

MT	A:	Hello, what are your hobbies?
	B:	My hobbies, umm, I used to dance a lot in high school, what are yours?
	A:	I used to paint, but these days I am just occupied with whatever my kids are occupied with at that moment.
OT	B:	Ooh that's nice, how many kids do you have?
	A:	I have two kids, one boy aged 6 and a daughter aged 3 What about you?
	B:	Yes, two twin girls aged 4.
	A:	Aww that's such a lovely age.
	B:	Ya it is, but they can also be a little handful at times.
MI	A:	Anyways, let's go back to the topic at hand, tell me more about your hobbies?

The above example shows how the topic transitioned between the two users, from hobbies which was their major topic given by a prompt, to talking about their kids. We see from the marked area that they transitioned from the major topic (MT) to an off-topic (OT) and rectified the topic shift as well. This shift occurs abruptly, with stark difference in the semantic space between the two topics. Such a topic diversion and rectification is a natural phenomenon in a human-human conversation.

3 Related work

A good conversation is one which focuses on a balance between staying on topic and changing it in an interactive multi-turn conversation system (See et al., 2019). Detection of what constitutes as on-topic can be viewed as segmentation of conversation into relevant and irrelevant of the conversation (Stewart et al., 2006). Earlier work in segmenting conversations into topics expected a high lexical cohesion within a topic segment (Hearst, 1997). However, we see that they fail to have regard of sentence-level dependencies leading to fragmented segmentation (Takanobu et al., 2018). Various supervised methods approached this task as a classification problem (Arguello and Rosé, 2006) but annotations for them can be expensive and not scalable for large datasets. Unsupervised methods on goal-oriented conversations also have limited ability to learn from the dataset (Joty et al., 2013). Modelling this problem into detection of global topic structure and local topic continuity (Takanobu et al., 2018) results in a weakly supervised approach, using a hierarchical LSTM, to analyse dialogue context and content. However, a major drawback in that method is that the topic sets are predefined and the utterances are bucketed into the same. In an unbounded natural conversation, specifying the topic set in advance is not a feasible task.

Our proposed topic segmentation would help us introduce a system initiative module by figuring out *when* to give refinement or guidance and *how* to best contribute in solving a user's problem (Horvitz, 1999), by detecting the major topic of the conversation and steering the user towards it in case of a diversion.

4 Annotation Framework

We use the human-transcribed conversations from the NXT-format Switchboard corpus (Calhoun et al., 2010) in our task. In this dataset, participants are given a topic prompt and were asked to converse with each other for around ten minutes. This dataset was chosen for annotation, amongst others, as some did not have *enough turns* to observe a topic shift [(Lowe et al., 2015), (Gliwa et al., 2019)] or had *fixed topics* of conversation [(McCowan et al., 2005), (Janin et al., 2003)] neither of which were favourable for us to model an off-topic shift detection for open-domain conversations.

In Switchboard, we observe the freedom with which the participants drift from the given topic prompt, leading to different off-topic threads in the conversation and several statements by the users to steer the conversation back to the original topic. To model this property, we annotated the dataset, into three labels - major, minor and off-topic tags. Dialogues are inherently hierarchical in structure, but we see that human annotators cannot definitively agree on a hierarchical segmentation (Passonneau and Litman, 1997). Thus we adopt a flat model of annotation where a strong shift from the original topic of conversation is annotated as off-topic and a subsidiary shift is labelled as minor topic.

- **Major Topic (MT)** - The utterances which belong to the topic with which the conversation commenced with and is largely talked about were tagged as major topic. Each conversation has a solitary Major topic.

- **Minor Topic (MiT)** - The utterances that are part of a sub-topic, which was a natural digression from the major topic but lies in the semantic space of the major topic, are tagged as minor topic. A conversation can consist of multiple Minor Topics.
- **Off-topic (OT)** - The utterances that are part of a complete digression of the topic at hand were tagged as off-topic. Each conversation could encompass multiple instances of Off Topic clusters.

A conversational speech is not as structured as written text; it consists of overlaps of turns between the participants and interruptions. That is why each turn is divided into an utterance consisting of a single independent clause (Meteer and Iyer, 1996). This also helps us in narrowing down each utterance to have a single topic of discussion and thus a single tag to belong to. For our ease of annotation, we have considered incomplete sentence as complete sentences and annotated accordingly. We have also made a conscious decision to drop one word sentences.

