Movie-DiC: a Movie Dialogue Corpus for Research and Development

Rafael E. Banchs
Human Language Technology
Institute for Infocomm Research
Singapore 138632
rembanchs@i2r.a-star.edu.sg

Abstract

This paper describes Movie-DiC a Movie Dialogue Corpus recently collected for research and development purposes. The collected dataset comprises 132,229 dialogues containing a total of 764,146 turns that have been extracted from 753 movies. Details on how the data collection has been created and how it is structured are provided along with its main statistics and characteristics.

1 Introduction

Data driven applications have proliferated in Computational Linguistics during the last decade. Several factors, such as the availability of more powerful computers, an almost unlimited storage capacity, the availability of large volumes of data in digital format, as well as the recent advances in machine learning theory, have significantly contributed to such a proliferation.

Among the many applications that have benefited from this data-driven boom, probably the most representative examples are: information retrieval (Qin *et al.*, 2008), machine translation (Brown *et al.*, 1993), question answering (Molla-Aliod and Vicedo, 2010) and dialogue systems (Rieser and Lemon, 2011).

In the specific case of dialogue systems, data acquisition can impose some challenges depending on the specific domain and task the dialogue system is targeted for. In some specific domains, in which human-human dialogue applications already exists, data collection is generally straight forward, while in some other cases, data design and collection can constitute a complex problem (Williams and Young, 2003; Zue, 2007; Misu *et al.*, 2009).

Depending on the objective being pursued, dialogue systems can be grouped into two major categories: task-oriented and chat-oriented systems. In the first case, the system is required to help the user to accomplish a specific goal or objective (Busemann *et al.*, 1997; Stallard, 2000). In the second case, the system objective is mainly entertainment oriented. Systems in this category are required to play, chitchat or just accompany the user (Weizenbaum, 1966; Wallis, 2010).

In this work, we focus our attention on dialogue data which is suitable for training chat-oriented dialogue systems. Different from task-oriented dialogue collections (Mann, 2003), instead of being concentrated on a specific domain or area of knowledge, the training dataset for a chat-oriented dialogue system must cover a wide variety of domains, as well as be able to provide a fair representation of world-knowledge semantics and pragmatics (Bunt, 2000). To this end, we have collected dialogues from movie scripts aiming at constructing a dialogue corpus which should provide a good sample of domains, styles and world knowledge, as well as constitute a valuable resource for research and development purposes.

The rest of the paper is structured as follows. Section 2 describes in detail the implemented collection process and the structure of the generated database. Section 3 presents the main statistics, as well as the main characteristics of the resulting corpus. Finally, section 4 presents our conclusions and future work plans.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 203–207,
Jeju, Republic of Korea, 8-14 July 2012.

2 Collecting Dialogues from Movies

As already stated in the introduction, our presented dialogue corpus has been extracted from movie scripts. More specifically, scripts freely available from The Internet Movie Script Data Collection (http://www.imsdb.com/) have been used. In this section we describe the implemented data collection process and the data structure finally used for the generated corpus.

As a first step of the collection construction, dialogues have to be identified and extracted from the crawled html files. Three basic types of information elements are extracted from the scripts: speakers, utterances and context.

The utterance and speaker information elements contain what is said at each dialogue turn and the corresponding character who says it, respectively. Context information elements, on the other hand, contain all additional information/texts appearing in the scripts, which are typically of narrative nature and explain what is happening in the scene.

Figure 1 depicts a browser snapshot illustrating the typical layout of a movie script and the most common spatial distribution of the aforementioned information elements.

It is important to mention that a lot of different variants to the format presented in Figure 1 can be actually encountered in The Internet Movie Script Data Collection. Because of this, our parsing algorithms had to be revised and adjusted several times in order to achieve a reasonable level of robustness that allowed for processing the largest possible amount of movie scripts.

Another important problem was the identification of dialogue boundaries. Some heuristics were implemented by taking into account the size and number of context elements between speaker turns.

A post-processing step was also implemented to either filter out or amend some of the most common parsing errors occurring during the extraction phase. Some of these errors include: corrupted formats, turn continuations, notes inserted within the turn, misspelling of speaker names, etc.

In addition to this, a semi-automatic process was still necessary to filter out movie scripts exhibiting extremely different layouts or invalid file formats. Approximately, 17% of the movie scripts crawled from The Internet Movie Script Data Collection had to be discarded. From a total of 911 crawled scripts, only 753 were successfully processed.

