Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy

Lina Rojas-Barahona[1], **Bo-Hsiang Tseng**[1], **Yinpei Dai**[1], **Clare Mansfield**[2]
Osman Ramadan[1], **Stefan Ultes**[1], **Michael Crawford** [3] **and**
Milica Gašić[1]
[1] University of Cambridge, [2] CM Insight, [3] Imperial College London
mg436@cam.ac.uk

Abstract

In recent years, we have seen deep learning and distributed representations of words and sentences make impact on a number of natural language processing tasks, such as similarity, entailment and sentiment analysis. Here we introduce a new task: understanding of mental health concepts derived from Cognitive Behavioural Therapy (CBT). We define a mental health ontology based on the CBT principles, annotate a large corpus where this phenomena is exhibited and perform understanding using deep learning and distributed representations. Our results show that the performance of deep learning models combined with word embeddings or sentence embeddings significantly outperform non-deep-learning models in this difficult task. This understanding module will be an essential component of a statistical dialogue system delivering therapy.

1 Introduction

Promotion of mental well-being is at the core of the action plan on mental health 2013–2020 of the World Health Organisation (WHO) (World Health Organization, 2013) and of the European Pact on Mental Health and Well-being of the European Union (EU high-level conference: Together for Mental Health and Well-being, 2008). The biggest potential breakthrough in fighting mental illness would lie in finding tools for early detection and preventive intervention (Insel and Scholnick, 2006). The WHO action plan stresses the importance of health policies and programmes that not only meet the need of people affected by mental disorders but also protect mental well-being. The emphasis is on early evidence-based non-pharmacological intervention, avoiding institutionalisation and medicalisation. What is particularly important for successful intervention is the frequency with which the therapy can be accessed (Hansen et al., 2002). This gives automated systems a huge advantage over conventional therapies, as they can be used continuously with marginal extra cost. Health assistants that can deliver therapy, have gained great interest in recent years (Bickmore et al., 2005; Fitzpatrick et al., 2017). These systems however are largely based on hand-crafted rules. On the other hand, the main research effort in statistical approaches to conversational systems has focused on limited-domain information seeking dialogues (Schatzmann et al., 2006; Geist and Pietquin, 2011; Gasic and Young, 2014; Fatemi et al., 2016; Li et al., 2016; Williams et al., 2017).

In this paper we introduce a new task: understanding of mental health concepts derived from Cognitive Behavioural Therapy (CBT). We present an ontology that is formulated according to Cognitive Behavioural Therapy principles. We label a high quality mental health corpus, which exhibits targeted psychological phenomena. We use the whole unlabelled dataset to train distributed representations of words and sentences. We then investigate two approaches for classifying the user input according to the defined ontology. The first model involves a convolutional neural network (CNN) operating over distributed words representations. The second involves a gated recurrent network (GRU) operating over distributed representation of sentences. Our models perform significantly better than chance and for instances with a large number of data they reach the inter-annotator agreement. This understanding module will be an essential component of a statistical dialogue system delivering therapy.

The paper is organised as follows. In Section 2 we give a brief background of the statistical approach to dialogue modelling, focusing on dialogue ontology and natural language understanding. In Section 3 we review related work in the area of automated mental-health assistants. The sections that

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 44–54
Brussels, Belgium, October 31, 2018.

follow represent the main contribution of this work: a CBT ontology in Section 4, a labelled dataset in Section 5, and models for language understanding in Section 6. We present the results in Section 7 and our conclusion in Section 8.

2 Background

A dialogue system can be treated as a trainable statistical model suitable for goal-oriented information seeking dialogues (Young, 2002). In these dialogues, the user has a clear goal that he or she is trying to achieve and this involves extracting particular information from a back-end database. A structured representation of the database, the *ontology* is a central element of a dialogue system. It defines the concepts that the dialogue system can understand and talk about. Another critical component is the natural language understanding unit, which takes textual user input and detects presence of the ontology concepts in the text.

2.1 Dialogue ontology

Statistical approaches to dialogue modelling have been applied to relatively simple domains. These systems interface databases of up to 1000 entities where each entity has up to 20 properties, i.e. *slots* (Cuayáhuitl, 2009). There has been a significant amount of work in spoken language understanding focused on exploiting large knowledge graphs in order to improve coverage (Tür et al., 2012; Heck et al., 2013). Despite these efforts, little work has been done on mental health ontologies for supporting cognitive behavioural therapy on dialogue systems. Available medical ontologies follow a symptom-treatment categorisation and are not suitable for dialogue or natural language understanding (Bluhm, 2017; Hofmann, 2014; Wang et al., 2018).

2.2 Natural language understanding

Within a dialogue system, a natural language understanding unit extracts meaning from user sentences. Both classification (Mairesse et al., 2009) and sequence-to-sequence (Yao et al., 2014; Mesnil et al., 2015) models have been applied to address this task.

Deep learning architectures that exploit distributed word-vector representations have been successfully applied to different tasks in natural language understanding, such as semantic role labelling, semantic parsing, spoken language understanding, sentiment analysis or dialogue belief tracking (Collobert et al., 2011; Kim, 2014; Kalchbrenner et al., 2014; Le and Mikolov, 2014a; Rojas Barahona et al., 2016; Mrkšić et al., 2017).

In this work we consider understanding of mental health concepts of as a classification task. To facilitate this process, we use distributed representations.