4.1 Annotation Guidelines

The annotation process starts with the annotators identifying topic shifts in a conversation and bracketing the utterances. Each bracket is then mapped to an annotation tag of major, minor or off topic as seen in conversation 6. The annotators were given the following guidelines
(i) Annotators are advised to go through the entire conversation first before beginning the annotation process to get a better understanding of the topic flow. (ii) In most instances, conversations begin with a major topic bracket. (iii) Minor and off topic brackets are not further segmented. (iv) Minor topic bracket is always preceded by a major topic bracket.
A document tailing these guidelines along with appropriate examples was given to the annotators for reference. We have annotated the dataset [4] using three independent annotators and each utterance belonged to either major, minor or off-topic. The Cohen's kappa score or the inter evaluator agreement is 0.64 for our annotation, which indicates reliability.

Topic tag	**Frequency**
Major Topic	3206(30.4%)
Minor Topic	4759(45.2%)
off-topic	2560(24.4%)

Table 1: Frequencies of major, minor and off topic utterances in the dataset.

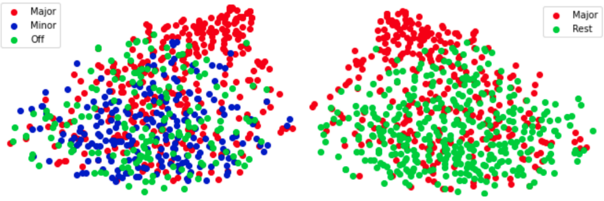

Figure 1: Image (left) shows the t-SNE representation of MT vs MiT vs OT classes whereas the (right) shows the t-SNE representation of MT vs rest classes.

We had observed that the major issue for disagreement lie in whether to tag a conversation as minor or off-topic. In cases of confusion, annotators were advised to tag the turn as minor-topic since the degree of digression from the major topic is subjective in nature. This resulted in the increase of minor topic tags over rest.

5 Experiments and Results

Prior to designing the topic classifier, we wanted to understand the characteristics of Switchboard corpus and visualize the classes that we have defined in Section 4. We plotted the t-SNE embeddings(Van der Maaten and Hinton, 2008) for the 3 classes in Fig 1(left). We observe that minor and off-topic classes are entangled and thus decided to merge these two classes into a *rest* class. The t-SNE plot for the data with the merged class can be seen in Fig 1(right), and the classes are now less entangled. Our task is now a binary classification task with the two classes being *major* and *rest*. This is further backed by the poor results obtained on the application of classification models to classify each classes individually, which we omit for brevity.

5.1 Methodology

Our task is to segment the conversation and label each segment with the tag of major or rest. More formally, given a conversation X having

[4] The dataset and annotation guidelines are available at this link

Model	Precision	Recall	F1 score
SVM	0.55	0.59	0.56
LightGBM	**0.65**	**0.69**	**0.66**
BERT-base	0.69	0.69	0.69
RoBERTa-base	0.77	0.63	0.69
XLNet-base	**0.84**	**0.72**	**0.77**

Table 2: LightGBM gives best results amongst the baselines. XLNet-base gives best results overall.

utterances $x_1, x_2, \ldots, x_n$ and the topic set $S = \{major, rest\}$. Our task is to segment these utterances into major topic or rest i.e., a binary classification task. To achieve this, we started with classical machine learning algorithms like SVM and LightGBM (Ke et al., 2017) and then we tested the latest sequence classification deep learning models like BERT (Devlin et al., 2018).

SVM and LightGBM are the two baselines calculated to compare against BERT and its variants. We have not used TextTiling, which is commonly used for dialog segmentation tasks as one of our baselines, because TextTiling measures the similarity of each adjacent sentence pair and uses valleys of similarities for segment detection. This is useful for datasets which have conversations with well defined topic shifts but the conversations in Switchboard do not have that property.

BERT and its variant models (RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2020)) are transformer based deep learning models. RoBERTa improves the training procedure by removing the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs. XLNet on the other hand is a bidirectional transformer, that uses better training methodology, larger data and more computational power to improve upon BERT. Our model was evaluated against precision, recall and F1 score. We see that good precision is a reliable metric to measure against. Our prime focus is on detection of the topic shift away from major topic, thus high precision gives us a better system to identify when topic shift occurs and label it accordingly.

5.2 Results

We fine-tune BERT by taking a pre-trained model, adding an additional untrained classifier layer and training this new model for our task. This is done because pre-trained Transformer model weights already encode a lot of information about our language which is helpful in cases where the datasets are small. For the sequence classification task, we use a special [CLS] token at the beginning of our sentence-chain which encodes the information of the sentence-chain into it. This token is used in the final layer to classify whether a sentence-chain belongs to a major topic or rest class. On observing the results, we find that the XLNet-based model outperforms BERT, RoBERTa and the baselines. We hypothesize that XLNet performs better than BERT and RoBERTa because it does not suffer from the problem of a fixed maximum length for tokens. Both BERT and RoBERTa allow maximum 512 tokens in a sentence whereas XLNet has no such limitation. This indicates a better coverage of utterances which consist of more than 512 tokens, a phenomenon observed many times in the dataset. During training entire context of the conversation is taken into account and the model is trained using the labels used for each sentence chain belonging to that conversation. While evaluating the model, a conversation is taken and every sentence chain is tested whether it belongs to major topic or not.

