Computational Pragmatics
Project: Utterance2Vec

Lautaro Quiroz
10849963

Roger Wechsler
10850007

David Woudenberg
10069143

Abstract

Traditionally, automatic Dialog Act Tagging has been based on training machine learning classifiers using hand-crafted features extracted from data. 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 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 have dialog acts as an input for their computations. Examples of these include speech recognition, speech synthesis, summarization, and of course, human-machine dialog systems. As a result, correctly identifying dialog act tags is fundamental for such tasks.

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 possibilities (e.g. the classes Agree or Acknowledge) whereas classes like Statement can contain 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.

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 purpose of this work, we extract feature representations for entire utterances in an unsupervised fashion. For that we 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 no semantic relations. We expect this encapsulated extra 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 previous word embedding models by (Mikolov et al., 2013).

[RW: I would put the following two paragraphs about embeddings into 3.1 Utterance Embeddings][DW: I agree but it could also go to the related work?] Neural networks for semantics has gained relevant importance in the past years, since the publication of the first neural networks word embedding models. One of the main reasons for this to happen, is the fact that the resulting vector representations are able to effectively capture the meaning of words given their context. Figure 1 shows a typical example of this character-

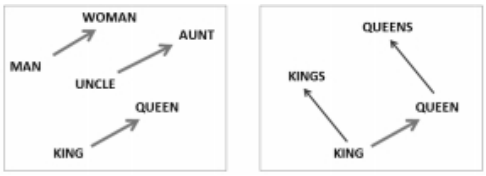

Figure 1: Word2vec semantic relations.

istic. [DW: We need to explain what can be seen in that figure]

In these models, word embeddings are learned by predicting words within a context (depending on the neural network model -i.e. CBOW or Skip-gram, it could be that the network tries to predict a word given a context, or the context given a word). Due to the co-occurrence of words in texts, given the use of a language, these models are capable of successfully coming up with mathematical representations of words. This idea of word co-occurrences is not longer valid when we take sentences as units; a sentence can have -potentially- any other valid sentence before and after. Instead, a neural network for learning representations for sentences has been proposed, in which the sentence structures are learned from the words that are included in it (Le and Mikolov, 2014). In this case, word embeddings and sentence embeddings are trained simultaneously, where the sentence is represented with a vector, shared across the words within the phrase. This sentence representation is maintained and updated in parallel to the word vectors. This unsupervised model makes it possible to map sentences (utterances) of any length to vectors of fixed length. Moreover, an interesting characteristic of this model is that any (unannotated) corpus can be used as input for training, allowing us to create huge training sets, based on several dialog corpora from different domains and with different characteristics.

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 set up by previous work, we compare the results against a simple baseline that uses a bag-of-words (BOW) representation for each utterance. [DW: as the paragraphs above might be the complicated part of our research we should spend some time getting this as clear and readable as possible]

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 specify the datasets used in our experiments, present the model to infer utterance embeddings and we detail the properties of our classification pipeline. An exhaustive analysis of the results of the experiments is presented in Section 4. Finally, Section 5 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. [LQ: is this true?] 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. 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. 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.

[DW: Explaining (Le and Mikolov, 2014) might be nice HERE? or maybe before (Stolcke et al., 2000) in a separate subsection?]

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.

[DW: We need to explicitly state how we differ from all this previous work]

3 Methodology

3.1 Utterance embeddings

In order to get vector representations from utterances, we use an extension of the word embeddings neural network proposed by Mikolov(Mikolov et al., 2013). Originally, these networks had two main architectures, know as *Continuous-bag-of-words* and *Skip-gram*. In the first case, they were optimized so as to predict the next word given its context, while in the latter, a word is input and the context is predicted. Due to word co-occurrences, these models are able to effectively capture the meaning of the words. The co-occurrence property present at the word level is no longer valid when handling phrases. For sentence embeddings a novel approach was recently introduced [LQ: cite], in which a similar training algorithm is followed. In this case, two structures are maintained (one for words, and one for sentence representations); the word structure is shared across all sentences, while the sentence structure is only valid for the current paragraph. The task is the same as before: given a certain window, the model is optimized to predict the missing word; but this time, the context representation is constructed using the individual

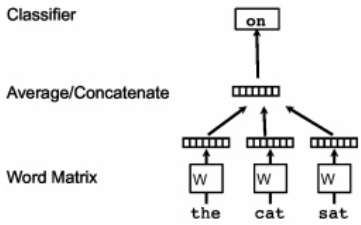

Figure 2: Paragraph vector PV-DM architecture.

word vectors as well as the paragraph vector. This training schema is called *Paragraph Vector Distributed memory (PV-DM)*, and it is the one we use to train our model. Figure 2 shows the *PV-DM* architecture.

Once the model is trained, it can be queried so as to get the vector representations for each already seen sentence. It can also be used to *infer* vector mappings of *unseen* sentences. This second step will be crucial to get representations for utterances in our test dataset.

A great advantage of these approach is that it is completely unsupervised, and thus, we can use any amount of unlabelled data we want.

Word2vec pretrained vectors

The implementation we use for getting paragraph vectors allows us to use already trained word embeddings. The way in which this feature works follows can be summarized as: (a) No new words are added to the vocabulary; (b) Intersecting words adopt pretrained values ; (c) Non intersecting words are left alone. For our experiments, we try both alternatives, training word and paragraph embeddings from scratch using several dialog corpora as input, and also using freely available embeddings[1].

3.2 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

[1] https://code.google.com/p/word2vec/

divide the dataset into a training and a validation 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 transcriptions between two participants. 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.

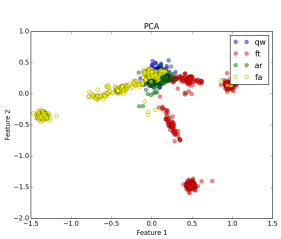

Figure 3: 2D PCA.

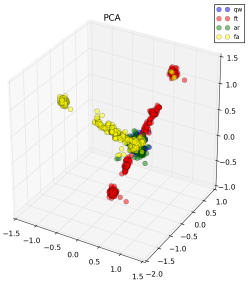

Figure 4: 3D PCA.

3.3 Classifiers

[LQ: Mention which classifiers we use, and with which params]

4 Results

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

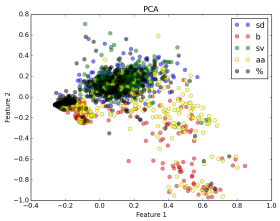

Figure 5: 2D PCA.

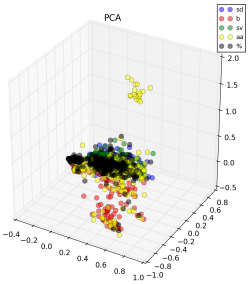

Figure 6: 3D PCA.

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.

5 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.