CONVERSATION DIALOG CORPORA FROM TELEVISION AND MOVIE SCRIPTS

Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
E-mail: {*lasguido.kp9,ssakti,neubig,tomoki,s-nakamura*}*@is.naist.jp*

ABSTRACT

Example-based dialogue systems often require natural conversation templates as examples for response generation. However, in previous work most conversation corpora have been created by hand and do not well portray actual conversations between two people. One way to overcome this problem is to record and transcribe real human-to-human conversation. However, this work is tedious and time consuming. In this work, we utilize conversation scripts from television and movies. We extract conversations from television and movie scripts from the web and perform various types of filtering. In order to ensure that the conversation is performed by two speakers, we introduce a unit of conversation called a tri-turn (a trigram conversation turn) which allow us to filter conversations with more than two speakers. In the end, our conversation corpora contains 86,719 query-response pairs that represent conversation turns performed by two speakers talking to each other[1].

Index Terms— dialog corpora, tri-turn unit, dialog system resource

1. INTRODUCTION

An essential component of social life is conversation, one of the main ways by which humans can gain access to the content of others minds [1]. Through the advance of technology in many aspect of our daily lives, the issue of communication via natural and spontaneous speech between human beings and machines is also becoming more important [2].

Heretofore, natural language dialog systems mostly focused on two main dialogue genres: (1) specific-task dialog (for instance ATIS flight reservation [3], DARPA Communicator dialog travel planning [4]) and (2) non-specific-task dialog (for instance chatterbot systems like Eliza [5] or Alice [6]). Moreover a dialog systems can also be described by the amount of human intervention used in their construction, ranging from entirely hand-made to completely data-driven. Seminal work often limited interactions to a specific scenario (e.g. a Rogerian psychotherapist [5]) or were based on complex, knowledge-rich rule-based systems for generating responses, which required large amounts of human effort to create or add new rules [6].

Data-driven approach for deploying a dialog system are becoming popular as a lightweight alternative to create broad-coverage chat-oriented dialog systems [7, 8, 9, 10]. Data-driven approaches generally use dialog examples that are semantically indexed to a database. Proper responses for every user input are generated based on these dialog examples. In particular, the data used in these systems usually consists of query-response pairs, where the query is representative of the user's input to the system, and the response is representative of the system's response.

As a result, to achieve a good coverage on various types of natural conversation, recording of a large data set of real human-to-human conversation is necessary, which is tedious and time consuming. A generic solution to address this problem is using handmade scripted dialog scenarios, which may lead to unnatural conversation. Another approach is constructing dialog examples from available log databases, for instance a conversation between human subjects and a Wizard of OZ in WOZ system [11], or human-to-human text conversation in Twitter[2] [12].

Regardless, it is still difficult to cover all possible patterns that may exist in real human-to-human conversation. Currently, most data-driven approach systems rely on either canned responses by providing error messages [13] or templates for generation when a reponse cannot be found, sometimes resulting in a completely incomprehensible response [14]. Furthermore, given recent success of machine translation in various NLP tasks [15, 16, 17], some other work investigates machine translation as an approach for response generation [10, 18].

In previous work [18], we proposed a method to utilize human-to-human conversation examples from drama television and movies script data. In this paper, we provide a full description of the collection methodology of the language re-

[1]We will released and update gradually our corpus at `http://isw3.naist.jp/~lasguido-l/me/resources.html#dialog-conversation-pair`.

[2]http://twitter.com/

sources used in the previous work, and provide a fuller analysis of the content contained therein. The aim of our work is to provide a dialog-pair (query-response) corpora of human-to-human conversation. In order to ensure the content extracted from raw drama television and movie script files consist of appropriate dialog-pair examples, we use a unit of conversation called a tri-turn, and semantic similarity filtering based on matches over WordNet synsets.

In detail, we can break down the process of dialog corpus construction into two main steps: (1) raw data collection and preprocessing, which removes unnecessary information and normalizes the text, and (2) dialog-pair extraction and semantic similarity filtering, which ensures the conversation is done between two people talking to each other. The explanation about raw data collection and preprocessing is available in Section 3 and we describe the process of dialog-pair extraction and semantic similarity filtering in Section 4.