6 Case Study

The system response generated in this case study is a System Initiative (SI) given to a snippet of the Switchboard corpus, prompting the user to go back to the major topic of the conversation, when it detects a topic shift from it.

Setup The major bottleneck in generating a SI response is the detection of MT in an open-domain conversation. Since there are no predefined topics at hand, we see that one manner of MT detection could be using word importance scores which are scored using a bidirectional LSTM in the range of 0 to 5. (Kafle and Huenerfauth, 2018)

Major Topic Detection Our assumption in this case study was that the set of words with word importance scores > 4, in the first K turns of the conversation, contain the major topic in them. We test our assumption using the human-annotated major topics of the conversation. We evaluate the extracted Bag of Words (BoW) and the annotated data using cosine similarity score. After sampling for values of K ranging from 0 to 40, we see that the major topic is detected best when $K = 15$.

MT

A: So, do you fish?
B: Oh, yeah. My dad has a lake cabin.
B: and so we go there for the small lake, uh, just outside of the Dallas Fort Worth area.
A: Oh, that's nice

OT

A: I, I, You see, I'm from west Texas.
B: Oh, are you? Where are you from?
A: Lubbock
B: Oh, I'm from Midland.
A: Oh, another west Texan.
B: I went to college at Tech,

SI

Sys: *Do you want to go back to topic of fishing?*

Observation We observe the BoW extracted using word importance scores has a cosine similarity of 0.652 on an average with the human-annotated MT of the dataset. This helps us in generating a SI that can contribute towards the user's objective. We use a simple template-based response and add the component of major topic, to generate a user guided SI to steer the conversation back in case of a topic shift. The turn at which this SI should occur, is detected using our XLNet-based model to identify a shift from the major topic of the conversation. This helps us to support the user in their task and add a collaborative feature to the interactive agent.

7 Conclusion

In this paper, we looked at generating a system initiative module in a conversational system that does not interrupt the user and also works towards achieving the common goal of the user. We present a dataset that helps in training an XLNet-based model to correctly detect a digression from the major topic of the conversation. We have also looked at an application of this model as a case study where we detect topic shift and generate a system initiative for the rectification of the same. A predictable limitation of our system lies in not detecting minor and off-topic individually. This categorisation would help in giving a leeway in case of a shift to a minor topic thread and a system rectification initiative in case of a shift to an off-topic thread .

References

Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. *arXiv preprint arXiv:1705.00746*.

Jaime Arguello and Carolyn Rosé. 2006. Topic-segmentation of dialogue. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 42–49.

Alexander Bartl and Gerasimos Spanakis. 2017. A retrieval-based dialogue system utilizing utterance and context embeddings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1120–1125. IEEE.

Nicholas J Belkin and Alina Vickery. 1985. *Interaction in information systems: A review of research from document retrieval to knowledge-based systems*. 025.04 BEL. CIMMYT.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Eric Horvitz. 1999. Uncertainty, action, and interaction: In pursuit of mixed-initiative computing. *IEEE Intelligent Systems*, 14(5):17–20.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and*

Signal Processing, 2003. Proceedings.(ICASSP'03)., volume 1, pages I–I. IEEE.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.

Sushant Kafle and Matt Huenerfauth. 2018. A corpus for modeling word importance in spoken dialogue transcripts. *arXiv preprint arXiv:1801.09746*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

Staffan Larsson. 2017. User-initiated sub-dialogues in state-of-the-art dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 17–22.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100. Citeseer.

Marie Meteer and Rukmini Iyer. 1996. Modeling conversational speech for speech recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Rebecca J Passonneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Robin Stewart, Andrea Danyluk, and Yang Liu. 2006. Off-topic detection in conversational telephone speech. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 8–14.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Feng-Lin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *IJCAI*, pages 4403–4410.

George Terzopoulos and Maya Satratzemi. 2020. Voice assistants and smart speakers in everyday life and in education. *Informatics in Education*, 19(3):473–490.

Johanne R Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2):102162.

Yunli Wang and Cyril Goutte. 2018. Real-time change point detection using on-line topic models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2505–2515.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.