Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 358–361
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

A Counselling Corpus in Cantonese

John Lee, Tianyuan Cai, Wenxiu Xie, Lam Xing
Department of Linguistics and Translation
City University of Hong Kong
{jsylee, tianycai, wenxixie, lamxing}@cityu.edu.hk

Abstract

Virtual agents are increasingly used for delivering health information in general, and mental health assistance in particular. This paper presents a corpus designed for training a virtual counsellor in Cantonese, a variety of Chinese. The corpus consists of a domain-independent subcorpus that supports small talk for rapport building with users, and a domain-specific subcorpus that provides material for a particular area of counselling. The former consists of ELIZA style responses, chitchat expressions, and a dataset of general dialog, all of which are reusable across counselling domains. The latter consists of example user inputs and appropriate chatbot replies relevant to the specific domain. In a case study, we created a chatbot with a domain-specific subcorpus that addressed 25 issues in test anxiety, with 436 inputs solicited from native speakers of Cantonese and 150 chatbot replies harvested from mental health websites. Preliminary evaluations show that Word Mover's Distance achieved 56% accuracy in identifying the issue in user input, outperforming a number of baselines.

Keywords: Cantonese, chatbot, counselling, test anxiety

1. Introduction

Virtual agents are increasingly used for delivering health information in general, and mental health assistance in particular. While chatbots may not qualify to replace human counsellors, users have been found to be more willing to disclose information to a chatbot than to a human (Lucas et al., 2014). Woebot, for example, has been shown to be effective in reducing symptoms of depression (Fitzpatrick et al., 2017).

In the "guided conversation" format, virtual counsellors ask pre-determined questions and allow users to choose from suggested responses (Fitzpatrick et al., 2017; Casas et al., 2018). Other chatbots are designed to handle free-form input from users. PAL (Liu et al., 2013) and TeenChat (Huang et al., 2015) are two examples among those operating in Chinese. Targeting teenagers and young adults, these two chatbots offer advice on social topics such as family relations and love affairs. In response to user input, it selects the most relevant answer from a knowledge base, taking into account user characteristics such as gender, marital status and age.

Since no single virtual counsellor can adequately address all mental health issues or serve all user populations, counselling services at schools or mental health organizations may be interested in developing their own bots to provide tailored advice in specialized counselling domains. While existing chatbots contain valuable linguistic and counselling resources, there is often no straightforward way to re-use or adapt them for related counselling tasks.

To facilitate development of virtual counsellors, this paper presents a corpus that is designed to be modifiable and extendable by any counsellor, who will henceforth be referred to as the "administrator". Designed for chatbots in the example-based framework, it contains a *domain-independent* subcorpus with small talk materials for rapport building with users, an essential component in counselling sessions across all domains. It also has a *domain-specific* subcorpus, which is to be populated by the administrator with example user inputs and chatbot replies for the domain concerned.

We developed the corpus in Cantonese, which is considered the "most widely known and influential variety of Chinese other than Mandarin" (Matthews and Yip, 2011). Less frequently used in formal written communication than Mandarin, the dominant variety in mainland China, Cantonese is supported by relatively few resources for natural language processing. Although Cantonese and Mandarin are genetically related, having both developed from Middle Chinese, they are mutually unintelligible in their spoken form and have significant differences in their written form (Wong and Lee, 2018). The differences between the major Chinese varieties have been described as being "at least on the order of the different languages of the Romance family" (Hannas, 1997).

The rest of the paper is organized as follows. After an overview of corpus design in the next section, we describe the content and role of the domain-independent subcorpus (Section 3) and the domain-specific subcorpus (Section 4).[1] We then present a case study on constructing a chatbot for test anxiety (Section 5).

2. Corpus Design

Our corpus, which consists of post-reply sentence pairs, is designed to support the construction of virtual counsellors in any domain. Given a user input, a chatbot can identify the example post in the corpus that is semantically most similar, and then retrieve the corresponding reply.

According to the five-stage framework for counselling sessions adopted by Inoue et al. (2012), the Introduction stage establishes rapport between the therapist and the client. In the Elaboration stage, the therapist explores the client's situation and tries to find clues for a solution. This may be followed by a Resistance stage if the client feels uncomfortable or resists the therapy. The therapist then proposes actions toward a solution in the Intervention stage. Finally, the Solution stage concludes the session.

[1] The corpus is available for research purposes on request to the first author.

