Why “Blow Out”? A Structural Analysis of the Movie Dialog Dataset

Richard Searle
Eccentric Data, Cambridge
richard@eccentricdata.com

Megan Bingham-Walker
Eccentric Data, Cambridge
megan@eccentricdata.com

Abstract

A long-term goal of machine learning is to build an intelligent dialogue agent that is capable of learning associations within data and using them to understand and answer questions and make relevant recommendations. The Facebook Movie Dialog Dataset (MDD) was recently proposed in Dodge et al. (2016) to evaluate the comparative performance of dialogue agent systems. However, a structural analysis of the data for the recommendation tasks suggests that there may be some flaws in the design of the dataset.

1 Introduction

There has been a recent upsurge of commercial interest in the development of intelligent dialogue systems to answer questions, provide personalized recommendations and deliver services across a range of different domains. Some of the challenges that an intelligent dialogue agent will need to overcome in order to fulfill these objectives are: to be able to learn incrementally from heteregenous sources; to be able to adapt to changes in context; and to hold information in a long-term structured memory for rapid recall (Bordes et al., 2015).

Facebook recently proposed the Movie Dialog dataset (MDD) in Dodge et al. (2016) to encourage new research on this topic by providing a benchmark to evaluate the specific strengths and weaknesses of such systems. The MDD is part of the bAbI project of research into methods for automatic text understanding and reasoning (Weston et al., 2015). The MDD supports four question answering tasks featuring information retrieval and movie recommendation.

We had already developed an efficient, structural approach to incrementally learn clusters of associated products from high-dimensional, time-series data, for marketing personalized structured products to the clients of a financial services company. A structured product is an investment product that comprises a basket of underlying financial instruments, for example, equities, debt issuance, commodities, currencies or a combination of globally traded securities. In this case, the sparse hyper-dynamic nature of the product combinations and customer preferences necessitated an automated, algorithmic ranking metric. User preferences are conditioned with respect to the characteristics of a particular product offer, whereas individual products are a function of their underlying components. This decoupling of user preferences from the attributes of the structured product proves problematic when trying to identify an affinity between discrete users and potential product offers using techniques such as latent factorization.

Since our prior work constituted a generic, unsupervised approach, we were curious to see how it could perform against recommendation tasks in other domains and whether the structural model could be extended to act as a long-term memory for question answering.

One of the benefits of a structural approach, encompassing graphical learning representation, is that it is possible to track the reasons why certain recommendations are made, which can be beneficial to users (Tintarev and Masthoff, 2007) and enables accurate evaluation of the performance of the system. We report results from the application of a dynamic structural model as a contextual memory for Task 2 and Task 3, as defined in Dodge et al. (2016). Analysis of our findings reveals a possible flaw in the design of the Facebook Movie Dialog Dataset which may explain the results obtained by our method and those previously

Proceedings of the 1st Workshop on Representation Learning for NLP, pages 215–221,
Berlin, Germany, August 11th, 2016.

reported by Dodge et al. (2016).

2 Recommendation Tasks

2.1 Task 2 - Undirected recommendations

Task 2 aims to test whether an intelligent dialogue agent is able to make valid personalized recommendations for a user.

Dodge et al. (2016) selected ∼11k movies from the MovieLens dataset. They randomly sampled ∼110k users and for each of these users selected 1-8 movies that the user had rated 5. A statement was generated to express the user's opinion of these movies from a template, and this forms the statement input text (see Table 1 below). From the list of other movies that the same user had also rated 5, they randomly selected one to be the target answer representing a valid recommendation based on the composition of the basket of movies in the statement.

Statement
Gentlemen of Fortune, Revanche,
Eternal Sunshine of the Spotless Mind,
Prometheus, Fanny and Alexander, The Hurt
Locker, and 127 Hours are films I really like.
Would you recommend something to watch?
Answer
Blow Out

Table 1: Task 2 Example.

Data from ∼110k users was sampled multiple times to generate 1M training samples, data from ∼1k users was used to generate 10k development samples and data from a further ∼1k users was used to generate 10k test samples.

