Designing a non-goal oriented Question Answering System for Soccer

Neeraj Ganu
Department of Computer Science
Stony Brook University
nganu@cs.stonybrook.edu

Nikhil Doifode
Department of Computer Science
Stony Brook University
ndoifode@cs.stonybrook.edu

Abstract

Non-goal oriented multi-hop question answering systems lack the ability to generate answers with grounded facts. We wanted to address the problem of generating well-grounded responses by integrating knowledge graphs into the dialogue system's response generation process. The proposed Recurrent Neural Network (RNN) model which uses the soccer conversation dataset, can also integrate knowledge graphs into the response generation process for domain knowledge, producing well-articulated, knowledge grounded responses. We also wanted to address the problem of updating knowledge graphs since teams information changes frequently in soccer. The proposed sentient gate uses knowledge graph more than the baseline model when factoid question is asked. The results show that our model performs better than other state-of-the-art models.

1 Introduction

Non goal oriented Question Answering (QA) systems are a first step towards chit-chat scenarios where humans engage in conversations with bots over non-trivial topics. This type of systems can benefit from added additional domain knowledge. So our objective is to create a system which will be able to handle factoid as well as non-factoid queries like chit-chats or opinions on the soccer related subjects/domains. In case of factoid queries, we want to generate well-articulated responses which are knowledge grounded while preserving co-references across the dialogue contexts. For example below snipped shows how the system should behave.

User utterance: Who is the captain of Argentina ?
Expected Response: Lionel Messi is the captain
User utterance: Do you know the name of their coach ?
Expected Response: Lionel Scaloni is the coach

In this snippet, system answers the first question with a well-articulated answer which is grounded in truth and for the second question system should understand that 'their' means Argentina given the context.

Eric et al [1] introduced an in-car dialogue dataset for multi-domain, task-oriented dialogues along with a knowledge graph which can be used to answer questions about the task the user wants to be assisted with. The dataset used in that system consists of dialogues from the following domains: calendar scheduling, weather information retrieval, and point-of interest navigation. For non-goal oriented dialogues, [5] proposed a dataset in the movie domain which contains short dialogues for factoid question answering over movies or for recommendations. They also provide a knowledge graph consisting of triples as (s, r, o). Where s is the subject, r stands for relations and o being the object. The movie dialogues can utilize this provided knowledge graph for recommendation and question answering purposes. [2] proposed a soccer related dataset and system which uses knowledge graph to answer factoid question.

The in-car system [1] only tackles the problem of factual and well-articulated response generation in dialogues for goal oriented tasks. [5] system only tackles the problem of factual response generation in dialogues for non-goal oriented tasks, but not well articulated ones. [2] system doesn't use the knowledge graph correctly meaning in case the if certain factoid question is answered as part of training dataset the system trust the answer from model more than the knowledge graph. The system also does not answer the question of knowledge graph updating since in soccer the teams information changes frequently and we should update the KG.

Question answering of domain specific non-goal oriented conversations and generating well-articulated responses is addressed by using hierarchical RNN based architecture models. Hierarchical RNN are used for incorporating more contextual information in the response generation process. For keeping the responses grounded in truth we are integrating Knowledge Graphs in the response generation [1]. The [2] system uses sentient gate for choosing between the output of RNN model and Knowledge graph we made some changes in the sentient gate which makes the system trust the Knowledge graph more. For the task of updating the Knowledge graphs we tried to use the systems like GUpdater [9]. GUpdater is built upon graph neural networks (GNNs) with a text-based attention mechanism to guide the updating message passing through the KG structures.

We used the [2] KG-Copy network as our baseline model. For updating the knowledge graph, we looked at a few approaches. Out of them GUpdater [9] and Web Scraping based were more relevant to the task at hand. GUpdater is based on transaction based approach where after each transaction the whole knowledge graph is updated which is too complex for the system we are trying to implement. So we used Wikipedia based web scraper which scraps information about the teams and updates the knowledge graph. For handling sentient gate problem we looked at the solution of [1] where the system doesn't use sentient gate which is possible by increasing the similarity score of matching embedding of question asked to system and knowledge graph.