Category	Description	Example user input	Example chatbot reply
Chitchat	Short social expressions	你好 'hi' 唔該 'thanks' 拜拜 'goodbye'	你好 'hi' 唔洗客氣 'you're welcome' 得閒再搵我啦 'see you later'
	Small talk	你叫咩名？ 'What's your name?'	我叫做... 'My name is ...'
Dialog	General dialog	你的工作是什麼？ 'What's your job?'	陪聊 'Talking to people.'
Encouragers	ELIZA expressions	我覺得 ... 'I feel ...'	點解覺得 ...? 'Why do you feel ... ?'
Advice requests	Explicit request for advice	我想聽下你既建議 'I want to hear your advice'	(To be drawn from the domain-specific subcorpus)

Table 1: Domain-independent subcorpus (Section 3): example post-reply pairs

Category	# post-reply pairs
Chitchat	123
Advice requests	41
Encouragers	103
Dialog	276,000

Table 2: Size of the domain-independent subcorpus (Section 3), with breakdown into the main categories

The *domain-independent* subcorpus (Section 3) provides content for "off-task," social exchanges that are common across many counselling domains, especially during the Introduction stage. In subsequent stages, as users discuss their specific issues, the chatbot is expected to detect the issue types and incorporate appropriate advice in its response. For this purpose, the *domain-specific* subcorpus (Section 4) is to be populated by the administrator with example user inputs and possible advice to address them.

3. Domain-independent Subcorpus

The domain-independent subcorpus targets off-task interactions, such as greetings and chitchat. These interactions are typically dominant in the Introduction stage of a counselling session, but they can also be interweaved with on-task interactions in the subsequent stages. Their purpose is to move the conversation along, while keeping the users engaged and preparing them to disclose their feelings and issues.

Several example post-reply pairs in this subcorpus are shown in Table 1. Expected to be re-usable for most counselling domains and target users, these pairs belong to one of the four categories below. Table 2 shows the size of the domain-independent subcorpus with a breakdown into these categories.

3.1. Chitchat

The chitchat category covers common user expressions such as greetings, thank-you and good-bye, to which formulaic responses from the chatbot is usually sufficient; short yes/no answers (e.g., 係呀 'yes') to the preceding question; as well as small talk such as user enquiries about the chatbot's personal information, including name, age, occupation, family members, favorite food or colors, etc. The chatbot's persona is the only content that is expected to be customized in this subcorpus.

3.2. Encouragers

Encouragers consist of backchannel or empathic replies, mostly defined with ELIZA-like regular expressions (Weizenbaum, 1983).

3.3. Advice Requests

Advice requests are user inputs that explicitly ask the chatbot for counselling advice.

3.4. Dialog

If the user input does not match any of the above categories, the chatbot can backoff to a large-scale dialog database. We compiled a set of over 276,000 Cantonese and Chinese post-reply pairs, to handle general or off-task user input. These pairs are taken from two conversation corpora: the *Xiaohuangji* corpus[2] and the ChatterBot corpus[3].

4. Domain-specific Subcorpus

Following the Introduction stage of the counselling session, the chatbot engages users in discussing their thoughts and experiences, and then gives counselling advice. The domain-specific subcorpus consists of example user inputs and appropriate pieces of advice. Unlike those in the domain-independent subcorpus (Section 3), these post-reply pairs are less likely to be relevant to other domains. The onus is therefore put on the administrators to supply these pairs for their target counselling domain.

Most existing virtual counsellors clearly define their areas of competence. For example, PAL addresses the topics "husband and wife", "family relations", "love affairs", "adolescence", "feeling and mood", and "mental tutors" (Liu et al., 2013), while TeenChat detect stress in the areas of "study", "self-cognition", "interpersonal", "affection" and "general" (Huang et al., 2015). The domain-specific subcorpus likewise expects the administrator to define a set of issues, but at a finer granularity, such that each issue can be mapped to specific pieces of advice.

4.1. Symptom Statements

We will refer to each issue as a "symptom". In the domain of test anxiety, for example, symptoms may include headaches, worries about failure, and worries about parental reaction, which are among the most relevant issues identified in a psychology study (Wren and Benson, 2004).