An important aspect of the MDD defined by Dodge et al. (2016) is that the number of users in each dataset is expanded by a factor of ∼10 through random sampling with replacement. The resultant datasets do not, however, identify unique users and, thus, preferred aggregations of movies by user. This characteristic of the MDD prevents user profiling and restricts the selection of recommended movies in Task 2 and Task 3 to a probabilistic similarity measure.

2.2 Task 3 - QA and contextual recommendations

Task 3 evaluates a short three-stage dialogue involving a combination of QA and recommendations that draws on additional contextual information. The first question is in the same format as Task 2, with the addition of a selection criterion - the writer and director Brian De Palma, in the example shown in Table 2. The second question is a factoid question about the response to the first question. The third question is a request for a follow-up recommendation, which refers to the context provided in the first question and a secondary contextual criterion defined by the statement in question 3 (see Table 2 below).

Statement 1
Gentlemen of Fortune, Revanche,
Eternal Sunshine of the Spotless Mind,
Prometheus, Fanny and Alexander, The Hurt
Locker, and 127 Hours are films I really like.
I'm looking for a Brian De Palma movie.
Answer 1
Blow Out
Statement 2
Who does that star?
Answer 2
John Travolta, John Lithgow,
Nancy Allen, Dennis Franz
Statement 3
I prefer Robert De Niro movies.
Can you suggest an alternative?
Answer 3
Hi, Mom!

Table 2: Task 3 Example.

3 Related Work

A network-based representational model has been demonstrated to be an accurate and efficient method for information retrieval, inference and reasoning for question answering (Berant et al., 2013; Berant et al., 2014; Hixon et al., 2015; Guu et al., 2015). Recent research has demonstrated some success in learning undirected graphical structural models from data (Lake and Tenenbaum, 2010; Mao et al., 2015) and it has also been shown that associations in data, intrinsic to a network-based architecture, form an important element of human learning (Spelke and Kinzler, 2007; Gershman, 2015).

We sought to augment a structural network-based representation of entities (in this case movies) and their attribute features by using the training data to learn statistical relationships between the entities from their recommendation histories. This form of relationship learning is consistent with the treatment of each sample as a unique user outcome and enables a top-*k* for *k* = number of hits (@100 for Task 2 and @10 for Task 3) ranking of the distance metrics between movie titles. The resultant ranking is used to assess the prediction accuracy of the recommendations made by the system.

4 Methodology

It is possible to conceive two primitive methods for recommending a movie based on the sample training data:

1. recommendation according to user movie preferences, by movie attribute (commonly achieved through the use of techniques such as latent matrix factorization), and
2. recommendation according to movie similarity, by movie attribute or user co-preference (as defined by a probabilistic distance measure between items or attribute features).

Since unique user data was lost in the construction of the MDD, it is not possible to associate movies with communities of users, which is a technique that we have found to aid both computational efficiency and accuracy of personalized user profiling in financial services. Instead, recommendations must be generated from either the similarity of movie attributes or the frequency of users co-rating movies as a 5. We will show that the latter method is the only viable approach to selection of a candidate recommendation using the MDD. Dodge et al. (2016) do not disclose the basis for their calculation of discrete and cumulative accuracy for Task 3 and the example data, shown in Table 1 and Table 2, does not discriminate between films with identical titles (e.g. versions of "20,000 Leagues Under the Sea" were released in 1916, 1954, and 1997). Consequently, we report accuracy of recommendation based on hits@100 or hits@10 with respect to correct identification of the movie title only. In Task 3, question 2 refers to a specific movie and for this task element we report the accuracy of our system under the assumption that a unique film was recommended with additional discriminatory training data (i.e. a hits@1 selection with a definitive release year, e.g. "20,000 Leagues Under the Sea" + 1954). Since, Task 3, question 2 cannot be accurately answered without a definitive answer to question 1, we do not report a mean accuracy as we consider this to be a misleading representation of the effective accuracy of a system over the task. Instead, we report the cumulative accuracy, α_{T3}, over Task 3 given by the formula:

$$\alpha_{T3} = (P(\alpha_{Q1}) * P(\alpha_{Q2})) * P(\alpha_{Q3}) \quad (1)$$

The cumulative accuracy α_{T3} is shown in Table 3 for our structural approach and is applied to the results reported by Dodge et al. (2016). We suggest that the cumulative accuracy α_{T3} represents a more realistic evaluation of the predicted accuracy of system responses in a dialog exchange of the type characterizing Task 3, whereby the success of a system is governed by the conditional probability introduced by the question sequence.