2. RELATED WORK

There have been several related works in developing conversation dialog corpora from movie scripts. Walker, et. al. [19] collect and publish a corpus of movie script conversations from the The Internet Movie Script Database website[3]. In this work, they annotate the conversation script by various features (for instance: speaker gender, movie genre, word length, and sentiment strength). On the other hand, [20] introduce IRIS (Informal Response Interactive System), a chat-oriented dialog system based on movie script databases. However, this system did not filter any uncorrelated consecutive utterances from the conversation movie script database, which is a central point of our work. As mentioned by the author, this causes failures and diminishes the ability to maintain a consistent conversation.

Compare to the previous work, we attempt to perform filtering to select appropriate utterances that represent real conversation between two people. In addition, we also perform tri-turn and semantic similarity filtering to maintain the integrity of the corpus.

3. RAW DATA COLLECTION

In the first step in our corpus collection, we construct a dialog corpora from raw HTML gathered on the web. Preprocessing of the movie scripts is done by transforming raw HTML files into easily readable text format.

A film script is a conversational manuscript that depicts the conversations and actions between actors in a movie. The detailed process of raw data collection can be seen in the Figure 1. Figure 1(a) illustrates an example of one movie scene with four actors talking to each other. The corresponding raw

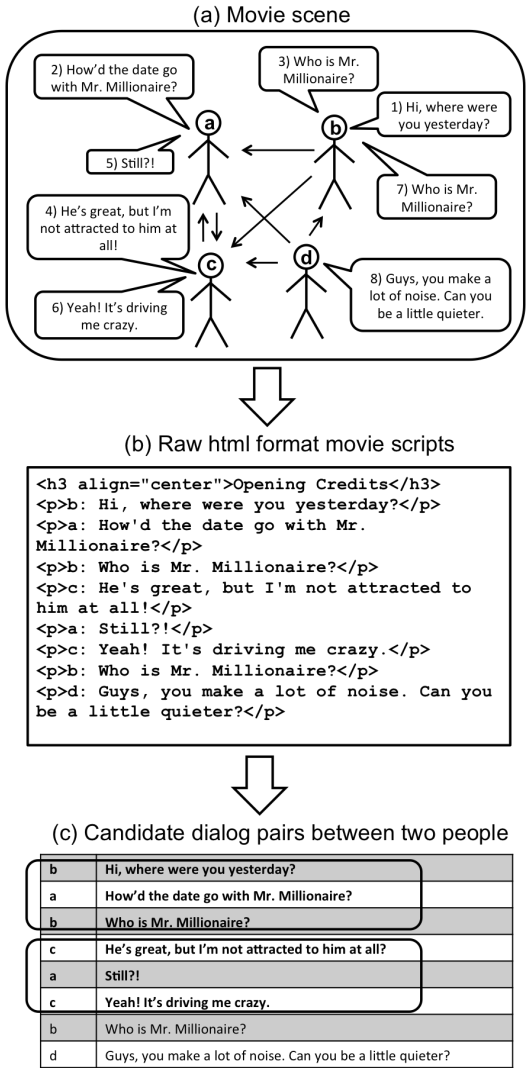


Fig. 1. Dialog corpora construction from movie scripts.

movie scripts that are available from the web are usually written in HTML files shown in Figure 1(b). The dialogues between actors are arranged in chronological order. Since we use a variety of sources of movie scripts that have various formats, we implemented several parsing algorithms to fetch the information from the raw movie conversation. We also remove unnecessary explanatory information about the movie scenes.

Moreover, we also define two basic types of information about each dialog: actors and utterances. The utterances are the actual content of each dialog turn in the movie scripts. The actor refers to the character name in the movies. This actor and utterance information will be utilized to construct a dialog corpus.

[3]http://www.imsdb.com/

4. DIALOG PAIR EXTRACTION AND SEMANTIC SIMILARITY FILTERING

As a result of the process in the previous section, conversational dialogues contained in movie scenes do not have a clear indication of which utterances are responses to a particular utterance. Therefore, it is important to find a solution that is able to construct appropriate dialog-pair examples from raw movie script files. To ensure that the dialog example database contains only query-response pairs, we utilize two methods for selection of the dialog data: trigram turn sequences, or *tri-turns* and semantic similarity filtering [18].

