Proceedings of HLT-NAACL 2003
Short Papers , pp. 94-96
Edmonton, May-June 2003

Latent Semantic Analysis for dialogue act classification

Riccardo Serafin
Computer Science
University of Illinois
Chicago, IL, USA
rseraf1@uic.edu

Barbara Di Eugenio
Computer Science
University of Illinois
Chicago, IL, USA
bdieugen@uic.edu

Michael Glass
Mathematics and Computer Science
Valparaiso University
Valparaiso, IN, USA
Michael.Glass@valpo.edu

Abstract

This paper presents our experiments in applying Latent Semantic Analysis (LSA) to dialogue act classification. We employ both LSA proper and LSA augmented in two ways. We report results on DIAG, our own corpus of tutoring dialogues, and on the CallHome Spanish corpus. Our work has the theoretical goal of assessing whether LSA, an approach based only on raw text, can be improved by using additional features of the text.

1 Introduction

Dialogue systems need to perform dialog act classification, in order to understand the role the user's utterance plays in the dialog (e.g., a question for information or a request to perform an action), and to generate an appropriate next turn. In recent years, a variety of empirical techniques have been used to train the dialogue act classifier (Reithinger and Maier, 1995; Stolcke et al., 2000; Walker et al., 2001).

In this paper, we propose Latent Semantic Analysis (LSA) as a method to train the dialogue act classifier. LSA can be thought as representing *the meaning of a word as a kind of average of the meanings of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains* (Landauer et al., 1998). LSA learns from co-occurrence of words in collections of texts. It builds a semantic space where words and passages are represented as vectors. Their similarity is measured by the cosine of their contained angle in the semantic space. LSA is based on Single Value Decomposition (SVD), a mathematical technique that causes the semantic space to be arranged so as to reflect the major associative patterns in the data, and ignores the smaller, less important influences.

LSA has been successfully applied to many tasks: e.g, to assess the quality of student essays (Foltz et al., 1999) and to interpret the student's input in an Intelligent Tutoring system (Graesser et al., 2000). However, there is no research on applying LSA to dialogue act classification.

LSA is an attractive method because it is relatively straightforward to train and use. More importantly, although it is a statistical theory, it has been shown to mimic a number of aspects of human competence / performance (Landauer et al., 1998). Thus, it appears to somehow capture and represent important components of meanings.

We also have a theoretical goal in investigating LSA. A common criticism of LSA is that its "bag of words" approach ignores any other linguistic information that may be available, e.g. order and syntactic information: to LSA, *man bites dog* is identical to *dog bites man*. We suggest that an LSA semantic space can be built from the co-occurrence of arbitrary textual features. We propose to place in the bag of words other features that co-occur in the same text. We are calling LSA augmented with features "FLSA" (for "feature LSA"). The only relevant prior work is (Wiemer-Hastings, 2001), that adds part of speech tags and some syntactic information to LSA.

This paper describes the corpora and the methods we used, and the results we obtained. To summarize, plain LSA seems to perform well on large corpora and classification tasks. Augmented LSA seems to perform better on smaller corpora and target classifications.

2 Corpora

We report experiments on two corpora, DIAG and Spanish CallHome.

DIAG is a corpus of computer mediated tutoring dialogues between a tutor and a student who is diagnosing a fault in a mechanical system with the DIAG tutoring system (Towne, 1997). The student's input is via menu, the tutor is in a different room and answers via a text window. The DIAG corpus comprises 23 dialogues for a total of

607 different words and 660 dialogue acts. It has been annotated for a variety of features, including four dialogue acts[1] (Glass et al., 2002): *problem solving*, the tutor gives problem solving directions; *judgement*, the tutor evaluates the student's actions or diagnosis; *domain knowledge*, the tutor imparts domain knowledge; and *other*, when none of the previous three applies.

The Spanish CallHome corpus (Levin et al., 1998; Ries, 1999) comprises 128 unrestricted phone calls in Spanish, for a total of 12066 different words and 44628 dialogue acts. The Spanish CallHome annotation augments a basic tag such as *statement* along several dimensions, such as whether the statement describes a psychological state of the speaker. This results in 232 different dialogue act tags, many with very low frequencies. In this sort of situations, tag categories are often collapsed when running experiments so as to get meaningful frequencies (Stolcke et al., 2000). In CallHome37, we collapsed statements and backchannels, obtaining 37 different tags. CallHome37 maintains some subcategorizations, e.g. whether a question is yes/no or rhetorical. In CallHome10, we further collapse these categories. CallHome10 is reduced to 8 dialogue acts proper (eg statement, question, answer) plus the two tags ``%'' for abandoned sentences and ``x'' for noise.

3 Methods

We have experimented with four methods: LSA proper, which we call plain LSA; two versions of clustered LSA, in which we 'cluster' the document dimension in the Word-Document matrix; FLSA, in which we incorporate features other than words to train LSA (specifically, we used the preceding *n* dialogue acts).

Plain LSA. The input to LSA is a Word-Document matrix with a row for each word, and a column for each *document* (for us, a document is a unit such as a sentence or paragraph tagged with a dialogue act). Cell $c(i, j)$ contains the frequency with which $word_i$ appears in $document_j$. Clearly, this w*d matrix will be very sparse. Next, LSA applies SVD to the Word-Document matrix, obtaining a representation of each document in a k dimensional space: crucially, k is much smaller than the dimension of the original space. As a result, words that did not appear in certain documents now appear, as an estimate of their correlation to the *meaning* of those documents. The number of dimensions k retained by LSA is an empirical question. The results we report below are for the best k we experimented with.