Baseline model proposes a dataset of 2,990 conversations for non-goal oriented dialogues in the domain of soccer, over various club and national teams. We modified some questions in the test dataset to check whether the KG updates and sentient gate updates are working or not. We checked the output generated from baseling model against our model. Our system is evaluated against other state-of-the-art architectures and baseline model for knowledge grounded dialogue systems. The evaluation is done based on both knowledge groundedness using entity-F1 score and also standard, automated metrics (BLEU) for evaluating dialogue systems.

The main outcomes of this project are:

1. Added a Wikipedia based web scraper which updates knowledge graph with latest information

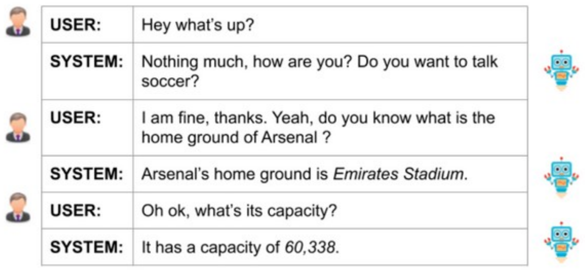
2. Modified sentient gate to choose the output of knowledge graph more than the output of RNN model

3. Modified similarity function so the similarity score between matching question asked to system and knowledge graph is better than the RNN output.

2 Task

The input to this system is a question or a series of questions(for context). The system will answer these questions based on its understanding of the language and knowledge of soccer gathered from training data and the Knowledge Graph(KG). Main challenge for the system are learning to comprehend non goal oriented conversations and keep context. The other complexity is to provide knowledge of soccer to the user.

Standard approaches have previously tried to use RNN and Knowledge graphs with great success for general conversations[1] and domain specific tasks[2]. More recent approaches have also tried using language models with knowledge graphs[3].



USER:	Hey what's up?
SYSTEM:	Nothing much, how are you? Do you want to talk soccer?
USER:	I am fine, thanks. Yeah, do you know what is the home ground of Arsenal ?
SYSTEM:	Arsenal's home ground is *Emirates Stadium.*
USER:	Oh ok, what's its capacity?
SYSTEM:	It has a capacity of *60,338.*

Figure 1: Task conversation example

2.1 Baseline Model

The baseline system that we referred to used a KG-copy network from Chaudhuri et al.[2]. The system is represented in Figure 2. The main components of this system are the Recurrent Neural Network (RNN) and the KG embeddings. It is influenced by the copynets approach[4].

2.1.1 Encoder RNN

They used a RNN layers as encoder mechanism for the model. It uses a single long-short term memory(LSTM) layer. It takes input word sequence and encodes it as a fixed length vector irrespective of input sequence length. It uses attention mechanism to decide weights for values of hidden states of words appearing before it.

2.1.2 Decoder RNN

The decoder is also an attention based LSTM layer. The input is the context and the hidden layer encoding from the encoder at time step T. The attention weights are then calculated by concatenating the hidden states seen so far and the decoder LSTM output. Finally, the wighted context representation is calculated as the summation of the product of the attention weights and outputs of the decoder LSTM.

This is referred to as Vocab Distribution in Figure 1.

2.1.3 KG embedding

First a simple averaging of the input query is done, this generates a embedding for the current input context. Then KG embeddings are calculated by taking an embedding of the local KG's subject entity and relation labels for each triple. Note that the paper mentions that this averaging is logically done only for nouns and verb phrases in each triple.

2.1.4 Sentient Gate

The sentient gate is a decider for: at time step T should the output depend on the KG or the training conversations. They write the final objective function as the probability of predicting the next word during decoding based on the encoder hidden-states and KG.

2.2 The Issues

The main issues with the approach used in the baseline method were that it would require updating of the Knowledge graph to keep the knowledge relevant. Another problem is that the sentient gate needs to be very smart about when to choose a word from the conversation knowledge and the knowledge graph.