3 Related work

The aim of building an automated therapist has been around since the first time researchers attempted to build a dialogue system (Weizenbaum, 1966). Automated health advice systems built to date typically rely on expert coded rules and have limited conversational capabilities (Rojas-Barahona and Giorgino, 2009; Vardoulakis et al., 2012; Ring et al., 2013; Riccardi, 2014; DeVault et al., 2014; Ring et al., 2016). One particular system that we would like to highlight is an affectively aware virtual therapist (Ring et al., 2016). This system is based on Cognitive Behavioural Therapy and the system behaviour is scripted using VoiceXML. There is no language understanding: the agent simply asks questions and the user selects answers from a given list. The agent is however able to interpret hand gestures, posture shifts, and facial expressions. Another notable system (DeVault et al., 2014) has a multi-modal perception unit which captures and analyses user behaviour for both behavioural understanding and interaction. The measurements contribute to the indicator analysis of affect, gesture, emotion and engagement. Again, no statistical language understanding takes place and the behaviour of the system is scripted. The system does not provide therapy to the user but is rather a tool that can support healthcare decisions (by human healthcare professionals).

The Stanford Woebot chat-bot proposed by (Fitzpatrick et al., 2017) is designed for delivering CBT to young adults with depression and anxiety. It has been shown that the interaction with this chat-bot can significantly reduce the symptoms of depression when compared to a group of people directed to a read a CBT manual. The conversational agent appears to be effective in engaging the users. However, the understanding component of Woebot has not been fully described. The dialogue decisions are based on decision trees. At each node, the user is expected to choose one of several predefined responses. Limited language understanding was in-

troduced at specific points in the tree to determine routing to subsequent conversational nodes. Still, one of the main deficiencies reported by the trial participants in (Fitzpatrick et al., 2017) was the inability to converse naturally. Here we address this problem by performing statistical natural language understanding.

4 CBT ontology

To define the ontology we draw from principles of Cognitive Behavioural Therapy (CBT). This is one of the best studied psychotherapeutic interventions, and the most widely used psychological treatment for mental disorders in Britain (Bhasi et al., 2013). There is evidence that CBT is more effective than other forms of psychotherapy (Tolin, 2010). Unlike other, longer-term, forms of therapy such as psychoanalysis, CBT can have a positive effect on the client within a few sessions. Also, due to it being highly structured, it is more easily amenable by computer interpretation. This is why we adopted CBT as the basis of our work.

Cognitive Behavioural Therapy is derived from Cognitive Therapy model theory (Beck, 1976; Beck et al., 1979), which postulates that our emotions and behaviour are influenced by the way we think and by how we make sense of the world. The idea is that, if the client changes the way he or she thinks about their problem, this will in turn change the way he or she feels, and behaves.

A major underlying principle of CBT is the idea of cognitive distortions, and the value in challenging them. In CBT, clients are helped to test their assumptions and views of the world in order to check if they fit with reality. When clients learn that their perceptions and interpretations are distorted or unhelpful they then work on correcting them. Within the realm of cognitive distortion, CBT identifies a number of specific self-defeating thought processes, or thinking errors. There is a core of around 10 to 15 thinking errors, with their exact titles having some fluidity. A strong component of CBT is teaching clients to be able to recognize and identify the thinking errors themselves, and ultimately discard the negative thought processes and 're-think' their problems.

We consider the main analytical step in this therapy: an adequate decoding of these 'thinking error' concepts, and the identification of the key emotion(s) and the situational context of a particular problem. Therefore, our ontology consists of *thinking errors*, *emotions*, and *situations*.

4.1 Thinking errors

Notwithstanding slight variations in number and terminology, the list of *thinking errors* is fairly well standardised in the CBT literature. We present one such list in Table 1. However, it is important to note that there is a fair degree of overlap between different *thinking errors*, for example, between *Jumping to Negative Conclusions* and *Fortune Telling*, or between *Disqualifying the Positives* and *Mental Filtering*. In addition, within the data used – and as is likely to be the case in any data of spontaneous expressions of psychological upset – a single problem can exhibit several *thinking errors* simultaneously. Thus, the situation is much more challenging than in simple information-seeking dialogues, where ontologies are typically clearly defined and there is no or very little overlap between concepts.

4.2 Emotions

In addition to *thinking errors*, we define a set of *emotions*. We mainly focus on negative emotions, relevant to people in psychological distress. In CBT, emotions tend to be divided into positive and negative, or helpful/healthy and unhelpful/unhealthy emotions (Branch and Willson, 2010). The set of emotions for this work evolved over time in the early days of annotation. Although we initially agreed to focus on 'unhealthy' emotions, as defined by CBT, there seemed also to be a place for the 'healthy' emotion *Grief/sadness*. Overall, the list of emotions used was drawn from a number of sources, including CBT literature, the annotators' own knowledge of what they work with in psychological therapy, and the common emotions that were seen emerging from the data early on in the process. Note that more than one emotion might be expressed within an individual problem – for example *Depression* and *Loneliness*. The list of *emotions* is given in Table 2.

4.3 Situations

While our main emphasis was on *thinking errors* and *emotions*, we also defined a small set of *situations*. The list of *situations* again evolved during the early days of annotation, with a longer original list being reduced down, for simplicity. Again, it is possible for more than one situation (for example *Work* and *Relationships*) to apply to a single problem. The considered *situations* are given in Table 3.

Thinking Error	Frequency	Exhibited by...
Black and white (or all or nothing) thinking	20.82%	Only seeing things in absolutes No shades of grey
Blaming	8.05%	Holding others responsible for your pain Not seeking to understand your own responsibility in situation
Catastrophising	11.87%	Magnifying a (sometimes minor) negative event into potential disaster
Comparing	3.27%	Making dissatisfied comparison of self versus others
Disqualifying the positive	6.15%	Dismissing/discounting positive aspects of a situation or experience
Emotional reasoning	13.31%	Assuming feelings represent fact.
Fortune telling	25.70%	Predicting how things will be, unduly negatively
Jumping to negative conclusions	44.16%	Anticipating something will turn out badly, with little evidence to support it
Labelling	10.51%	Using negative, sometimes highly coloured, language to describe self or other Ignoring complexity of people
Low frustration tolerance "I can't bear it"	16.03%	Assuming something is intolerable, rather than difficult to tolerate or a temporary discomfort
Inflexibility "should/need/ought"	8.08%	Having rigid beliefs about how things or people must or ought to be
Mental filtering	5.50%	Focusing on the negative Filtering out all positive aspects of a situation
Mind-reading	14.60%	Assuming others think negative things or have negative motives and intentions
Over-generalising	12.69%	Generalising negatively, using words like always, nobody, never, etc
Personalising	5.85%	Interpreting events as being related to you personally and overlooking other factors

Table 1: Taxonomy for *thinking errors* and how they are exhibited.