5 Building the Structural Model

Algorithm 1 below illustrates how the structural model is built from the movie knowledge base, training and development data. The variable *b* is defined as the movies in the basket of movies in the statement. The variable *r* is the recommended movie in the answer.

Algorithm 1 Building Structural Model

```
procedure CreateGraph($K$)
    load movie knowledge base
    for movie in knowledge base do
        $K \leftarrow$ add node movie
        $K \leftarrow$ add movie attributes
    load task training data as d
    for task in MDD tasks do
        for user in $d$ do
            for $b$ in basket and $r$ in rec do
                $K \leftarrow$ add edge($b$, $r$)
                $K \leftarrow$ count users, freq rec
            for edge in $K$ do
                distance metric m($b$, $r$)
    freeze system memory $K$ for testing
```

Algorithm 2 below provides a workflow for how a prediction is generated from the test data for Tasks 2 and 3. In accordance with the experimental protocols, Task 2 top-*k* is reported for hits@100 and Task 3 top-*k* is reported for hits@10.

On completion of training and development cycles, the structural model comprising the system

Algorithm 2 Recommend using item similarity

```
procedure MAKEREC(basket, criteria)
    load task test data as sample
    for s in sample do
        for $b$ in basket and $c$ in candidates do
            $rec[s]$ = edge($b$,$c$) if $c[crit]$=True
        for $r$ in rec[$s$] do
            $rec \in rec[s]$ if metric $e$[m] in top-$k$
```

memory is frozen. However, in practice, the system memory should continuously update to reflect user feedback on the recommended movies, a capability that is embedded in our dynamic model but not applied here. Such a capability is consistent with some of the challenges that an intelligent dialogue agent will need to overcome, as noted by Bordes et al. (2015).

6 Results and Analysis

We report our results in conjunction with those reported in Dodge et al. (2016) in Table 3.

Our graphical model renders the information retrieval Task 3, question 2 and Task 1 trivial, subject to valid data being held in the long-term memory. Following initial concern regarding the accuracy of our system, we are satisfied that our structural approach produces a valid model as a basis for movie recommendations using the statistical relationships encoded within the graph. This begs the question as to why the recommendation accuracy of our system and those produced by the method of Dodge et al. (2016) is so low?

Dodge et al. (2016) suggests that the reason why results for Task 2 are lower than for the information retrieval Task 1 and comparable Task 3, question 2 is due to missing labels, as a consequence of the sampling methodology they describe. We contend, however, that the basis of recommendation imposed by the MDD is flawed and that the frequency of occurrence of movies established by the effective 1M user population generates super-nodes that penalize valid answers with sparse structural association. Furthermore, our structural model of the MDD enables detailed analysis of the causes of recommendation error, as shown in Table 6 in the Appendix. Examination of the statistical relationships between answer recommendations and the statement basket of movies reveals that in many cases the sparse association and characteristic attributes of the answer provides no statistical basis for its inclusion as a recommendation in preference to other candidate films.

We consider that the MDD construction of the basket-recommendation relationship by arbitrary selection of films rated 5 by a user does not indicate the suitability of the proposed answer as a recommendation, hence the title of our paper. For the Task 3 example shown in Table 1, the movie “Blow Out” was not rated 5 by any user that also rated the basket movies 5. For the example in Table 2, “Blow Out” could not, therefore, be recommended by our system without recourse to joint training and is included within the “No association” error for Task 3 in Table 6.