To choose the best tag for a document in the test set, we compare the vector representing the new document with the vector of each document in the training set. The tag of the document which has the highest cosine with our test vector is assigned to the new document.

Clustered LSA. Instead of building the Word-Document matrix we build a Word-Tag matrix, where the columns refer to all the possible dialog act types (tags). The cell $c(i, j)$ will tell us how many times $word_i$ is used in documents that have tag_j. The Word-Tag matrix is w*t instead of w*d. We then apply Plain LSA to the Word-Tag matrix.

SemiClustered LSA. In Clustered LSA we lose the distribution of words in the documents. Moreover, if the number of tags is small, such as for DIAG, SVD loses its meaning. SemiClustered LSA tries to remedy these problems. We still produce the k-dimensional space applying SVD to the Word-Document matrix. We then reduce the Word-Tag matrix to the k dimensional space using a transformation based on the SVD of the Word-Document matrix. Note that both Clustered and SemiClustered LSA are much faster at test time than plain LSA, as the test document needs to be compared only with t tag vectors, rather than with d document vectors ($t << d$).

Feature LSA (FLSA). We add extra features to plain LSA. Specifically, we have experimented with the sequence of the previous n dialogue acts. We compute the input WordTag-Document matrix by computing the Word-Document matrix, computing the Tag-Document matrix and then concatenating them vertically to get the $(w+t)$*d final matrix. Otherwise, the method is the same as Plain LSA.

4 Results

Table 1 reports the best results we obtained for each corpus and method. In parentheses, we include the k dimension, and, for FLSA, the number of previous tags we considered.

In all cases, we can see that Plain LSA performs much better than baseline, where baseline is computed as picking the most frequent dialogue act in each corpus. As concerns DIAG, we can also see that SemiClustered LSA improves on Plain LSA by 3%, but no other method does.

As regards CallHome, first, the results with plain LSA are comparable to published ones, even if the comparison is not straightforward, because it is often unclear what the target classification and features used are. For example, (Ries, 1999) reports 76.2% accuracy by using neural networks augmented with the sequence of the n previous speech acts. However, (Ries, 1999) does not mention the target classification; the reported baseline appears compatible with both CallHome37 and CallHome10. The training features in (Ries, 1999) include part-of-speech (POS) tags for words, which we do not have. This may

[1] They should be more appropriately termed *tutor moves*.

Corpus	Plain	Clustered	SemiClustered	FLSA
Diag (43.64%)	75.73% (50)	71.91% (3)	78.78% (50)	74.26% (1,150)
CallHome37 (42.69%)	65.36% (50)	22.08% (10)	31.39% (300)	62.59% (1, 50)
CallHome10 (42.69%)	68.91% (25)	61.64% (5)	58.38% (300)	66.57% (1, 100)

Table 1: Result Summary

explain the higher performance. Including POS tags into our FLSA method is left for future work.

No variation on LSA performs better than plain LSA in our CallHome experiments. In fact, clustered and semiclustered LSA perform vastly worse on the larger classification problem in CallHome37. It appears that, the smaller the corpus and target classification are, the better clustered and semiclustered LSA perform. In fact, semiclustered LSA outperforms plain LSA on DIAG.

Our experiments with FLSA do not support the hypothesis that adding features different from words to LSA helps with classification. (Wiemer-Hastings, 2001) reports mixed results when augmenting LSA: adding POS tags did not improve performance, but adding some syntactic information did. Note that, in our experiments, adding more than one previous speech act did not help.

5 Future work

Our experiments show that LSA can be effectively used to train a dialogue act classifier. On the whole, plain LSA appears to perform well. Even if our experiments with extensions to plain LSA were mostly unsuccessful, they are not sufficient to conclude that plain LSA cannot be improved. Thus, we will pursue the following directions. 1) Further investigate the correlation of the performance of (semi)clustered LSA with the size of the corpus and / or of the target classification. 2) Include other features in FLSA, e.g. syntactic roles. 3) Redo our experiments on other corpora, such as Map Task (Carletta et al., 1997). Map Task is appropriate because besides dialogue acts it is annotated for syntactic information, while CallHome is not. 4) Experiment with FLSA on other tasks, such as assessing text coherence.

Acknowledgements

This work is supported by grant N00014-00-1-0640 from the Office of Naval Research.

References

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

P. W. Foltz, D. Laham, and T. K. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

M. Glass, H. Raval, B. Di Eugenio, and M. Traat. 2002. The DIAG-NLP dialogues: coding manual. Technical Report UIC-CS 02-03, University of Illinois - Chicago.

A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, and the Tutoring Research Group. 2000. Autotutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

L. Levin, A. Thymé-Gobbel, A. Lavie, K. Ries, and K. Zechner. 1998. A discourse coding scheme for conversational Spanish. In *Proceedings ICSLP*.

N. Reithinger and E. Maier. 1995. Utilizing statistical dialogue act processing in Verbmobil. In *ACL95, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*.

K. Ries. 1999. HMM and Neural Network Based Speech Act Detection. In *Proceedings of ICASSP 99*.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

D. M. Towne. 1997. Approximate reasoning techniques for intelligent diagnostic instruction. *International Journal of Artificial Intelligence in Education*,8.

M. A. Walker, R. Passonneau, and J. E. Boland. 2001. Qualitative and quantitative evaluation of DARPA communicator dialogue systems. In *ACL01, Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

P. Wiemer-Hastings. 2001. Rules for syntax, vectors for semantics. In *CogSci01, Proceedings of the Twenty-Third Annual Meeting of the Cognitive Science So*