Emotion	Frequency	Exhibited by ...
Anger (/frustration)	14.76%	Feelings of frustration, annoyance, irritation, resentment, fury, outrage
Anxiety	63.12%	Any expression of fear, worry or anxiety
Depression	20.72%	Feeling down, hopeless, joyless, negative about self and/or life in general
Grief/sadness	5.70%	Feeling sad, upset, bereft in relation to a major loss
Guilt	3.37%	Feeling blameworthy for a wrongdoing or something not done
Hurt	19.88%	Feeling wounded and/or badly treated
Jealousy	3.12%	Antagonistic feeling towards another either wish to be like or to have what they have
Loneliness	7.41%	Feeling of alone-ness, isolation, friendlessness, not understood by anyone
Shame	5.68%	Feeling distress, humiliation, disgrace in relation to own behaviour or feelings

Table 2: Taxonomy for *emotions* and how they are exhibited.

Situation	Frequency
Bereavement	2.65%
Existential	21.93%
Health	10.61%
Relationships	67.58%
School/College	8.28%
Work	6.10%
Other	5.53%

Table 3: Taxonomy for *situations*.

5 The corpus

The corpus consists of 500K written posts that users anonymously posted on the Koko platform[1]. This platform is based on the peer-to-peer therapy proposed by (Morris et al., 2015). In this set-up, a user anonymously posts their problem (referred to as the *problem*) and is prompted to consider their most negative take on the problem (referred to as the *negative take*). Subsequently, peers post responses that attempt to offer a re-think and give a more positive angle on the problem. When first developed, this peer-to-peer framework was shown to be more efficacious than expressive writing, an intervention that is known to improve physical and

[1] https://itskoko.com/

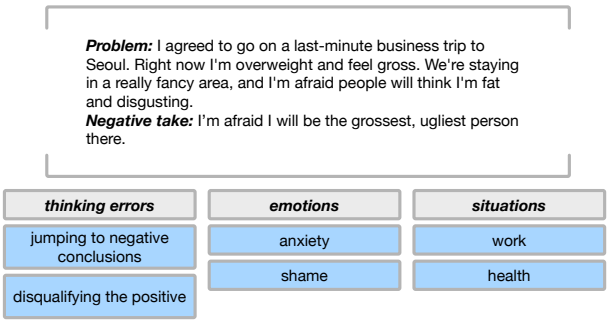

Figure 1: An example of an annotated Koko post.

emotional well-being (Morris et al., 2015). Since then, the app developed by Koko has collected a very large number of posts and associated responses. Initially, any first-time Koko user would be given a short introductory tutorial in the art of 're-thinking'/'re-framing' problems (based on CBT principles), before being able to use the platform. This however changed over time, as the age of the users decreased, and a different tutorial, emphasizing empathy and optimism, was used (less CBT-based than the 're-thinking'). Most of the data annotated in this study was drawn from the earlier phase. Figure 1 gives an annotated post example.

5.1 Annotation

A subset of posts was annotated by two psychological therapists using a web annotation tool that we developed. The annotation tool allowed annotators to have a quick view of the posts, showing up to 50 posts per page, to navigate through posts, to check pending posts and to annotate them by adding or removing *thinking errors*, *emotions* and *situations*. All annotations were stored in a MySQL database.

Initially 1000 posts were analysed. These were used to define the ontology. Then 4035 posts were labelled with *thinking errors*, *emotions* and *situations*. It takes an experienced psychological therapist about one minute to annotate one post. Note that the same post can exhibit multiple *thinking errors*, *emotions* and *situations*, which makes the whole process more complex. We randomly selected 50 posts and calculated the inter-annotator agreement. The inter-annotator agreement was calculated using a contingency table for thinking error, emotion and situation, showing agreement and disagreement between the two annotators. Then, Cohen's kappa was calculated discounting the possibility that the agreement may happen by chance. The result is shown in Table 4. The main reason for the low agreement in *thinking errors* (61%) is due to the unbounded number of *thinking errors* per post. In other words, the annotators typically have three or four *thinking errors* in common but one of them might have detected one or two more. Still, the agreement is much higher than chance, so we think that while challenging, it is possible to build a classifier for this task. The distributions of labelled posts with multiple sub-categories for three super-categories are shown in Figure 2

Concept	Thinking error	Situation	Emotion
Kappa	0.61 ± 0.09	0.92 ± 0.08	0.90 ± 0.07

Table 4: Cohen's kappa with a 95% confidence interval

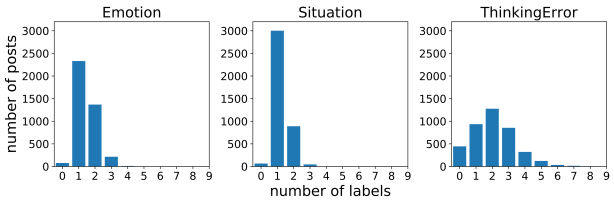

Figure 2: Distribution of posts for each category.