For Task 2, “Blow Out” is associated with three of the basket films; “Eternal Sunshine of the Spotless Mind” (ESSM), “Fanny and Alexander” (FA), and “The Hurt Locker” (THL). However, as shown in Table 5, the sparse association of “Blow Out” with the basket films excludes it from the top-k for k = 100 strongest recommendations as the similarity metric for “Blow Out” falls below the minimum threshold for the top-k recommendations. The weak similarity metric generated between “Blow Out” and the basket movies is a product of the disparity between the number of users rating the basket movie as 5 and the fact that only 158 users rated “Blow Out” as 5.

Importantly, the genre information in Table 4 also illustrates that criteria-based association of recommendations is prevented by the heterogeneous nature of the MDD basket compositions. We evaluated this approach in the development of our system, but found reduced correlation of movies when compared to the use of frequency of user rating as the basis for a valid distance metric.

Furthermore, where the correct answer is identified by the system as a potential candidate, the imposition of a statistically valid top-k ranking excludes the majority of answers in favour of super-nodes that feature more prominently within films rated 5 by all users. The distribution of these dominant movie titles suggests a causal link between the reported accuracy of our system and those described by Dodge et al. (2016). The improvement shown in the results of Dodge et al. (2016) over our own may be attributable to the difference between our, definitive, structural method, and the alternative, parametric methods described in their research.

We consider that the hash lookup employed by

Methods	**Recs Task**	**QA+Recs**	**QA+Recs**	**QA+Recs**	**QA+Recs**
	Task 2	Task 3	Task 3	Task 3	Task 3
		Question 1	Question 2	Question 3	Cumulative
	hits@100	hits@10	hits@1	hits@10	
LSTM	27.1	35.3	14.3	9.2	3.2
Supervised Embeddings	29.2	56.7	76.2	38.8	22.0
MemN2N	28.6				
MemN2N (2 hops)		53.4	90.1	88.6	47.3
Structural Model	20.0	46.2	100.0	70.5	32.6

Table 3: Test results for Task 2, benchmarked against Dodge et al. (2016, Table 6) and test results for the individual questions in Task 3 benchmarked against Dodge et al. (2016, Table 9). Results reported as percentage accuracy.

Basket movie	**Movies co-rated 5**	**Users rating movie 5**	**Users rating "Blow Out" 5**	**"Blow Out" recommendations**	**Genre**
ESSM	4511	17485	2	1	*unknown*
FA	2160	1801	1	0	Drama
THL	1919	1283	1	0	War

Table 4: Structural association of "Blow Out" with basket movies for Task 2 example shown in Table 1.

Dodge et al. (2016) may introduce the possibility of conflation error by virtue of the inclusion of candidate movie titles on the basis of their semantic or syntactic structure. The embedding of movie titles without recourse to their probabilistic association with the expressed basket of films liked by the user may yield false positives in the case of the MDD, which will augment the evaluated accuracy of a system. In practice, however, the repeatability and accuracy of such a system may prove problematic.

7 Conclusion

Our experience of personalized user profiling in the financial services sector and analysis of the application of our method to the MDD tasks suggests that a combination of different methods may represent the most efficient path to effective, contextual personalized recommendations. In particular, the use of parametric candidate selection and relaxation of the strict statistical association required for candidate films helps to overcome issues of dominance in sparse, high-dimensional datasets. Critical factors in the success of both supervised and unsupervised approaches to recommendation are, however, the primacy and individual characteristics of the user and distinct user communities that support latent factorization methods. We believe that a simple reconfiguration of the MDD to reflect these characteristics would enable a more informative analysis of competing methods and technologies and thus contribute to fulfilling the objectives for intelligent dialog agents as set out by Bordes et al. (2015).

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Brad Huang, Christopher D. Manning, Abby Vander Linden, Brittany Harding, and Peter Clark. 2014. Modeling Biological Processes for Reading Comprehension. In *2014 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1510, Doha. Association for Computational Linguistics.

Antoine Bordes, Jason Weston, Sumit Chopra, Tomas Mikolov, Arman Joulin, and Léon Bottou. 2015. Artificial Tasks for Artificial Intelligence. In *2015 International Conference on Learning Representations*, San Diego.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *2016 International Conference on Learning Representations*, San Juan.