Soccer has constantly changing dynamics. Players get transferred from one club to the other constantly. Any soccer conversation needs to reflect these facts and thus the automated replies need to be grounded in facts that are up to date. Figure 4 shows the Knowledge graph for the Arsenal Football Club but since then their Chairman has resigned and have hired a new coach,

The model itself is trained on very little data and for it to make decisions on when to go for semantic completion of sentences as compared to fact based data is very important. Just replying based on semantics may be fine in a normal conversation but not in conversations about a specific domain such as soccer. It also needs to avoid hard-coding of answers to question that may appear multiple times in the data set.

3 Approach

3.1 Update the Knowledge Graphs

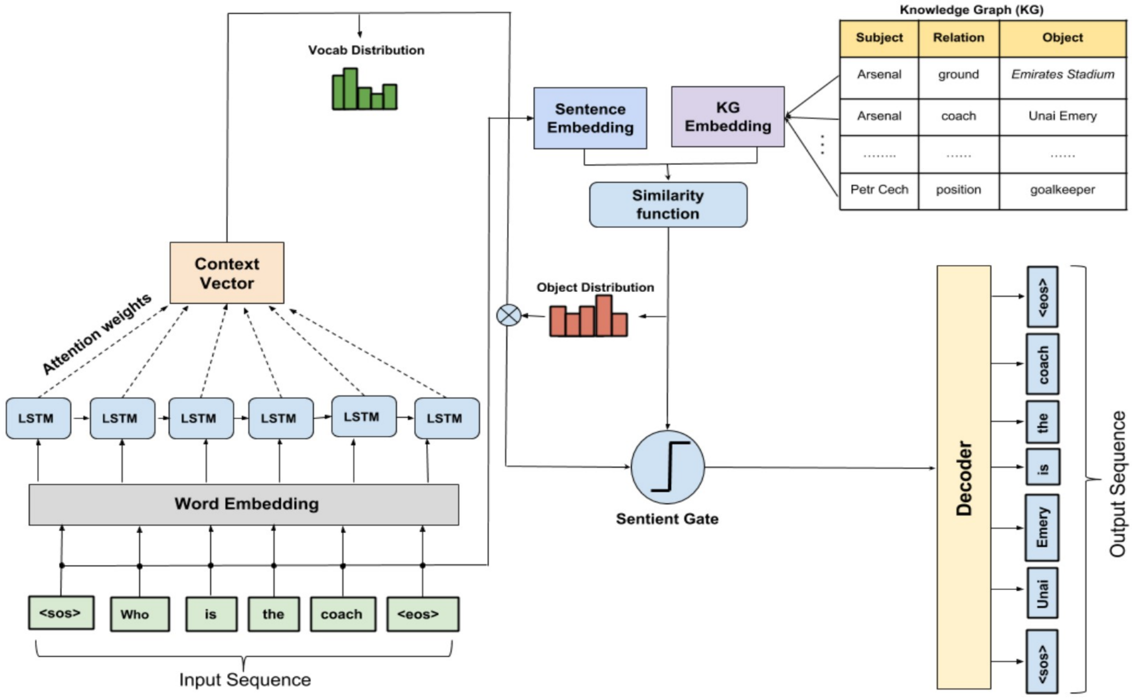
We used web scraping to scrape the data from Wikipedia for the knowledge graphs. Wikipedia has a favourable format to scrape the information. They have a page for all of the clubs and countries that we considered. We had to decide the relations that we wanted to include in the KG, this was also influenced by the original KG and what information Wikipedia has available.

Once the information can be scraped we need to arrange that as triples in this KG. Sometimes additional information was needed such as the number of appearances for a certain player. In such cases we have to further go to the player's personal Wikipedia page to get the information.

This however creates some challenges, some player pages don't have the information in the correct format. Skipping these fields is the only option in these cases. Example of an updated graph is shown in Figure 3. Note: Arsenal currently don't have a chairman.

