Dialog Act classification using utterance embeddings

Lautaro Quiroz
10849963

Roger Wechsler
10850007

David Woudenberg
10069143

Abstract

In this work, we experiment with a new feature representation for dialog act tagging by learning distributional embeddings for utterances. We train a distributional semantic model that allows us to infer vector representations for entire utterances in an unsupervised fashion and use them to train several classifiers in order to tag utterances in the Switchboard corpus.

1 Introduction

Discourse structure analysis is essential for understanding spontaneous dialogs and developing human-computer dialog systems. An essential part of discourse structure analysis is the identification of dialog act classes (e. g. *questions*, *self-talks*, *statements*, *backchannels*). As defined by Austin (1962), dialog acts present linguistic abstractions of the illocutionary force of speech acts and model the communicative intention of an utterance in a conversation. There are several tasks that are require utterances to be tagged with dialog acts. Examples of these include speech recognition, speech synthesis, summarization, and of course, human-machine dialog systems. The correct identification of dialog act tags for utterances is thus an important research topic.

Table 1 shows an example of dialog acts from the Switchboard corpus we are trying to classify. The table already gives an idea that some of the 43 dialog acts might have a rather closed set of possible realisations (e.g. the classes *Agree* or *Acknowledge*) whereas classes like *Statement* can contain utterances of almost any content. Although the individual words in an utterance are important cues, we argue that the meaning of an utterance as a whole is essential for tagging it correctly.

For the purpose of this work, we extract feature representations for entire utterances in an unsupervised fashion. We therefore build a distributional semantic model that learns vector representations for entire utterances and then use these vector features as inputs for different machine learning classifiers, expecting that the embeddings can model the meaning of an entire utterance. Several techniques can be used for mapping text units to a high dimensional real value space. Utterance embeddings have the attractive property of representing an entire textual sequence as a vector while taking word order into account, as opposed to the classical *Bag-of-words* approach in which word order is not preserved and in which resulting vectors show only little semantic relations. We expect this additional information to play an important role in the classification task. In order to infer those embeddings we use the *paragraph2vec* framework recently introduced by (Le and Mikolov, 2014), which is based on the earlier word embedding models by (Mikolov et al., 2013).

Tag	**Speaker / Utterance**
Wh-Question	A: how old is your daughter?
Statement-non-opinion	B: she's three.
Summarize	oh , A: so she's a little one.
Agree	B: yes.
Acknowledge	A: yeah.
Statement-non-opinion	B: she's, she's little.

Table 1: SWDA dialog excerpt.

For the actual dialog act tagging we treat the problem as a multi-class classification task and classify utterances both in isolation as well as in the context of the previous utterances. We evaluate the tagging accuracy and compare different models. Besides the baselines provided by research from previous work, we compare the results against a simple baseline that uses a bag-of-words (BOW) representation for each utterance.

The outline of this report is set as follows: In Section 2, we present relevant approaches that aim at classifying dialog acts and briefly describe their main characteristics and results. In Section 3 we describe the concept of utterance embeddings and their training in more detail. Section 4 specifies the datasets and the details of our classification pipeline. An exhaustive analysis of the results of the experiments is presented in Section 5. Finally, Section 6 includes conclusions drawn from this work as well as issues and future work. [LQ: check]

2 Related Work

Several approaches have been proposed for classifying dialog acts. Most of them rely on supervised trained models, and use hand crafted features extracted for all utterances. Some recent work shows that using distributional representations for dialog act classification outperforms these methods. We briefly present some of the most relevant work in this section.

The authors of Stolcke et al. (2000) predict dialog acts by modeling a conversation as a Hidden Markov Model (HMM). A sequence of dialog acts is represented as a *discourse model* where the probability of the next dialog act depends on the *n* previous dialog acts. They integrate this model with a *language model* for each separate dialog act, which computes the possibility of the occurrence of all *word n-grams* in an utterance given a certain dialog act tag. In Stolcke et al. (2000) models are also trained on the actual speech signals, where the 'language model' is trained on prosodic and acoustic evidence. [DW: do we need to tell this last bit?] When considering the models trained on the dialog transcripts we can see that in this an utterance is represented as a bag of n-grams. We will try to find a representation that captures the composition of an utterance in a better way.