[2]Accessed at https://github.com/skdjfla/dgk_lost_conv

[3]Accessed at https://github.com/gunthercox/chatterbot-corpus

Symptom	Symptom statement	Counselling item
Headache	我頭疼 'I have a headache'	原來係咁, 會唔會係得唔夠? 'I see, did you get enough sleep?'
Worries about failure	如果唔合格咁點算啊？ 'What if I fail?'	依家擔心都冇用順其自然啦 'No need to worry now, let it be!'
Worries about parental reaction	又會被屋企人話 'I'll get critized at home'	只要你盡左力父母一定會明白既 "If you tried hard, your parents would understand'
Severe	想死 'I want to die'	你依家可以打比x 8478搵counsellor傾下先！ 'Call x8478 now to chat with a counsellor'

Table 3: Domain-specific corpus (Section 4): example symptoms, symptom statements and counselling items for the test anxiety domain

For each symptom, the administrator is to provide a set of symptom statements, i.e., typical user inputs that express that symptom. Table 3 shows example symptom statements for the domain of test anxiety (Wren and Benson, 2004).

4.2. Counselling Items

The chatbot should be able to address each symptom with appropriate pieces of advice, which we will refer to as "counselling items". The administrator is to provide these items, which may be practical advice collected from existing mental health materials. They can also be contact information of human counsellors, which would be appropriate for more severe symptoms such as thoughts of suicide.

The administrator can optionally provide a set of sentences for broaching the counselling topic[4], to be deployed when users engage exclusively in small talk for an extended period, to steer them back to on-task discussion.

4.3. Threshold Score

The administrator may adjust the aggressiveness of the chatbot in advice giving by setting the threshold score for symptom detection. The typical system computes a similarity score between the user input and each symptom statement in the subcorpus (Section 5.2). If the score exceeds the threshold, a counselling item for the symptom is returned. A higher threshold leads to a chatbot that is more restrained in giving advice and more inclined to listen. A lower threshold, in contrast, increases its propensity to give counsel.

5. Case Study

To evaluate the proposed chatbot framework, we compiled a domain-specific subcorpus for the domain of test anxiety (Section 5.1) and experimented with a number of semantic similarity measures (Section 5.2). Together with the domain-independent subcorpus, we instantiated a chatbot that addresses test anxiety (Figure 1).

Many websites and pamphlets already offer advice and remedies for test anxiety, an issue that affects many students. Despite the wealth of counselling materials static media, however, there has been little attempt to leverage them for virtual counselling. To the best of our knowledge, the only reported chatbot that specifically deals with test

[4]e.g., 你有冇關於學業既野想同我傾下？ 'Do you want to talk about any academic issue?'

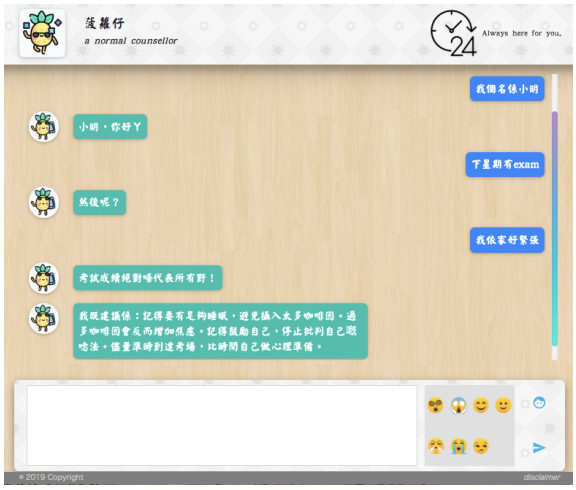

Figure 1: Example conversation in our Cantonese chatbot for test anxiety

anxiety is the *Exam-Stress Counselor and Academic Planner* (Rudra et al., 2012). This system uses a deterministic finite state automaton to analyze user inputs, which are expected to conform to a number of fixed patterns. Our chatbot, in contrast, is designed for interactions with free-form text.

5.1. Compilation of Domain-specific Subcorpus

Our set of symptoms were the 25 most relevant issues in test anxiety identified by Wren and Benson (2004). For symptom statements, we solicited user inputs for each symptom from 12 subjects, all native speakers of Cantonese. After reading the original symptom description in English written by Wren and Benson (2004), the subjects were asked to give paraphrases in Cantonese. They provided a total of 436 sentences, which served as the evaluation dataset.

For counselling items, we collected advice on test anxiety from websites linked to the home page of the counselling department at our university. For severe symptoms requiring human intervention, we supplied the phone number of the department. The subcorpus contained a total of 150 counselling items, with an average of 5.8 items per symptom.

