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

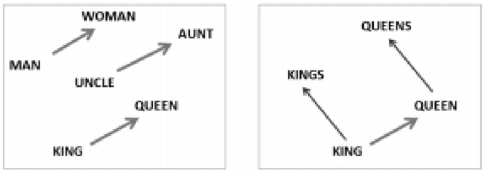

Figure 1: Word2vec semantic relations.

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.

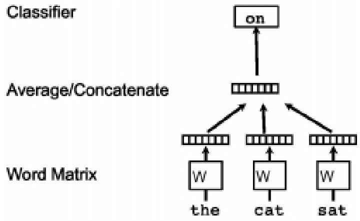

Figure 2: Paragraph vector PV-DM architecture.

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

We evaluate our utterance embedding features for dialog act taggin in two different settings: 1) We tag each utterance independent of the context it appears in. 2) We include the context of the previous utterance by concatenating the utterance embeddings of the previous utterance and the current utterance to obtain a very simplified discourse model as discussed in Section 2. In addition, we experiment with two different tagset granulairities. The dataset, the actual classifiers and the tagsets are described in the following two subsections. [RW: Baselines]

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

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

4.1 Dialog datasets

The dialog act classifiers are trained in a supervised fashion, which required labelled data. For that purpose, we evaluate the actual dialog act tagging on the Switchboard Dialog Act corpus (SwDA) [RW: cite]. The corpus is a compilation of telephone transcripts between two interlocutors. It contains a total of 205,000 utterances and 1.4 million words. Each utterance is assigned one of 42 possible dialog act tags (according to teh Discourse Annotation and Markup System of Labelling - DAMSL). In our experiments, we also use a coarser tagset where we manually map the original tags into [RW: how many?] classes. The mapping is described in [RW: appendix or where?]. Table 1 shows an example of a small excerpt from the corpus with the respective dialog act labels.

We use the same training and test split as described in Stolcke (2000), which uses 1,115 dialogs for training and 19 dialogs for testing and which has been widely used in previous work. The utterances are preprocessed by removing linguistic annotations and any interpunctuation except for periods, question marks and exclamation marks. Additionally, utterances that complete previously interrupted utterences, which are marked with a continuation tag +, were concatenated with their initial segment, which also contains the correct dialog act label for the complete utterance. That preprocessing step was also performed by Milajevs and Purver (2014).

As mentioned before, the training of the utterance embeddings is completely unsupervised and does not require any labelled data. For that reason, we experiment with expanding our data for the training of the utterance representations. In addition to the Switchboard corpus, we use the spoken dialog data from the British National Corpus (BNC) [RW: cite]. The resulting dataset from the BNC contains approximately 1 million utterances. The BNC utterances, however, different inherently from the Switchboard utterances as they are usually longer and normalized to full grammatical sentences. The utterances in the SWDA dataset are more fragmentary and thus usually shorter.

4.2 Classifiers

[LQ: Mention which classifiers we use, and with which params] For the actual classification task we use three different classifiers which we treat as

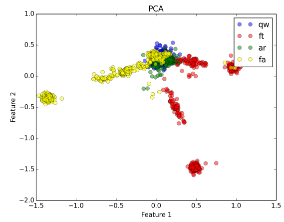

Figure 3: 2D PCA.

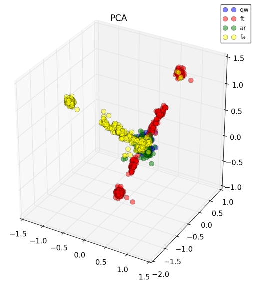

Figure 4: 3D PCA.

"black-boxes," i. e. we do not attempt to evaluate their hyperparameters.

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

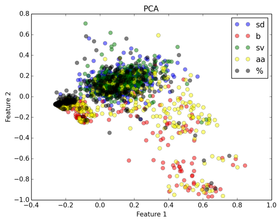

Figure 5: 2D PCA.

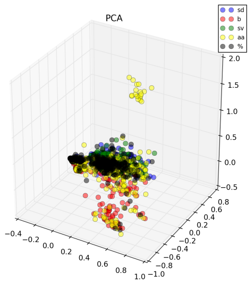

Figure 6: 3D PCA.

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.

6 Conclusion

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

References

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.

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.