In Kalchbrenner and Blunsom (2013) a Recurrent Convolutional Neural Network (RCNN) is trained in a supervised manner on a corpus, which achieves state of the art results on the dialog act tagging task. The RCNN learns both a *discourse model* and a *sentence model* from a specific corpus, where the utterance representation is derived from individual word vectors, which are chosen randomly. We feel that this representation can become a lot richer if it is learned as in Le and Mikolov (2014), where word vectors have some distributional meaning. Another strength of this approach is that it is possible to train these utterance embeddings in an unsupervised manner, making it possible to additional, possibly unannotated, corpora.

An investigation on the contribution of distributional semantic information to the dialog act tagging task was conducted in Milajevs and Purver (2014). It was found that such information did improve tagging when compared to simple bag of words approaches. However only very simple distributional representations were investigated in this research, words were represented as a vector of their co-occurrences. Utterances as point wise multiplications or additions of these vectors, which implicates the loss of any compositional features. The work of Milajevs and Purver (2014) was not able to outperform the earlier work on dialog act tagging presented before.

In our work we will use the same *discourse model* as Stolcke et al. (2000) and represent the entire dialog as an HMM. However for our *sentence model* we will train different classifiers based on the embeddings, which are learned using the techniques from Le and Mikolov (2014). We will explain how we construct utterance embeddings in the next section. In section 4 we explain how we use these representations to build a *sentence model* as well and briefly repeat the *discourse model* as an HMM.

[DW: This ok?] [RW: I agree. And then this can serve as a transition to the utt2vec explanation in the next section.]

3 Utterance Embeddings

[RW: This section still contains some redundant parts and needs to be restructured.] [DW: I did a first attempt at fixing it, but it still needs some (small) updates]

3.1 Word Embeddings

Neural networks for distributional semantics has gained relevant importance in the past years since the publication of the first neural network word embedding models. One of the main reasons was the discovery that dense high-dimensional vector representations of words are able to capture semantic relations between words (Mikolov et al., 2013). Figure 1 shows a typical example where simple vector addition and subtraction lead to analogies such as: The vector difference between *woman* and *man* is the same as between *aunt* and *uncle*.

In distributional semantic models, word embeddings are learned by predicting words within a

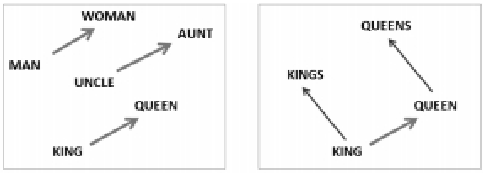

Figure 1: Word2vec semantic relations.

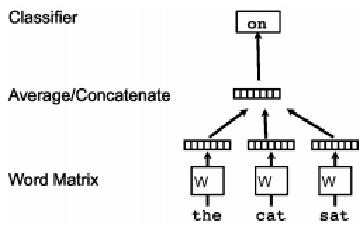

Figure 2: Paragraph vector PV-DM architecture.

context of other words. Based on the fact that similar words occur in similar contexts, these models are capable of successfully representing words in high-dimensional vectors.

Originally, these approaches had two main architectures, know as *Continuous-bag-of-words* and *Skip-gram*. In the first case, the neural networks were optimized to predict the next word given its context, whereas in the latter, the context is predicted given a certain word. Due to word co-occurrences, these models are able to effectively capture a semantic representation of the words.

3.2 Utterance embeddings

The approach from Mikolov et al. (2013) is no longer feasible on a level for word sequences such as entire utterances, as the vocabulary of all possible utterances is infinitely large, as we can construct utterances of any length. In order to overcome this sparsity problem, a slightly different framework called *paragraph2vec*[1] for learning representations of entire word sequences has been proposed by Le and Mikolov (2014), in which vectors are learned for a sequence from the words within the sequence. For our case, the model thus trains utterance embeddings as well as word embeddings.

During training, two structures are maintained: one for words, and one for utterance representations. The word structure is shared across the entire model, but an utterance embedding is trained for each single utterance individually. The training task is the same as before: Given a certain window, the model is optimized to predict the missing word. However this time, the context representation is constructed using the individual word vectors together with the utterance vector. This training schema is called *Paragraph Vector Distributed memory (PV-DM)*. Figure 2 shows the *PV-DM* architecture.

[1]Note that despite its name, the framework is able to learn embeddings for entire word sequences of any length like sentences, paragraphs or entire documents.