6 Deep learning model

6.1 Distributed representations

The task of decoding *thinking errors* and *emotions* is closely related to the task of sentiment analysis. In sentiment analysis we are concerned with positive or negative sentiment expressed in a sentence. Detecting thinking errors or emotions could be perceived as detecting different kinds of negative sentiment. Distributed representations of words, sentences and documents have gained success in sentiment detection and similarity tasks (Le and Mikolov, 2014a; Maas et al., 2011; Kiros et al., 2015). A key advantage of these representations is that they can be obtained in an unsupervised manner, thus allowing exploitation of large amounts of unlabelled data. This is precisely what we have in our set-up, where only a small portion of our posts is labelled.

We utilise GloVe (Pennington et al., 2014) word vectors, which have previously achieved competitive results in a similarity task. We train the word vectors on the whole dataset and then use a convolutional neural network (CNN) to extract features from posts where words are represented as vectors.

We also consider distributed representation of sentences. A particularly competitive model is the skip-thought model, which is obtained from an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage

(Kiros et al., 2015). On similarity tasks it outperfoms the simpler doc2vec model (Le and Mikolov, 2014a). An approach that represents vectors by weighted averages of word vectors and then modifies them using PCA and SVD outperforms skip-thought vectors (Arora et al., 2017). This method however does not do well on a sentiment analysis task due to down-weighting of words like "not". As these often appear in our corpus, we chose skip-thought vectors for investigation here.

The skip-thought model allows a dense representation of the utterance. We train skip-thought vectors using the method described in (Kiros et al., 2015). The automatically generated post shown in Fig 3 demonstrates that skip-thought vectors can convey the sentiment well in accordance to context. We then train a gated recurrent unit (GRU) network using the skip-thoughts as input.

i 'm so depressed . i 'm worthless . No one likes me i 'm try being nice but . No light at every point i 'm unpopular and i 'm a <NUM> year old potato . my most negative take is that i 'll never know how to be as socially as a quiet girl. i will stop talking to how fragile is and be any ways of normal people .

Figure 3: An example of a generated post using skip-thought vectors initialised with "I'm so depressed".

6.2 Convolutional neural network model

The convolutional neural network (CNN) model is preferred over a recurrent neural network (RNN) model, because the posts are generally too long for an RNN to maintain memory over words. The convolutional neural network (CNN) used in this work is inspired by (Kim, 2014) and operates over pre-trained GloVe embeddings of dimensionality d. As shown in Fig 4, the network has two inputs, one for the *problem* and the other for the *negative take*. These are represented as two tensors. A convolutional operation involves a filter $\mathbf{w} \in \mathcal{R}^{ld}$ which is applied to l words to produce the feature map. Then, a max-pooling operation is applied to produce two vectors: $\mathbf{p}$ for *problem* and $\mathbf{n}$ for *negative take*. The reason for this is that the *negative take* is usually a summary of the post, carrying stronger sentiment (see Figure 1). We use a gating mechanism to combine $\mathbf{p}$ and $\mathbf{n}$ as follows:

$$\mathbf{g} = \sigma(\mathbf{W}_p\mathbf{p} + \mathbf{W}_n\mathbf{n} + \mathbf{b}) \quad (1)$$

$$\mathbf{h} = \mathbf{g} \odot \mathbf{p} + (\mathbf{1} - \mathbf{g}) \odot \mathbf{n} \quad (2)$$

Here, σ is the sigmoid function, $\mathbf{W}_p$, $\mathbf{W}_n$ and $\mathbf{W}$ are weight matrices, $\mathbf{b}$ is a bias term, $\mathbf{1}$ is a vector of ones, $\odot$ is the element-wise product, and $\mathbf{g}$ is the output of the gating mechanism. The extracted feature $\mathbf{h}$ is then processed with a one-layer fully-connected neural network (FNN) to perform binary classification. The model is illustrated in Fig 4.

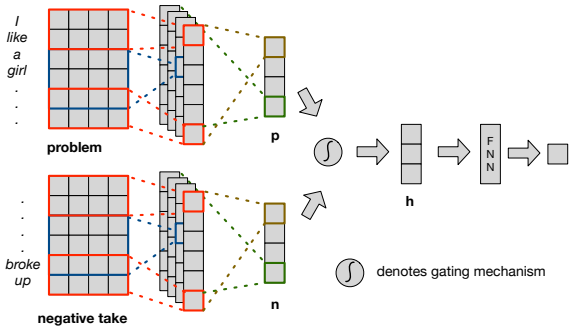

Figure 4: CNN with gating mechanism.

6.3 Gated recurrent unit model

We use the gated recurrent unit (GRU) model to process skip-thought sentence vectors, for two reasons. First, most posts contain less than 5 sentences, so a recurrent neural network is more suitable than a convolutional neural network. Second, since our corpus only comprises very limited labelled data, a GRU should perform better than a long short-term memory (LSTM) network as it has less parameters.

Denote each post as $P = \{s_1, s_2, ..., s_t, ...\}$, where s_t is the t^{th} sentence in post P. First, we use an already trained GRU to extract skip-thought embeddings $\mathbf{e_t}$ from the sentences s_t. Then, taking the sequence of sentence vectors $\{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_t}, ...\}$ as input, another GRU is used as follows:

$$\mathbf{z_t} = \sigma(\mathbf{W}_z\mathbf{h}_{t-1} + \mathbf{U}_z\mathbf{e_t} + \mathbf{b_z}) \quad (3)$$

$$\mathbf{r_t} = \sigma(\mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{U}_r\mathbf{e_t} + \mathbf{b_r}) \quad (4)$$

$$\tilde{\mathbf{h}}_\mathbf{t} = \tanh(\mathbf{W}(\mathbf{r_t} \odot \mathbf{h}_{t-1}) + \mathbf{U}\mathbf{e_t} + \mathbf{b_h}) \quad (5)$$

$$\mathbf{h_t} = \mathbf{z_t} \odot \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{z_t}) \odot \tilde{\mathbf{h}}_\mathbf{t} \quad (6)$$

$\mathbf{W}_z, \mathbf{U}_z, \mathbf{W}_r, \mathbf{U}_r, \mathbf{W}, \mathbf{U}$ are recurrent weight matrices, $\mathbf{b_z}, \mathbf{b_r}, \mathbf{b_h}$ are bias terms, $\odot$ is the element-wise dot product, and σ is the sigmoid function.