4.1. Tri-turn Filtering

A tri-turn is defined as three turns in a conversation between two actors X and Y that has the pattern X-Y-X (As shown in Figure 1(c)). In other words, within a tri-turn the first and last dialog turn are performed by the same actor and the second dialog turn is performed by the other actor. Next the query-response pairs are made by separating the tri-turn pattern X-Y-X into two pairs, X-Y and Y-X.

We found that when we observed this pattern, in the great majority of the cases this indicated that the first and second utterances (X-Y pair), as well as the second and third utterances (Y-X pair), formed a proper input-response pair as shown in the c-a-c tri-turn in Figure 2.

4.2. Semantic Similarity Filtering

However even after tri-turn filtering, noisy cases which contain uncorrelated turns still exist (see the first two utterances of the b-a-b tri-turn in Figure 2), this happens because the speakers are not actually speaking to each-other. To address this problem, we perform further filtering using the semantic similarity measure described as follows.

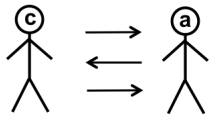
Semantic similarity (similar to the approach introduced in [21]) is used to ensure a semantic relationship between each dialog turn in the dialog-pair data. As shown in Equation (1), the similarity is computed between WordNet[4] synsets in each dialog turn.

$$sem_{sim}(S_1, S_2) = \frac{2 \times |S_{syn1} \cap S_{syn2}|}{|S_{syn1}| + |S_{syn2}|} \quad (1)$$

S_{syn1} and S_{syn2} respectively are groups of WordNet synsets for each word in the sentences S_1 and S_2 that are linked by a complex network of lexical relations. The similarity of sentence pair X-Y where S_1 = X and S_2 = Y can be obtained by calculating the relations between S_{syn1} and S_{syn2}. $|S_{syn1} \cap S_{syn2}|$ is the number of co-occurring WordNet synsets and $|S_{syn1}| + |S_{syn2}|$ is a total number of effective WordNet synsets. Dialog pairs with high similarity are then extracted and included into the database.

[4]http://wordnet.princeton.edu/

Actor	Correlated tri-turn
c	He's great, but I'm not attracted to him at all!
a	Still?!
c	Yeah! It's driving me crazy.



Actor	Un-correlated tri-turn
b	Hi, where were you yesterday?
a	How'd the date go with Mr. Millionaire?
b	Who is Mr. Millionaire?

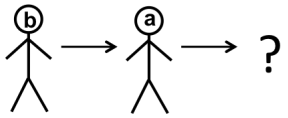


Fig. 2. Example of a tri-turn with two actors.

5. DATA ANALYSIS

We obtain our raw data from the Friends TV show[5], The Internet Movie Script Database[6], and The Daily Script[7]. Parsing the raw HTML data is done with the Perl CPAN HTML-Parser[8] and the filtering system is built in the Python environment using the Python NLTK tools[9].

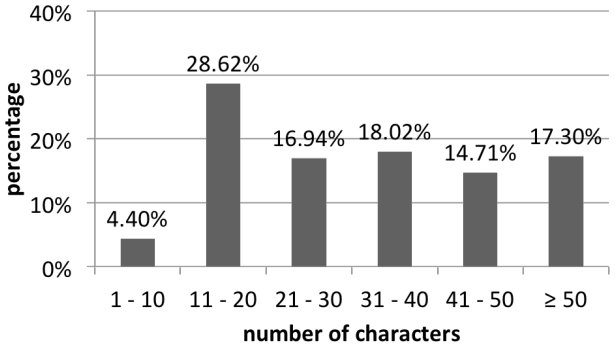


Fig. 3. Percentage of total characters involved in one movie.