Once the model is trained, it can be queried in order to get the fixed-length vector representations for each observed utterance. It can also be used to *infer* vectors for *unseen* utterances. This step is crucial to obtain the embeddings for utterances in our test dataset.

The approach presented by Le and Mikolov (2014) has two very nice advantages. Firstly, it is completely unsupervised, and thus we can use any amount of unannotated data that is available. Moreover, the model allows the use of pretrained word embeddings for initialization instead of initializing the parameters randomly.[2] This is useful if there is not enough data available to learn reliable word embeddings together with the sequence embeddings. For our experiments, we try both methods: training word and paragraph embeddings entirely from scratch using different dialog corpora as input, as well as using the freely available pretrained word embeddings based on the *Google News Corpus*[3] to initialize the model.

4 Methodology

[DW: This needs a brief introduction with set up of this section]

4.1 Dialog datasets

[LQ: Describe the switchboard corpus. It's 42 tags.] [LQ: Describe the BNC corpus.] Throughout the classification pipeline, we make use use different dialog corpora. With the purpose of training word embedding models, options include training using a combination of: the Switchboard corpus, the British National corpus, and pretrained word vectors.

For the evaluation step is mandatory to utilize labelled data, for this reason, we train and evaluate our classifiers on the Switchboard corpus. We

[2]Words that are not found in the pretrained model are still initialized randomly.

[3]`https://code.google.com/p/word2vec/`

divide the dataset into a training and a test set, containing % and % of the total length, respectively. [LQ: add percentages]

In the next subsections we present a description of each dataset.

The Switchboard corpus

The Switchboard Dialog Act corpus (SwDA) consists of a compilation of telephone transcripts between two interlocutors. It contains a total 205.000 utterances and 1.4 million words While many features are defined for each utterance unit, the most important for our work is the tag attribute. Each utterance is associated to a label, which summarizes syntactic, semantic, and pragmatic information. The corpus contains a total of 200 tags, which can be further aggregated into 44 main classes. Table 2 shows examples for the five most common tags. Table 3 present examples of utterances contained in the SwDA corpus.

[LQ: add cite to SWDA]

Tag	Description	Example	%
st	Statement-non-opinion	Me, I'm in the legal department.	36
b	Acknowledge	Acknowledge Uh-huh.	19
sv	Statement-opinion	I think it's great.	13
aa	Accept	That's exactly it.	5
%	Turn exit	So,-	5

Table 2: SWDA's 5 most frequent tags.

qrr B.34.utt2: C or do you think that [we're, + we're,] F uh, all trying to keep up with a certain standard of living?
sv A.35.utt1: I think that's part of it too.
sv A.35.utt2: C But I do think, -
qy B.36.utt1: E I mean do you think,

Table 3: SwDA utterance examples.

The British National corpus

The British National corpus (BNC) is a collection of 100 million words sampled from different written and spoken sources, with the intention of representing the British English language. Some of the sources of these data include newspapers, articles, journals, books, letter, transcription of informal conversations, among others. This dataset contains a huge amount of unlabelled sentences. Table 4 presents sentence examples extracted from the BNC.

ADR 172 The kind of girl that even if you didn't know well you always said "hello" to and got a cheery wave and a smile back.
B72 966 For example, the sedimentary rocks that form the top geological layer in much of southern Britain may be only a few hundred metres thick in a few isolated sites.
B2E 714 Then Fulham got one of her worst raids of the war.

Table 4: BNC sentence examples.

4.2 Discourse Model

[DW: Please check the following on readability]

Just like Stolcke et al. (2000) we will model a dialog as an HMM. The hidden states will be the dialog acts and the observed quantities the uttered words and their speaker. To find the most likely tag for a single utterance we will find the best dialog act tag t for an utterance using equation 1.

$$\begin{aligned} t^* &= \underset{t}{argmax} P(t|u) \\ &= \underset{t}{argmax} \frac{P(u|t)P(t)}{P(u)} \\ &= \underset{t}{argmax} \underbrace{P(u|t)}_{\text{sentence model}} \underbrace{P(t)}_{\text{discourse model}} \end{aligned} \tag{1}$$

We discuss the sentence model in detail in section 4.3. As a discourse model we will assume that the *prior probability* of a tag depends on the preceding tags and their speakers. $P(t_i) = P(t_i|t_{i-1}...t_0, s_{i-1}...s_0)$. Where we assume that the discourse is a Markov Model of order k: $P(t_i) = P(t_i|t_{i-1}...t_{i-k}, s_{i-1}...s_{i-k})$.