Basket movie	**"Blow Out" similarity metric**	**"Blow Out" Pr(recommended by basket movie)**	**"Blow Out" Pr(same user rating 5)**
ESSM	0.00011	0.01351	0.01266
FA	0.00051	0	0.00633
THL	0.00694	0	0.00632
k rank min. threshold for hits@k = 100	0.00980	0.3	0.25
***k* rank**	3906	2911	3088

Table 5: k rank of "Blow Out" for example basket of movies for Task 2 example shown in Table 1 (similarity metric and probabilistic k^{th} rank from $K = 5235$ candidate movies shown).

Samuel J. Gershman. 2015. A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, 11(11):e1004567.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing Knowledge Graphs in Vector Space. In *2015 Conference on Empirical Methods in Natural Language Processing*, pages 318–327, Lisbon. Association for Computational Linguistics.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 851–861, Denver. Association for Computational Linguistics.

Brenden Lake and Joshua Tenenbaum. 2010. Discovering Structure by Learning Sparse Graphs. In *Cognition in Flux: Proceedings of the 32nd Cognitive Science Conference*, pages 778–783, Portland. Cognitive Science Society.

Qi Mao, Li Wang, Ivor W. Tsang, and Yijun Sun. 2015. A Novel Regularized Principal Graph Learning Framework on Explicit Graph Representation. *arXiv preprint arXiv:1512.02752*.

Elizabeth S. Spelke and Katherine D. Kinzler. 2007. Core knowledge. *Developmental Science*, 10(1):89–96.

Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *WPRSIUI Associated with ICDE'07*, pages 801–810. IEEE.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *2016 International Conference on Learning Representations*, San Juan.

A Test Data and Discussion

Table 6 below provides details of the test results obtained for Task 2 and Task 3 using the MDD and test protocol defined by Dodge et al. (2016). We identified three principal causes of error:

1. the absence of an association between basket movies and the answer movie, due to no common user rating a basket film and an answer film 5;
2. low k rank of the answer film excluding it from the top-k for k = 100 or k = 10 recommendations for the basket of movies liked by a user;
3. for Task 3; the absence of a selection criterion from the answer film, thereby excluding its inclusion in the list of possible candidate recommendations. This is consistent with the observation made by Dodge et al. (2016) regarding potential errors due to missing labels.

Using the movie knowledge base as a long-term memory, we discern 17,928 unique movies and construct >3M edges within our structural model for both Task 2 and Task 3.

We did not apply joint training over the tasks but note that this would yield an improvement in the results by reducing errors due to absence of association, as illustrated by Table 4 since "Blow Out" is only connected to "Eternal Sunshine of the Spotless Mind" for Task 2 and is not connected to any of the basket movies for the Task 3 example shown in Table 2.

Table 6 indicates that the vast majority of errors are attributable to the low statistical association between basket films, compositions of basket films, and the hypothesized recommendations generated by random selection from specific

	Recs Task	**QA+Recs**	**QA+Recs**	**QA+Recs**
	Task 2	Task 3	Task 3	Task 3
		Question 1	Question 2	Question 3
	hits@100	hits@10	hits@1	hits@10
Test samples	10000	4915	4915	4470
Correct answers	1996	2269	4915	3153
Errors:				
Total	8004	2646	0	1317
No association	232	146	0	716
Low *k* rank	7772	2428	0	532
Accuracy (%)	20.0	46.2	100.0	70.5

Table 6: Test results breakdown for Tasks 2 and 3.

users' movie ratings. We contend that it is the underlying methodology behind the construction of the MDD that leads to the poor accuracy reported in Table 6 and not the intrinsic design of our system.

We attribute the improved results of Dodge et al. (2016) shown in Table 3 to the inclusion of answer recommendations on the basis of parametric affinity with the basket movie titles, rather than their statistical relevance as a potential user selection.

This may occur through the conflation of movie titles with independent basket-answer instances on account of words within their titles or characteristic attributes. Although we explored alternative methods for defining a statistical association based on the propinquity of movie attributes or attribute ranking, we were unable to identify a rigorous methodology that improved our reported accuracy for either Task 2 or Task 3.