Finally, the last hidden state $\mathbf{h}_T$ is fed into a FNN with one hidden layer of the same size as input. The model is illustrated in Fig 5.

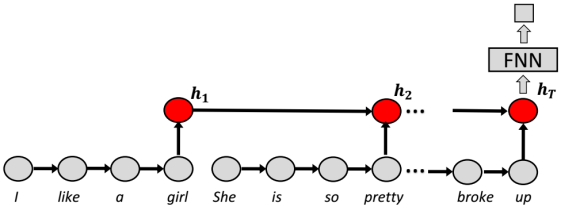

Figure 5: GRU with skip-thought vectors.

6.4 Training set-up

We first train 100 and 300 dimensions for both GloVe embeddings and skip-thought embeddings using the same mechanism as in (Pennington et al., 2014; Kiros et al., 2015). In some posts the length of sentences is very large, so we bound the length at 50 words. We do not treat the *problem* separately from the *negative take* as the GRU will anyway put more importance on the information that comes last. We split the labelled data in a $8:1:1$ ratio for training, validation and testing in a 10-fold cross validation for both GRU and CNN training. A distinct network is trained for each concept, i. e. one for *thinking errors*, one for *emotions* and one for *situations*. The hidden size of the FNN is 150.

To tackle the data bias problem, we utilise oversampling. Different ratios (1:1, 1:3, 1:5, 1:7) of positive and negative samples are explored.

We used filter windows of 2, 3, and 4 with 50 feature maps for the CNN model. For the GRU model, the hidden size is set at 150, so that both models have comparable number of parameters. Mini-batches of size 24 are used and gradients are clipped with maximum norm 5. We initialise the learning rate as 0.001 with a decay rate of 0.986 every 10 steps. The non-recurrent weights with a truncated normal distribution $(0, 0.01)$, and the recurrent weights with orthogonal initialisation (Saxe et al., 2013). To overcome over-fitting, we employ dropout with rate 0.8 and $l2$-normalisation. Both models were trained with Adam algorithm and implemented in Tensorflow (Girija, 2016).

7 Results

7.1 Baselines

For rule-based models, we chose a chance classifier and a majority classifier, where all the posts are treated as positive examples for each class. In addition, we trained two non-deep-learning models, the logistic regression (LR) model and the Support Vector Machine (SVM). Both of them take the bag-of-words feature as input and implemented in sklearn (Pedregosa et al., 2011). For completeness, we also trained 100 and 300 dimensions PV-DM document embeddings (Le and Mikolov, 2014b) as the distributed representations of the posts using the *gensim* toolkit (Řehůřek and Sojka, 2010), and employ FNNs to do the classification, the hidden size is set as 800 to ensure parameters of all deep learning models comparable. All the baseline models are trained with the same set-up as described in section 6.4.

7.2 Analysis

Table 5 gives the average F1 scores and the average F1 scores weighted with the frequency of CBT labels for all models under the oversampling ratio 1:1. It shows that GloVe word vectors with CNN achieves the best performance both in 100 and 300 dimensions.

Model	AVG. F1	Weighted AVG F1
Chance	0.203±0.008	0.337±0.008
Majority	0.24±0.000	0.432±0.000
LR-BOW	0.330±0.011	0.479±0.008
SVM-BOW	0.403±0.000	0.536±0.000
FNN-DocVec-100d	0.339±0.006	0.502±0.005
FNN-DocVec-300d	0.349±0.007	0.508±0.005
GRU-SkipThought-100d	0.401±0.005	0.558±0.004
GRU-SkipThought-300d	0.423±0.005	0.570±0.004
CNN-GloVe-100d	$\mathbf{0.443 \pm 0.007}$	0.576±0.005
CNN-GloVe-300d	0.442 ± 0.007	$\mathbf{0.578 \pm 0.006}$

Table 5: F1 scores for all models with 1:1 oversampling

Table 6 shows the F1-measure of the compared models that detect *thinking errors*, *emotions* and *situations* under the $1:1$ oversampling ratio. We only include the results of the best performing models, SVMs, CNNs and GRUs, due to limited space. The results show that both models outperform SVM-BOW in larger embedding dimensions. Although SVM-BOW is comparable to 100 dimensional GRU-Skip-thought in terms on average F1, in all other cases CNN-GloVe and GRU-Skip-thought overshadow SVM-BOW. We also find that CNN-GloVe on average works better than GRU-Skip-thought, which is expected as the space of words is smaller in comparison to the space of sentences so the word vectors can be more accurately trained. While the CNN operating on 100 dimensional word vectors is comparable to the CNN operating on 300 dimensional word vectors, the GRU-Skip-thought tends to be worse on 100 dimensional skip-thoughts, suggesting that sentence vectors generally need to be of a higher dimension to represent the meaning more accurately than word vectors.

Table 7 shows a more detailed analysis of the 300 dimensional CNN-GloVe performance, where both precision and recall are presented, indicating that oversampling mechanism can help overcome the data bias problem. To illustrate the capabilities of this model, we give samples of two posts and their predicted and true labels in Figure 6, which shows that our model discerns the classes reasonably well even in some difficult cases.