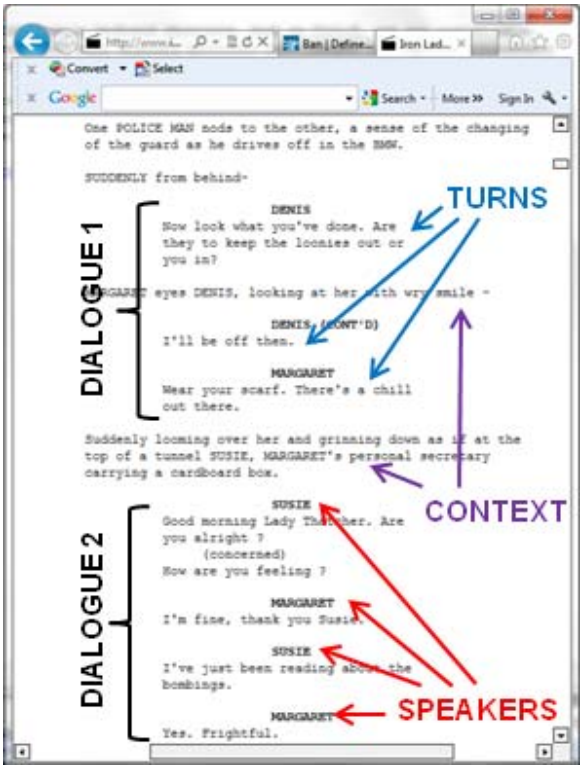

Figure 1: Typical layout of a movie script

The extracted information was finally organized in dialogical units, in which the information regarding turn sequences inside each dialogue, as well as dialogue sequences within each movie script was preserved. Figure 2 illustrates an example of the XML representation for one of the dialogues extracted from *Who Framed Roger Rabbit.*

```
<dialogue id="47" n_utterances="4">
   <speaker>VALIANT</speaker>
      <context></context>
      <utterance>You shot Roger.</utterance>
   <speaker>JESSICA RABBIT</speaker>
      <context>Jessica moves the box aside and
tugs on the rabbit ears. The rabbit head pops
off. Underneath is a Weasle. In his hand is the
Colt .45 Buntline.</context>
      <utterance>That's not Roger. It's one of
Doom's men. He killed R.K. Maroon.</utterance>
   <speaker>VALIANT</speaker>
      <context></context>
      <utterance>Lady, I guess I had you pegged
wrong.</utterance>
   <speaker>JESSICA RABBIT</speaker>
      <context>As they run down the
alley...</context>
      <utterance>Don't worry, you're not the
first. We better get out of here.</utterance>
</dialogue>
```

Figure 2: An example of a dialogue unit

3 Movie Dialogue Corpus Statistics

In this section we present the main statistics of the resulting dialogue corpus and study some of its more important properties. The final dialogue collection was the result of successfully processing 753 movie scripts. Table 1 summarizes the main statistics of the resulting dialogue collection.

Total number of scripts collected	911
Total number of scripts processed	753
Total number of dialogues	132,229
Total number of speaker turns	764,146
Average amount of dialogues per movie	175.60
Average amount of turns per movie	1,014.80
Average amount of turns per dialogue	5.78

Table 1: Main statistics of the collected movie dialogue dataset

Movies were mainly crawled from the action, crime, drama and thriller genres. However, as each movie commonly belongs to more than one single genre, much more genres are actually represented in the dataset. Table 2 summarizes the distribution of movies by genre (notice that, as most of the movies belong to more than one genre, the total summation of percentages exceeds 100%).

Genre	**Movies**	**Percentage**
Action	258	34.26
Adventure	133	17.66
Animation	22	2.92
Comedy	149	19.79
Crime	163	21.65
Drama	456	60.56
Family	31	4.12
Fantasy	82	10.89
Horror	104	13.81
Musical	18	2.39
Mystery	95	12.62
Romance	123	16.33
Sci-Fi	129	17.13
Thriller	329	43.69
War	25	3.32
Western	11	1.46

Table 2: Distribution of movies per genre

The first characteristic of the corpus to be analyzed is the distribution of dialogues per movie. This distribution is shown in Figure 3. As seen from the figure, the distribution of dialogues per movie is clearly symmetric around its mean value of 175 dialogues per movie. For most of the movies in the collection, a number of dialogues ranging from about 100 to 250 were extracted.