From the raw data, 28.62% of the collected movie scripts are played by 11 - 20 different characters. Only 4.40% collected movie scripts are played by 1 - 10 different charac-

[5]//ufwebsite.tripod.com/scripts/scripts.htm
[6]http://imsdb.com/
[7]http://dailyscript.com/
[8]http://search.cpan.org/dist/HTML-Parser/Parser.pm
[9]http://nltk.org

Head JSON	
turn_1	(Turn) first Turn in the tri-turn
turn_2	(Turn) second Turn in the tri-turn
turn_3	(Turn) third Turn in the tri-turn
syntax_distance_1	(double) represents the syntactic distance between turn_1 and turn_2
semantic_distance_1	(double) represents the semantic distance between turn_1 and turn_2
syntax_distance_2	(double) represents the syntactic distance between turn_2 and turn_3
semantic_distance_2	(double) represents the semantic distance between turn_2 and turn_3
Turn	
actor	the name who performed the dialog turn
sentence	the actual tokenized sentence
actual_sentence	the actual sentence after the filtering process (not tokenized)
postag	POS information of the sentence
ner	NE information of the sentence
dependency_grammar	normalized dependency trees of the sentence
semantic_set	WordNet synsets of the sentence
additional_info	additional non-dialog information (e.g. narration, actor expression, and actor emotion)
original_sentence	the original sentence from the script
turn_in_file	the sentence turn in the file
script_filename	the script filename

Table 2. Description of data in JSON format.

conversation scripts	1,786
dialog pairs	1,042,288
dialog pairs after filtering	86,719

Table 1. Conversation corpus details.

ters. Besides the main characters, the movie scripts are usually composed by cameos (e.g. *"a man in the radio," "man 1," "radio"*). These cameo characters contribute to increase the character variation in a single movie. Figure 3 shows in detail the total number of different characters involved in one movie.

The Friends TV show scripts are written in English and contain 5 seasons, with a total of 112 episodes. Each episode contains several scenes and each scene contains several dialog turns. The total number of scenes in the corpus are 1,437. The movie script data is from The Internet Movie Script Database and The Daily Script, captured June 2012. This resulted in a total of 1,786 conversation scripts with 1,042,288 dialog pairs. After performing dialog turn extraction and semantic similarity filtering, the total number of dialog pairs is 86,719. The summary of conversation corpora can be seen in the Table 1.

Additionally, we annotate every sentence in the dialog turn with the labels such as part-of-speech tags (POS), named entities (NE), and dependency trees. POS, NE, and dependencies are tagged by using the Python NLTK Brown corpus

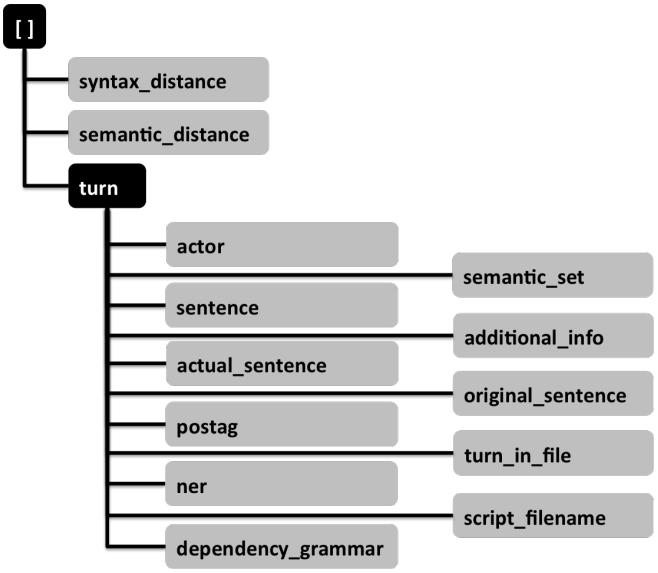


Fig. 4. Dialog structure in JSON format.

POS Tagger, Stanford NER[10], and the Stanford dependency parser[11]. We also add semantic and syntactic similarity distance between two sentences in the dialog pairs. The syntactic similarity distance obtained by calculating a syntactic similarity measure [21], given the dependency tree of a sentence as an input. Finally, we wrap each dialog-pair with all of its annotation in JSON[12] data format. The description of JSON

[10]http://nlp.stanford.edu/software/CRF-NER.shtml
[11]http://nlp.stanford.edu/software/stanford-dependencies.shtml
[12]http://json.org/

data format is shown in Figure 4 and Table 2.

6. CONCLUSION

In this work, we have created a new dialog-pair corpus from television and movie scripts. We gathered various source of conversation dialog from Friends TV Shows, The IMSDB, and The Daily Script. As future work, we plan to explore text conversation in Twitter. In some cases Twitter may contain a conversational text between two persons talking to each other. However, there are still possible challenges such as dealing with the nonstandard words or emoticons.