	Freq. Num.	SVM-BOW	100d		300d	
			CNN-Glove	GRU-Skip-thought	CNN-Glove	GRU-Skip-thought
Emotion						
Anxiety	2547	0.798±0.000	0.805±0.003	0.805±0.002	0.805±0.006	**0.816 ± 0.002**
Depression	836	0.564±0.000	0.605±0.003	0.568±0.001	**0.611 ± 0.008**	0.578±0.005
Hurt	802	0.448±0.000	0.505±0.007	0.483±0.003	**0.506 ± 0.005**	0.496±0.006
Anger	595	0.375±0.001	0.389±0.009	0.384±0.007	0.383±0.004	**0.425 ± 0.007**
Loneliness	299	0.558±0.000	0.495±0.008	0.445±0.007	**0.549 ± 0.009**	0.457±0.005
Grief	230	0.433±0.005	0.462±0.010	0.373±0.008	**0.462 ± 0.008**	0.382±0.005
Shame	229	0.220±0.000	**0.304 ± 0.011**	0.243±0.004	0.277±0.007	0.254±0.004
Jealousy	126	0.217±0.000	**0.228 ± 0.012**	0.159±0.004	0.216±0.005	0.216±0.009
Guilt	136	0.252±0.000	**0.295 ± 0.012**	0.186±0.007	0.279±0.014	0.225±0.008
AVG. F1 score for **Emotion**		0.429±0.001	**0.454 ± 0.008**	0.405±0.005	**0.454 ± 0.007**	0.428±0.006
Situation						
Relationships	2727	0.861±0.000	0.871±0.003	0.886±0.001	0.878±0.006	**0.889 ± 0.003**
Existential	885	0.556±0.000	0.591±0.002	**0.600 ± 0.005**	0.594±0.007	0.599±0.006
Health	428	0.476±0.000	**0.589 ± 0.003**	0.555±0.005	0.585±0.008	0.587±0.006
School_College	334	0.633±0.000	0.670±0.004	0.641±0.003	0.673±0.009	**0.680 ± 0.002**
Other	223	0.196±0.001	0.255±0.011	0.241±0.008	0.256±0.005	**0.281 ± 0.006**
Work	246	0.651±0.000	**0.663 ± 0.004**	0.572±0.006	0.661±0.011	0.639±0.006
Bereavement	107	0.602±0.000	0.637±0.021	0.402±0.024	**0.639 ± 0.021**	0.493±0.011
AVG. F1 score for **Situation**		0.568±0.000	0.611±0.007	0.557±0.007	**0.612 ± 0.010**	0.595±0.006
Thinking Error						
Jumping_to_negative_conclusions	1782	0.590±0.000	0.696±0.004	0.685±0.004	**0.703 ± 0.005**	0.687±0.002
Fortune_telling	1037	0.458±0.000	**0.595 ± 0.002**	0.558±0.004	0.585±0.006	0.564±0.005
Black_and_white	840	0.395±0.000	0.431±0.002	0.437±0.004	0.432±0.003	**0.441 ± 0.003**
Low_frustration_tolerance	647	0.318±0.000	0.322±0.007	0.330±0.003	0.313±0.005	**0.336 ± 0.001**
Catastrophising	479	0.352±0.000	**0.375 ± 0.002**	0.358±0.005	0.371±0.004	0.364±0.003
Mind-reading	589	0.360±0.000	0.404±0.005	0.353±0.011	**0.419 ± 0.006**	0.356±0.007
Labelling	424	0.399±0.001	0.453±0.007	0.335±0.004	**0.462 ± 0.004**	0.373±0.002
Emotional_reasoning	537	0.290±0.000	**0.319 ± 0.007**	0.285±0.005	0.306±0.006	0.293±0.008
Over-generalising	512	0.405±0.001	0.405±0.006	0.375±0.004	**0.418 ± 0.008**	0.389±0.004
Inflexibility	326	0.202±0.001	0.203±0.014	0.188±0.007	**0.218 ± 0.003**	0.208±0.005
Blaming	325	0.209±0.001	**0.304 ± 0.007**	0.264±0.002	0.277±0.003	0.274±0.004
Disqualifying_the_positive	248	0.146±0.000	0.194±0.007	0.176±0.005	0.187±0.003	**0.195 ± 0.005**
Mental_filtering	222	0.088±0.000	0.142±0.007	0.150±0.001	0.141±0.002	**0.155 ± 0.003**
Personalising	236	0.212±0.000	0.230±0.012	0.220±0.005	**0.236 ± 0.004**	0.221±0.005
Comparing	132	0.242±0.000	**0.289 ± 0.014**	0.177±0.008	0.255±0.009	0.227±0.007
AVG. F1 score for **Thinking Error**		0.311±0.000	**0.358 ± 0.007**	0.326±0.005	0.355±0.0050	0.339±0.004
AVG. F1 score		0.403±0.000	**0.443±0.007**	0.401±0.005	0.442±0.007	0.423±0.005
AVG. F1 score weighted with Freq.		0.536±0.000	0.576±0.005	0.558±0.004	**0.578±0.006**	0.570±0.004

Table 6: F1 score of the models trained with embeddings with dimensionality of 300 and 100 respectively.

Problem: I have lots of work to be done. I got a new phone and waste lots of time in it. I feel stressed because of my phone.
Negative take: I cheat my mom and waste my time.

Thinking Errors	*Emotions*	*Situations*
True labels		
- Low frustration tolerance	*- Anger*	*- Relationships* *- Existential* *- Work*
Predicted labels		
- Black and white thinking *- Disqualifying the positive*	*- Anger* *- Anxiety*	

Problem: I really miss my ex! She just stopped talking to me all of the sudden, I don't know what happened.
Negative take: I won't get to talk to her again ...