3.2 Sentient Gate changes

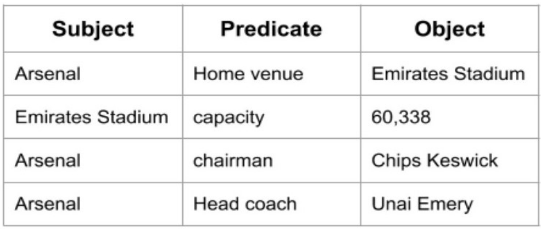
We observed that even after we updated the KG, the answers were not getting updated. This was because the sentient gate was no using KG facts to answer the questions. So we tried to change the sentient gate mechanism to use the facts more. This included changing of the similarity function, gate itself - dense layer and tried using both instead of choosing one.



Knowledge Graph (KG)

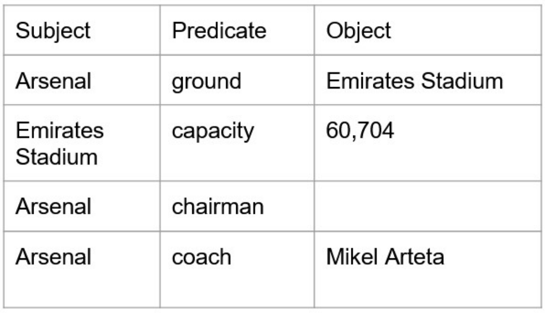
Subject	Relation	Object
Arsenal	ground	*Emirates Stadium*
Arsenal	coach	Unai Emery
....	......	....
Petr Cech	position	goalkeeper

Figure 2: KG-Copy Model Encoder-Decoder Architecture for Knowledge Grounded Re-sponse Generation.



Subject	Predicate	Object
Arsenal	Home venue	Emirates Stadium
Emirates Stadium	capacity	60,338
Arsenal	chairman	Chips Keswick
Arsenal	Head coach	Unai Emery

Figure 3: Arsenal Knowledge Graph snippet.



Subject	Predicate	Object
Arsenal	ground	Emirates Stadium
Emirates Stadium	capacity	60,704
Arsenal	chairman	
Arsenal	coach	Mikel Arteta

Figure 4: Updated Arsenal Knowledge Graph snippet.

3.3 Implementation Details

The scraping part of the project used basic knowledge of web design. The scraping was done in python using the BeautifulSoup package. We picked the clubs that were supported in the conversations database and in the original KGs.

For the model we used the architecture from the original paper. We changed the sentient gate to promote the use of KG for fact based questions. Some methods that we implemented:

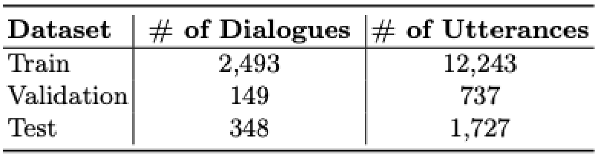
1. Changing similarity function
2. Changing embedding for KG
3. Adding layers for sentinel function

4 Evaluation

For evaluation of our system, we added some questions where we know knowledge graph is updated and hence the generated output should have information from the updated knowledge graph instead of old. Then we compared the result from the baseline model and our model to check the correctness of the system.

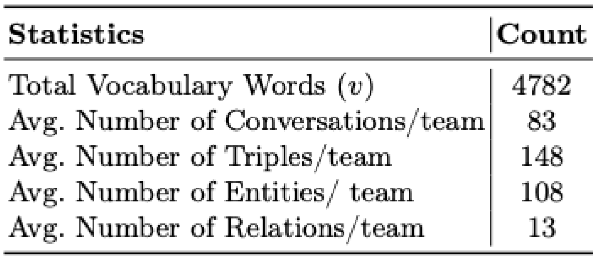
4.1 Dataset Details

The dataset of conversations over soccer is collected using AMT (Amazon Mechanical Turk) [6] in an wizard-of-oz style [7] setup. In such a setup, humans believe that they are interacting with machines, while the interaction is completely among humans. The turkers assigned to the system role were asked to use Wikipedia to answer the questions. We chose the knowledge graph of teams like Sweden, Spain, Senegal, Portugal, Nigeria, Mexico, Italy, Iceland, Germany, France, Croatia, Colombia, Brazil, Belgium, Argentina, Uruguay and Switzerland and the club teams selected were Barcelona, Real Madrid, Juventus, Manchester United, Paris Saint Germain, Liverpool, Chelsea, Atletico Madrid, Bayern Munich, Porto and Borussia Dortmund. The number of conversations are equally distributed across all teams. The statistics of the total number of conversations are given in Figure and .