7. REFERENCES

[1] R. M. Krauss and C. Y. Chiu, "Language and social behavior," in *Handbook of social psychology*, vol. 2. McGraw-Hill, Boston, 1997.

[2] John Holmes and Wendy Holmes, *Speech Synthesis and Recognition*, Taylor & Francis, Inc., Bristol, PA, USA, 2nd edition, 2001.

[3] E. Seneff, L. Hirschman, and V. Zue, "Interactive problem solving and dialogue in the ATIS domain," in *Proc. of the Fourth DARPA Speech and Natural Language Workshop*, 1991, pp. 354–359.

[4] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA communicator dialog travel planning systems: the June 2000 data collection," in *Proc. of EUROSPEECH*, 2000, pp. 1371–1374.

[5] J. Weizenbaum, "Eliza - a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[6] R.Wallace, *Be Your Own Botmaster*, A.L.I.C.E A.I. Foundation, 2003.

[7] S. Jung, C. Lee, and G.G. Lee, "Dialog studio: An example based spoken dialog system development workbench," in *Proc. of the Dialogs on dialog: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems. Interspeech 2006-ICSLP satellite workshop*, Pittsburgh, USA, 2006.

[8] C. Lee, S. Lee, S. Jung, K. Kim, D. Lee, and G.G. Lee, "Correlation-based query relaxation for example-based dialog modeling," in *Proc. of ASRU*, Merano, Italy, 2009, pp. 474–478.

[9] K. Kim, C. Lee, D. Lee, J. Choi, S. Jung, and G.G. Lee, "Modeling confirmations for example-based dialog management," in *Proc. of SLT*, Berkeley, California, USA, 2010, pp. 324–329.

[10] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proc. of EMNLP*, Edinburgh, Scotland, UK., July 2011, pp. 583–593, ACL.

[11] H. Murao, N. Kawaguchi., S. Matsubara, Y. Yamaguchi, and Y. Inagaki, "Example-based spoken dialogue system using WOZ system log," in *Proc. of SIGDIAL*, Saporo, Japan, 2003, pp. 140–148.

[12] F. Bessho, T. Harada, and Y. Kuniyoshi, "Dialog system using real-time crowdsourcing and Twitter large-scale corpus," in *Proc. of SIGDIAL*, Seoul, South Korea, 2012, pp. 227–231.

[13] C. Lee, S. Jung, S. Kim, and G. G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," *Speech Commun.*, vol. 51, no. 5, pp. 466–484, May 2009.

[14] N. Chambers and J. Allen, "Stochastic language generation in a dialogue system: Toward a domain independent generator.," in *Proc. of SIGDIAL*, Cambridge, Massachusetts, USA, 2004, pp. 9–18.

[15] A. Echihabi and D. Marcu, "A noisy-channel approach to question answering," in *Proc. of ACL - Volume 1*, Sapporo, Japan, 2003, ACL '03, pp. 16–23.

[16] Y. W. Wong and R. J. Mooney, "Learning for semantic parsing with statistical machine translation," in *Proc. of HLT-NAACL*, New York, NY, USA, 2006, pp. 439–446.

[17] Y. W. Wong and R. J. Mooney, "Generation by inverting a semantic parser that uses statistical machine translation.," in *Proc. of HLT-NAACL*, Rochester, NY, USA, 2007, pp. 172–179.

[18] L. Nio, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Combination of example-based and smt-based approaches in a chat-oriented dialog system," in *Proc. of ICE-ID*, Bali, Indonesia, 2013, IEEE.

[19] M. Walker, G. Lin, and J. Sawyer, "An annotated corpus of film dialogue for learning and characterizing character style," in *Proc. of LREC*, Istanbul, Turkey, May 2012.

[20] R. E. Banchs and H Li, "IRIS: a chat-oriented dialogue system based on the vector space model," in *ACL (System Demonstrations)*, 2012, pp. 37–42.

[21] D. Liu, Z. Liu, and Q. Dong, "A dependency grammar and wordnet based sentence similarity measure," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 1027–1035, 2012.