Thinking Errors	*Emotions*	*Situations*
True labels		
- Jumping to negative conclusions *- Low frustration tolerance*	*- Hurt* *- Grief*	*- Relationships*
Predicted labels		
- Jumping to negative conclusions *- Low frustration tolerance*	*- Hurt* *- Grief*	*- Relationships*

Figure 6: predictions of posts by 300 dim CNN-GloVe

Figure 7 gives the comparative performance of two models under different oversampling ratios. While oversampling is essential for both models, GRU-Skip-thought is less sensitive to lower oversampling ratios, suggesting that skip-thoughts can already capture sentiment on the sentence level. Therefore, including only a limited ratio of positive samples is sufficient to train the classifier. Instead, models using word vectors need more positive data to learn sentence sentiment features.

8 Conclusion

We presented an ontology based on the principles of Cognitive Behavioural Therapy. We then annotated data that exhibits psychological problems and computed the inter-annotator agreement.

We found that classifying *thinking errors* is a difficult task as suggested by the low inter-annotator agreement. We trained GloVe word embeddings and skip-thought embeddings on 500K posts in an unsupervised fashion and generated distributed representations both of words and of sentences. We

label	precision	recall	F1 score	accuracy
Anxiety	0.739±0.007	0.884±0.005	0.805±0.006	0.729±0.012
Depression	0.538±0.010	0.708±0.005	0.611±0.008	0.813±0.010
Hurt	0.428±0.005	0.620±0.004	0.506±0.005	0.763±0.011
Anger	0.313±0.005	0.491±0.000	0.383±0.004	0.769±0.012
Loneliness	0.479±0.010	0.643±0.008	0.549±0.009	0.923±0.006
Grief	0.437±0.013	0.490±0.000	0.462±0.008	0.937±0.005
Shame	0.219±0.008	0.378±0.004	0.277±0.007	0.891±0.007
Jealousy	0.170±0.002	0.296±0.012	0.216±0.005	0.935±0.006
Guilt	0.221±0.014	0.378±0.008	0.279±0.014	0.936±0.008
Relationships	0.847±0.005	0.912±0.007	0.878±0.006	0.829±0.011
Existential	0.516±0.008	0.700±0.004	0.594±0.007	0.789±0.009
Health	0.520±0.010	0.668±0.005	0.585±0.008	0.900±0.006
School_College	0.570±0.009	0.821±0.008	0.673±0.009	0.934±0.004
Other	0.209±0.004	0.331±0.007	0.256±0.005	0.894±0.007
Work	0.601±0.015	0.733±0.006	0.661±0.011	0.955±0.003
Bereavement	0.567±0.029	0.733±0.008	0.639±0.021	0.979±0.002
Jumping_to_negative_conclusions	0.643±0.005	0.775±0.004	0.703±0.005	0.711±0.009
Fortune_telling	0.486±0.006	0.737±0.004	0.585±0.006	0.733±0.010
Black_and_white	0.330±0.003	0.625±0.003	0.432±0.003	0.657±0.011
Low_frustration_tolerance	0.222±0.005	0.531±0.002	0.313±0.005	0.631±0.028
Catastrophising	0.291±0.005	0.509±0.000	0.371±0.004	0.796±0.012
Mind-reading	0.343±0.008	0.540±0.002	0.419±0.006	0.783±0.014
Labelling	0.376±0.004	0.597±0.003	0.462±0.004	0.853±0.007
Emotional_reasoning	0.241±0.006	0.417±0.004	0.306±0.006	0.748±0.017
Over-generalising	0.337±0.009	0.548±0.002	0.418±0.008	0.808±0.014
Inflexibility	0.162±0.002	0.336±0.006	0.218±0.003	0.807±0.012
Blaming	0.218±0.002	0.381±0.005	0.277±0.003	0.841±0.009
Disqualifying_the_positive	0.125±0.002	0.365±0.008	0.187±0.003	0.808±0.016
Mental_filtering	0.087±0.001	0.386±0.009	0.141±0.002	0.741±0.026
Personalising	0.179±0.003	0.345±0.007	0.236±0.004	0.871±0.009
Comparing	0.257±0.009	0.253±0.009	0.255±0.009	0.952±0.003

Table 7: Precision, recall, F1 score and accuracy for 300 dim CNN-GloVe with oversampling ratio 1:1

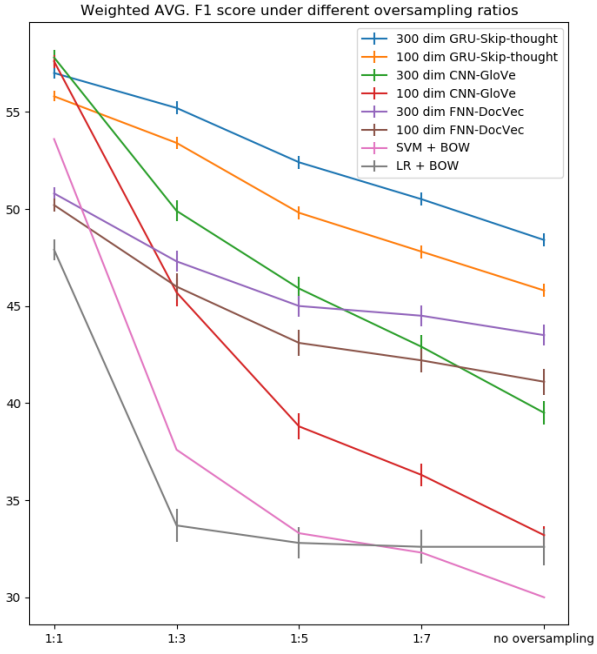

Figure 7: Weighted AVG. F1 for different models

then used the GloVe word vectors as input to a CNN and the skip-thought sentence vectors as input to a GRU. The results suggest that both models significantly outperform a chance classifier for all *thinking errors*, *emotions* and *situations* with CNN-GloVe on average achieving better results.