Dataset	# of Dialogues	# of Utterances
Train	2,493	12,243
Validation	149	737
Test	348	1,727

Figure 5: Statistics of Soccer Dataset



Statistics	Count
Total Vocabulary Words (v)	4782
Avg. Number of Conversations/team	83
Avg. Number of Triples/team	148
Avg. Number of Entities/ team	108
Avg. Number of Relations/team	13

Figure 6: KG statistics

4.2 Evaluation Measures

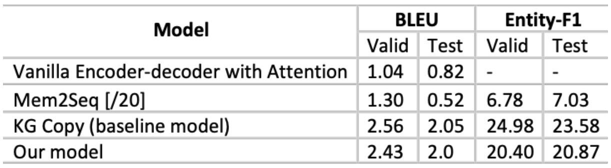
We compare our model with Baseline model, Mem2Seq and a vanilla encoder-decoder with attention. We report the average BLEU scores [8]. Although, BLEU is used for evaluating automatic machine translation that is language-independent and correlates highly with human evaluation. It has also been used in past literature for evaluating response generated by dialogue systems. We also use the average entity-F1 scores which evaluates the model's ability to generate relevant entities from the underlying knowledge base and to capture the semantics of the user initiated dialogue flow. We generate the results with these measures on valid and test soccer dataset.

4.3 Baselines

Baseline model uses a batch-size of 32 for 100 epochs for training. It applies Adam [10] for optimization with a learning rate of 1e-3 for the encoder and 5e-3 for the decoder. The size of the hidden layer of both the encoder and decoder LSTM is set to 64. Baseline model trains the decoder RNN with teacher-forcing [11]. The input word embedding layer is of dimension 300 and initialized with pretrained fasttext [12] word embeddings. A dropout [13] of 0.3 is used for the encoder and decoder RNNs and 0.4 for the input embedding.

Our model trains for 150 epochs while the hidden layer of both the encoder and decoder LSTM is set to 128. The reason for this was to force/train the model to use the KG more since we have updated the KG and the model should use the updated KG more for fact based questions.

4.4 Results



Model	BLEU		Entity-F1	
	Valid	Test	Valid	Test
Vanilla Encoder-decoder with Attention	1.04	0.82	-	-
Mem2Seq [/20]	1.30	0.52	6.78	7.03
KG Copy (baseline model)	2.56	2.05	24.98	23.58
Our model	2.43	2.0	20.40	20.87

Figure 7: Results on Soccer Dataset

The results show that our proposed model performs better than both the vanilla attention sequence-to-sequence models and Mem2Seq model across both metrics but slightly worse than the baseline model. Our model outperforms Mem2Seq by 1.48 in BLEU test score and 13.84% on entity-F1 test score. It performs better than the vanilla sequence-to-sequence model by 1.18 on the BLEU test metric. Our model is worse by 0.13 compared to Baseline model in BLEU test score and 2.71% on entity-F1 test score.

4.5 Analysis

Our model is able to generate the well-articulate responses which are well grounded in knowledge as demonstrated in the snippet below.

User utterance: who is the captain of argentina ?
Response: lionel messi is the captain
User utterance: do you know the name of their coach ?
Response: lionel scaloni is the coach

As you can see in the snippet above, the first response is factually correct and also a well-articulate one. In the last response, the model is able to identify that the pronoun ”their” refers to the team and is able to maintain a knowledge grounded, as well as articulate responses even for relatively long dialogue turns.

Error Analysis:
The first problem we faced in the test dataset itself. The test dataset has questions and answers for evaluation. So after updating the knowledge graph, if the test dataset has a question whose response was made according to the old knowledge graph and the system gives the output by using the updated knowledge graph. The following snippet demonstrates this problem.