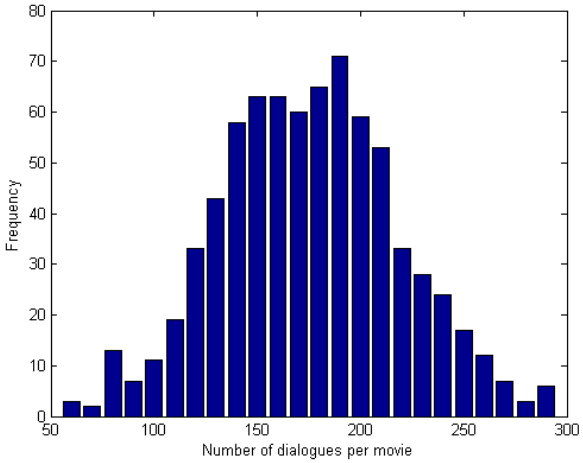

Figure 3: Distribution of dialogues per movie

The second property of the corpus to be studied is the distribution of turns per dialogue. This distribution is shown in Figure 4. As seen from the figure, this distribution approximates a power law behavior, with a large number of very short dialogues (about 50K two-turn dialogues) and a small amount of long dialogues (only six dialogues with more than 200 turns). The median of the distribution is 5.63 turns per dialogue.

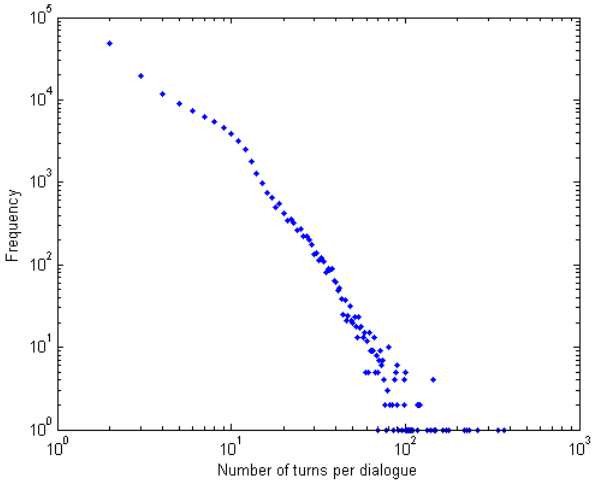

Figure 4: Distribution of turns per dialogue

The third property of the corpus to be described is the distribution of number of speakers per dia-

logue. This distribution is shown in Figure 5. As seen from the bar-plot depicted in the figure, the largest proportion of dialogues (around 60K) involves two speakers. The second largest proportion of "dialogues" (about 35K) involves only a single speaker, which means that this subset of the data collection is actually composed by monologues or single speaker interventions. The third and fourth larger proportions are those involving three and four speakers, respectively.

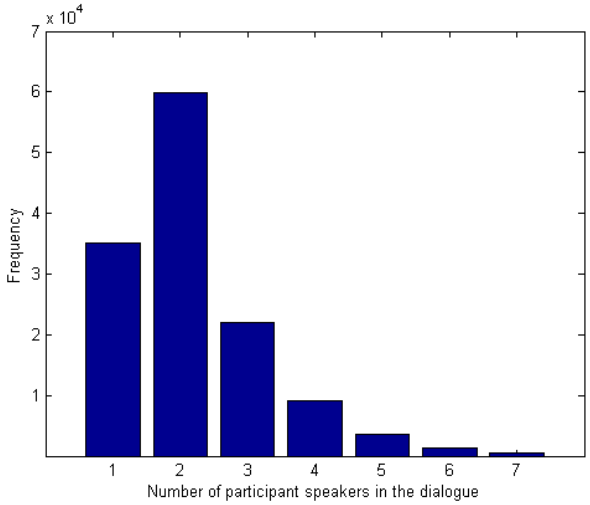

Figure 5: Distribution of number of speakers per dialogue

Finally, in Figure 6, we present a cross-plot between the number of dialogues and the number of turns within each movie script.

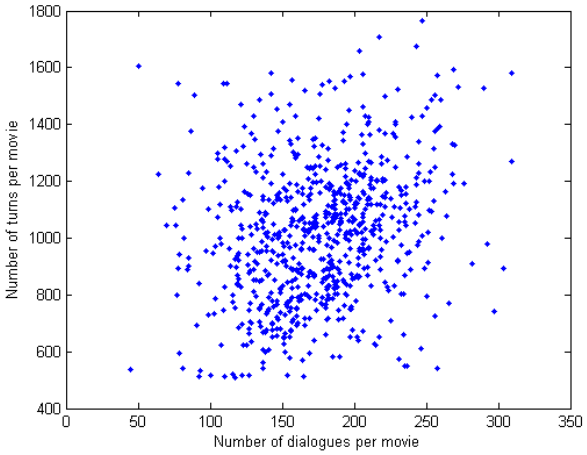

Figure 6: Cross-plot between the number of dialogues and turns within each movie script

As seen from the cross-plot, an average movie has between 150 and 200 dialogues comprising between 1000 and 1200 turns in total. The cross-plot also reveals some interesting extreme cases in the data collection.