Different dynamic programming techniques can be used to find the most probable sequence of tags $\mathbf{t}$ given a sequence of utterances $\mathbf{u}$ and their speaker $\mathbf{s}$, we will use the Viterbi decoding algorithm Ryan and Nudd (1993).

[DW: the following could go to the discussion, but i think it is good to note these assumptions straight away] An assumption made in this model is that two sequential utterances u_i and u_{i+1} are independent of each other. We know that in conversation interlocutors tend to align on different levels including lexical choices Danescu-Niculescu-Mizil et al. (2012). Also it is very likely that the same words will be used in the answer to a question as in the question. However we count on the fact that the dependence of u_i and t_i will be stronger than the is stronger than the independence violation of u_i and u_{i+1} Stolcke et al. (2000).

4.3 Sentence Models

[DW: Do we want to do this before or after the discourse model?] [LQ: Mention which classifiers we use, and with which params]

5 Results

5.1 Evaluating Sentence Models

[DW: results from the huge table we have online, find a nice way to summarize this]

5.2 Intersecting Sentence and Discourse Models

[DW: Baseline]

[DW: Using 'best' sentence models from last subsection]

5.3 Error Analysis

Utterance embeddings have the property of encapsulating the meaning of utterances from the words that compose them; as with the case of word embeddings, these representations present appealing relations from which similar words (in a semantic sense) appear close by in the high dimension space. We being by extracting samples from the SwDA corpus belonging to clear unrelated tags, and applying dimensionality reduction to their vector representations and plotting the results in 2 and 3 dimensions. The sampled utterances belong to the following categories: *Wh-question*, *Thanking*, *Reject*, and *Apology*. Considering the words that are used in these kind of units, sample points should be clearly separated. The utterance embeddings were extracted using a 300-dimensional model trained on the SwDA corpus interjected with Google pretrained vectors. Figure 5 and Figure 6 show scatter diagrams of the utterance embeddings after applying PCA.

Though, the classification task seems easy considering the previous tag examples, it gets more complicated when we handle other utterance choices. Figure 5 and Figure 6 show the scatter diagram after applying PCA to utterance embedding of the 5 most frequent tags (*Statement-non-opinion*, *Acknowledge*, *Statement-opinion*, *Accept*, *Turn exit*), which account for 78% of the corpus.

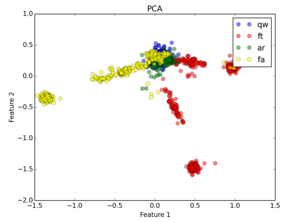

Figure 3: 2D PCA.

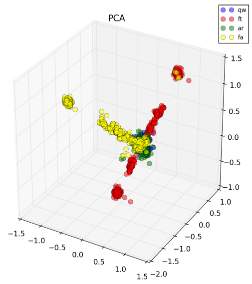

Figure 4: 3D PCA.

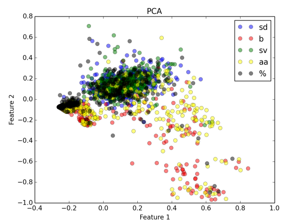

Figure 5: 2D PCA.

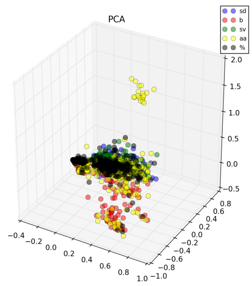

Figure 6: 3D PCA.

6 Conclusion

- repeat results, what do they imply for hypotheses?
- criticisms to what we did
- possible future work
- ...

References

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

L John. 1962. Austin. how to do things with words.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *CoRR*, abs/1306.3584.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Dmitrijs Milajevs and Matthew Purver. 2014. Investigating the contribution of distributional semantic information for dialogue act classification. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 40–47, Gothenburg, Sweden, April. Association for Computational Linguistics.

Matthew S. Ryan and Graham R. Nudd. 1993. The viterbi algorithm. Technical report, Coventry, UK, UK.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373, September.