User utterance: Who is the coach Arsenal?
Test Dataset Response: Unai Emery is the coach
Our response: Mikel Arteta is the coach

As you can see in the snippet above, the Test dataset has response "Unai Emery is the coach" which is according to the old knowledge graph whereas the correct response is "Mikel Arteta is the coach" which our system gives. So when we do the evaluation, this is classified as wrong response even though it is right. We thought about updating the test dataset but it requires a lot of manual work.

The model also produces too generic and non-factual responses for queries about opinions about best or favorite players as shown in following snippet.

User utterance: Who is senegal's best current player not including mane?
True response: keita balde diao
System response: i think it is the best player in the world cup
User utterance: who's your favorite player in Argentina?
True response: messi
System response: i think eden hazard is the best player

As you can see in the snippet above, for the first response, the model is not able to interpret the question and generates a generic response. For the second case, the generated response is not factual because the question is about Argentina, but Eden Hazard is from a different team (Belgium). We think this maybe because the vocabulary size is relatively large compare to the size of training dialogues and can be improved with more training data.

The model also often suffers when more complex quantitative and other reasoning skills are required to respond to the context as demonstrated in the snippet below.

User utterance: how many world cups has the brazil team won ?
True Response: brazil has won the fifa world cup five times
System response: they won the world cup
User utterance: who was the top scorer in the world cup for belgium?
True Response: eden hazard
System response: i think it was the top scorer for the world cup

As you can see in the snippet above, for the first response, the model needs to perform a count operation over the Knowledge Graph to answer it, which is currently unsupported. Similarly, for the second case the model would require better language understanding to respond. Sometimes the model also suffers from the problem of unknown words in the test set.

4.6 Code

All code and data for this paper are available on our Github repository:
https://github.com/nikhildoifode/NLP_project

5 Conclusions

We have shown that by updating KG graphs we can get our model to output updated knowledge, keeping it relevant and factual. While more data can help the model in making better decisions on how to respond, the current model displays good level of conversational fluency and domain knowledge. This work also shows that using Knowledge Graphs in fields where domain knowledge can't be easily learned from conversation can give more fact centred conversations. This can be of particular interest in fields such as medicine, economics and various sports where the common vernacular can have other meanings. Even though our model uses knowledge graph sparsely it performs better than the models which do not use knowledge graphs.

6 References

1. Eric, M., Krishnan, L., Charette, F., Manning, C.D.: Key-value retrieval networksfor task-oriented dialogue. In: Proceedings

of the 18th Annual SIGdial Meetingon Discourse and Dialogue. pp. 37–49. Association for Computational Linguistics(2017). https://doi.org/10.18653/v1/W17-5506

2. Chaudhuri D., Rony M.R.A.H., Jordan S., Lehmann J. (2019) Using a KG-Copy Network for Non-goal Oriented Dialogues. In: Ghidini C. et al. (eds) The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science, vol 11778. Springer, Cham

3. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P. (2020). K-BERT: Enabling Language Representation with Knowledge Graph. ArXiv, abs/1909.07606.

4. Gu, J., Lu, Z., Li, H., Li, V.O.: Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Associationfor Computational Linguistics (Volume 1: Long Papers). pp. 1631–1640. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-1154

5. Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., Weston, J.: Evaluating prerequisite qualities for learning end-to-end dialog systems. ICLR (2016)

6. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6(1), 3–5 (2011). https://doi.org/10.1177/1745691610393980, https://doi.org/10.1177/1745691610393980, pMID: 26162106

7. Rieser, V., Lemon, O.: Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. Proceedings of ACL-08: HLT pp. 638–646 (2008)

8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)

9. Tang, Jizhi Feng, Yansong, Zhao, Dongyan. (2019). Learning to Update Knowledge Graphs by Reading News. 2632-2641. 10.18653/v1/D19-1265.

10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

11. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation 1(2), 270–280 (1989)

12. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146 (2017)

13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)