For instance, movies with few dialogues but many turns are located towards the upper-left corner of the figure. In this zone we can find movies as: *Happy Birthday Wanda June*, *Hannah and Her Sisters* and *All About Eve*. In the lower-left corner of the figure we can find movies with few dialogues and few turns, as for instance: *1492 Conquest of Paradise* and *The Cooler*.

In the right side of the figure we find the lots-of-dialogues region. There we can find movies with lots of very short dialogues (lower-right corner), such as *Jimmy and Judy* and *Walking Tall*; or movies with lots of dialogues and turns (upper-right corner), such as *The Curious Case of Benjamin Button* and *Jennifer's Body*.

4 Conclusions and Future Work

In this paper, we have described Movie-DiC a Movie Dialogue Corpus that has been collected for research and development purposes. The data collection comprises 132,229 dialogues containing a total of 764,146 turns/utterances that have been extracted from 753 movies. Details on how the data collection has been created and how the corpus is structured were provided along with the main statistics and characteristics of the corpus.

Although strictly speaking, and by its particular nature, Movie-DiC does not constitute a corpus of real human-to-human dialogues, it does constitute an excellent dataset for studying the semantic and pragmatic aspects of human communication within a wide variety of contexts, scenarios, styles and socio-cultural settings.

Specific technologies and applications that can exploit a resource like this include, but are not restricted to: example-based chat bots (Banchs and Li, 2012), question answering systems, discourse and pragmatics analysis, narrative vs. colloquial style classification, genre classification, etc.

As future work, we intend to expand the current size of the collection from 0.7K to 2K movies, as well as to improve some of our parsing and post-processing algorithms for reducing the amount of noise still present in the collection and enhance the quality of the current version of the dataset.

Acknowledgments

The author would like to thank the Institute for Infocomm Research for its support and permission to publish this work.

References

Banchs R E, Li H (2012) IRIS: a chat-oriented dialogue system based on the vector space model. In Proceedings of the 50th Annual Meeting of the ACL, demo session.

Brown P, Della Pietra S, Della Pietra V, Mercer R (1993) The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19(2):263-311.

Bunt H (ed) (2000) Abduction, belief, and context in dialogue: studies in computational pragmatics. J. Benjamins.

Busemann S, Declerck T, Diagne A, Dini L, Klein J, Schmeier S (1997) Natural language dialogue service for appointment scheduling agents. In Proceedings of the 5th Conference on Applied Natural Language Processing, pp 25-32.

Mann W (2003) The Dialogue Diversity Corpus. Accessed online on 16 March 2012 from: http://www-bcf.usc.edu/~billmann/diversity/DDivers-site.htm

Misu T, Ohtake K, Hori C, Kashioka H, Nakamura S (2009) Annotating communicative function and semantic content in dialogue act for construction of consulting dialogue systems. In Proceedings of the Int. Conf. of Spoken Language Processing

Molla-Aliod D, Vicedo J (2010) Question answering. In Indurkhya and Damerau (eds) Handbook of Natural Language Processing, pp 485-510. Chapman & Hall.

Qin T, Liu T, Zhang X, Wang D, Xiong W, Li H (2008) Learning to rank relational objects and its application to Web search. In Proceedings of the 17th International Conference on World Wide Web, pp 407-416.

Rieser V, Lemon O (2011) Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation. Springer.

Stallard D (2000) Talk'n'travel: a conversational system for air travel planning. In Proceedings of the 6th Conference on Applied Natural Language Processing, pp 68-75.

Wallis P (2010) A robot in the kitchen. In Proceedings of the ACL 2010 Workshop on Companionable Dialogue Systems, pp 25-30.

Weizenbaum J (1966) ELIZA – A computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1):36-45.

Williams J, Young S (2003) Using Wizard-of-Oz simulations to bootstrap Reinforcement-Learning-based dialog management systems. In Proceedings of the 4th SIGDIAL Workshop on Discourse and Dialogue.

Zue V (2007) On organic interfaces. In Proceedings of the International Conference of Spoken Language Processing.