5.2. Semantic Similarity Measures

To determine the symptom in the user input, the chatbot computes semantic similarity scores between the input and

Approach	Accuracy
Word Mover's Distance	**56.2%**
ELMo	47.5%
word2vec	47.7%

Table 4: Symptom detection accuracy

the symptom statements in the domain-specific subcorpus. We compared the performance of two semantic similarity measures with the baseline of using word2vec (Mikolov et al., 2013) embeddings alone.

Word Mover's Distance (WMD). We applied this distance metric by representing each word in the sentence with the word2vec vector embeddings. We then used WMD to measure the dissimilarity between two sentences, as expressed by the minimum cumulative distance that the embedded words of one sentence need to travel to match those in the other sentence (Kusner et al., 2015). The symptom statement that yields the shortest distance is returned.

ELMo. Shown to improve the state-of-the-art in a variety of NLP tasks, ELMo is a contextualized word representation that models not only the characteristics of word usage, but also how the usage varies across linguistic contexts, i.e., polysemy (Peters et al., 2018). The word vectors are learned functions of the internal states of a deep bidirectional language model, pre-trained on a large text corpus.

5.3. Symptom Detection Accuracy

We conducted a leave-one-out evaluation on the symptom statements in our domain-specific subcorpus (Section 5.1) to measure accuracy in symptom detection. The chatbot automatically determined the symptom for each of the 436 statements, by retrieving the most similar statement among those in the remainder of the subcorpus, with respect to the similarity measures described in Section 5.2. As shown in Table 4, Word Mover's Distance achieved the best performance in symptom detection, at 56.2% accuracy. It outperformed both ELMo (47.5%) and the word2vec baseline (47.7%).

6. Conclusions

We have presented the first text corpus that is designed for virtual counselling in Cantonese. The corpus consists of a domain-independent subcorpus of chitchat and general dialog materials, which are intended to be re-usable for any counselling domain; and a domain-specific subcorpus, to be populated by administrators with typical user inputs and counselling advice in the domain concerned. We reported a case study on test anxiety, and evaluated the chatbot's accuracy in symptom detection. It is hoped that this corpus will facilitate development of virtual counsellors to serve Cantonese speakers.

7. Acknowledgements

This work was partially supported by CityU Internal Funds for ITF Projects (no. 9678104).

8. Bibliographical References

Casas, J., Mugellini, E., and Khaled, O. A. (2018). Food Diary Coaching Chatbot. In *Proc. ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp)*, pages 1676–1680.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2):e19.

Hannas, W. C. (1997). *Asia's Orthographic Dilemma*. University of Hawaii Press, Honolulu, HI.

Huang, J., Li, Q., Xue, Y., Cheng, T., Xu, S., Jia, J., and Feng, L. (2015). TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress. *LNCS*, 9085:133–145.

Inoue, M., Hanada, R., Furuyama, N., Irino, T., Ichinomiya, T., and Massaki, H. (2012). Multimodal Corpus for Psychotherapeutic Situation. In *Proc. LREC Workshop on Multimodal Corpora for Machine Learning*, pages 18–21.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *Proc. International Conference on Machine Learning*, pages 957–966.

Liu, Y., Liu, M., Wang, X., Wang, L., and Li, J. (2013). PAL: A Chatterbot System for Answering Domain-specific Questions. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 67–72.

Lucas, G. M., Gratch, J., King, A., and Morency, L. P. (2014). It's Only a Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior*, 37:94–100.

Matthews, S. and Yip, V. (2011). *Cantonese: A Comprehensive Grammar*. Routledge, New York.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proc. International Conference on Learning Representations (ICLR)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proc. NAACL-HLT*.

Rudra, T., Li, M., and Kavakli, M. (2012). ESCAP: Towards the Design of an AI Architecture for a Virtual Counselor to Tackle Students' Exam Stress. In *Proc. 45th Hawaii International Conference on System Sciences*.

Weizenbaum, J. (1983). ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 26(1):23–28.

Wong, T.-S. and Lee, J. (2018). Register-sensitive Translation: A Case Study of Mandarin and Cantonese. In *Proc. Association for Machine Translation in the Americas (AMTA)*.

Wren, D. G. and Benson, J. (2004). Measuring Test Anxiety in Children: Scale Development and Internal Construct Validation. *Anxiety, Stress, & Coping: An international Journal*, 17(3):227–240.