Areas of future investigation include richer distributed representations, or a fusion of distributed representations from word-level, sentence-level and document-level, to acquire more powerful semantic features. We also plan to extend the current ontology with its focus on *thinking errors*, *emotions* and *situations* to include a much lager number of concepts. The development of a statistical system delivering therapy will moreover require further research on other modules of a dialogue system.

Acknowledgements

This work was funded by EPSRC project Natural speech Automated Utility for Mental health (NAUM), award reference EP/P017746/1. The authors would also like to thank anonymous reviewers for their valuable comments. The code is available at `https://github.com/YinpeiDai/NAUM`

References

S. Arora, Y. Liang, and T. Ma. 2017. A simple but tough-to-beat baseline for sentence embedding. In *ICLR*.

A.T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. New York, International Universities Press.

A.T. Beck, J. Rush, B. Shaw, and G Emery. 1979. *Cognitive Therapy of Depression*. New York, Guildford Press.

Charissa Bhasi, Rohanna Cawdron, Melissa Clapp, Jeremy Clarke, Mike Crawford, Lorna Farquharson, Elizabeth Hancock, Miranda Heneghan, Rachel Marsh, and Lucy Palmer. 2013. Second Round of the National Audit of Psychological Therapies for Anxiety and Depression (NAPT).

Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30.

Robyn Bluhm. 2017. The need for new ontologies in psychiatry. *Philosophical Explorations*, 20(2):146–159.

R. Branch and R. Willson. 2010. *Cognitive Behavioural Therapy for Dummies*. Wiley.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Heriberto Cuayáhuitl. 2009. *Hierarchical reinforcement learning for spoken dialogue systems*. Ph.D. thesis, University of Edinburgh, Edinburgh.

D DeVault, R Artstein, G Ben, T Dey, E Fast, A Gainer, K Georgila, J Gratch, A Hartholt, M Lhommet, G Lucas, S Marsella, F Morbini, A Nazarian, S Scherer, G Stratou, A Suri, D Traum, R Wood, Y Xu, A Rizzo, and L-P Morency. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multiagent Systems*.

EU high-level conference: Together for Mental Health and Well-being. 2008. European Pact on Mental Health and Well-being.

Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. In *Proceedings of SIGDIAL*.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 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).

M. Gasic and S. Young. 2014. Gaussian processes for pomdp-based dialogue manager optimization. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(1):28–40.

M Geist and O Pietquin. 2011. Managing Uncertainty within the KTD Framework. In *Proceedings of the Workshop on Active Learning and Experimental Design*, Sardinia (Italy).

Sanjay Surendranath Girija. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Nathan B. Hansen, Michael J. Lambert, and Evan M. Forman. 2002. The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, 9(3):329–343.

Larry P Heck, Dilek Hakkani-Tür, and Gökhan Tür. 2013. Leveraging knowledge graphs for web-scale unsupervised semantic parsing. In *Proceedings of Interspeech*, pages 1594–1598.

Stefan Hofmann. 2014. Toward a cognitive-behavioral classification system for mental disorders. *Behavior Therapy*, 45(4):576 – 587.

TR Insel and EM Scholnick. 2006. Cure therapeutics and strategic prevention: raising the bar for mental health research. *Molecular Psychiatry*, 11(1):11-17.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. 2015. Skip-thought vectors. *NIPS*.

Quoc Le and Tomas Mikolov. 2014a. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Quoc Le and Tomas Mikolov. 2014b. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Jiwei Li, Will Monroe, Alan Ritter, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of EMNLP*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young. 2009. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of ICASSP*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

RR Morris, Schueller SM, and Picard RW. 2015. Efficacy of a Web-Based, Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression: Randomized Controlled Trial. *J Med Internet Res*, 17(3).

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Giuseppe Riccardi. 2014. Towards healthcare personal agents. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, RFMIR '14, pages 53–56, New York, NY, USA. ACM.

Lazlo Ring, Barbara Barry, Kathleen Totzke, and Timothy Bickmore. 2013. Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ACII '13, pages 61–66, Washington, DC, USA. IEEE Computer Society.

Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. 2016. An affectively aware virtual therapist for depression counseling. In *CHI 2016 Computing and Mental Health Workshop*.

Lina M. Rojas Barahona, M. Gasic, N. Mrkšić, P-H Su, S. Ultes, T-H Wen, and S. Young. 2016. Exploiting sentence and context representations in deep neural models for spoken language understanding. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 258–267, Osaka, Japan. The COLING 2016 Organizing Committee.

Lina Maria Rojas-Barahona and Toni Giorgino. 2009. Adaptable dialog architecture and runtime engine (adarte): A framework for rapid prototyping of health dialog systems. *I. J. Medical Informatics*, 78(Supplement-1):S56–S68.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

J Schatzmann, K Weilhammer, MN Stuttle, and S Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *KER*, 21(2):97–126.

D.F. Tolin. 2010. Is cognitivebehavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review*, 30:710–720.

Gökhan Tür, Minwoo Jeong, Ye-Yi Wang, Dilek Hakkani-Tür, and Larry P Heck. 2012. Exploiting the semantic web for unsupervised natural language semantic parsing. In *Proceedings of Interspeech*.

L.P. Vardoulakis, L. Ring, B. Barry, C. Sidner, and T. Bickmore. 2012. Designing relational agents as long term social companions for older adults. In Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 289–302. Springer Berlin Heidelberg.

Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34 – 49.

Joseph Weizenbaum. 1966. Eliza, a computer program for the study of natural language communication between man and machine. *ACM*, 9(1):36–45.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of ACL*.

World Health Organization. 2013. Mental health action plan 2013 - 2020.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, G. Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 189–194.

SJ Young. 2002. Talking to Machines (Statistically Speaking). In *Proceedings of ICSLP*.