Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect?

Alexander Osherenko and Elisabeth André

Multimedia Concepts and Applications, Faculty of Applied Informatics
University of Augsburg, Germany
{osherenko,andre}@informatik.uni-augsburg.de

Abstract. Recently, there has been considerable interest in the automated recognition of affect from written and spoken language. In this paper, we investigate how information on a speaker's affect may be inferred from lexical features using statistical methods. Dictionaries of affect offer great promise to affect sensing since they contain information on the affective qualities of single words or phrases that may be employed to estimate the emotional tone of the corresponding dialogue turn. We investigate to what extent such information may be extracted from general-purpose dictionaries in comparison to specialized dictionaries of affect. In addition, we report on results obtained for a dictionary that was tailored to our corpus.

Keywords: Lexical Modality, Lexical Affect Sensing, Emotion Detection, Spontaneous Dialogues, Affect Dictionaries.

1 Introduction

Recently, there has been considerable interest in the automated recognition of affect from written and spoken language. The driving force behind this work is the observation that a computer system is more likely to be accepted by the human user if it is able to recognize his or her emotions and respond accordingly. Psychological studies reveal that the user's emotional state significantly affects his or her phrasing. For instance, if someone is in a state of high arousal, his or her phrasing tends to be more stereotypical and less diversified [6]. Weintraub observed in an experiment that speakers that were considered as more emotional used few non-personal references and a fair number of expressions of feeling [15]. In this paper, we investigate how information on a speaker's affect may be inferred from lexical features.

Dictionaries of affect offer great promise to lexical affect sensing since they contain information on the affective qualities of single words or phrases that may be employed to estimate the emotional tone of the corresponding dialogue turn. Dictionaries of affect are usually composed drawing on human common-sense knowledge about affect words or employing rating tools, such as semantic differential scales. Examples include the Whissell's Dictionary of Affect Language (DAL) [16], the LIWC2001 Dictionary [10] or the WordNet-Affect Database of ITS-IRST [14].

A number of approaches to affect sensing take advantage of the information included in affect dictionaries either exclusively or in addition to other features.

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 230–241, 2007.

Prendinger and colleagues [9] make use of the emotional senses in WordNet to estimate the emotional content of words in a document. Shaikh and colleagues [13] introduce an approach to affect sensing that is based on manually collected sentiment verbs and adjectives from WordNet. Zhang and colleagues rely on the Heise dictionary and WordNet to extract affect from speech in e-drama [19]. Nasukawa and Yi [8] conduct a sentiment analysis using a dictionary consisting of 3,513 affective adjectives, adverbs and nouns.

Affect dictionaries have also been proven useful to discriminate deceptive from non-deceptive speech. For instance, Hirschberg and colleagues [3] make use of the LIWC and DAL dictionaries to extract features from speech. They observe that the pleasantness score as well as the occurrence of positive emotion words seem to be a promising factor in predicting deception. Deceptive speech tends to have a greater pleasantness score and a greater proportion of positive emotion words than truthful speech.

Mairesse and Walker [7] aim at identifying a speaker's personality by means of a conversational analysis using the LIWC dictionary [10] as well as the MRC psycholinguistic database [1]. They observe that emotional stability is best predicated by features extracted from MRC while LIWC features show a better performance for other personality traits.

In this paper, we concentrate on affect recognition for spontaneous utterances based on the affective qualities of words. We investigate to what extent such information may be extracted from general-purpose dictionaries or dictionaries of affect. In particular, we focus on the following questions:

1. Do we get higher recognition rates if we restrict ourselves to word features that convey emotional content?
2. Do emotive annotations in dictionaries improve affect sensing for dialogue turns?
3. Are common words more useful to affect sensing than less common words?
4. Are dictionaries of affect more useful to affect sensing than general-purpose dictionaries?

In this paper, we answer these questions by evaluating two dictionaries of affect as well as one general-purpose dictionary. In addition, we create our own dictionary based on the corpus we investigate – a corpus containing transcriptions of spontaneous speech.

2 Dictionaries

In this study, we consider two affect dictionaries, the Whissell's Dictionary of Affect Language (DAL) [16] and the Linguistic Inquiry and Word Count Dictionary (LIWC) [10], as well as word frequency lists that are based on the general-purpose British National Corpus (BNC) and word frequency lists extracted from our own affect corpus.

DAL contains 8,742 words of different inflection that are characterized by their emotional connotation along three dimensions: evaluation, activation and imagery. Scores for the evaluation range from 1 (unpleasant) to 3 (pleasant), for the activation

range from 1 (passive) to 3 (active), for imagery range from 1 (difficult to form a mental picture of this word) to 3 (easy to form a mental picture). The scores have been determined by human judgment. In the following, the original scale is mapped from 1 to 3 to -1 to 1 for better readability.

The LIWC affect dictionary contains 2,251 word patterns that represent words ending with a wildcard e.g. the pattern *abandon** describes all inflections of word *abandon*. Unlike DAL, the LIWC does not characterize words by using continuous scales, but classifies entries categorically – there are word patterns of 68 different categories in LIWC. We are mostly interested in the category Affective or Emotional Processes (*Affect*) with 617 patterns referring to affective or emotional processes. *Affect* is subcategorized into positive emotions (*Posemo*) and negative emotions (*Negemo*). *Posemo* again has two subcategories: positive feelings (*Posfeel*), optimism and energy (*Optim*) while *Negemo* has three subcategories – anxiety or fear (*Anx*), anger (*Anger*), sadness (*Sad*). For example, the word *afraid* got the labels *Affect*, *Negemo* and *Anx*.

The DAL or LIWC dictionaries are compiled specifically for emotional analysis. In contrast, the BNC frequency list is a general-purpose list based on a 100 million word collection obtained from a wide range of written and spoken language sources representing contemporary British English [4]. Each word in the list is indicated with its frequency in the BNC corpus e.g. the word *children* occurs in the corpus 46,577 times. We compose several smaller frequency lists (*BNC-threshold*) out of the original BNC frequency list by discarding words for which the BNC frequency does not exceed the specified threshold. For instance, BNC-650 contains the subset of 8,603 words from the BNC corpus whereas each of words occurs more than 650 times in the whole 100M-word BNC. Additionally, we examine frequency lists (*SAL-n*) that are calculated based on the studied affect corpus SAL (see Section 4).

3 Corpus

In our study, the affective meaning of dialogue utterances from the SAL corpus is explored [5]. The freely available SAL corpus contains audio-visual data of four users communicating with one of four psychologically different characters: optimistic and outgoing (Poppy), confrontational and argumentative (Spike), pragmatic and practical (Prudence), depressing and gloomy (Obadiah) that try to draw the user into their own emotional state. The corpus includes 27 dialogues (672 turns) each of them annotated with FEELTRACE data by three annotators *dr*, *em*, *jd* [2]. The fourth annotator *cc* annotated only 23 out of 27 dialogues (569 turns). The FEELTRACE data are specified as coordinates in Osgood's Evaluation/Activation (E/A) space and are mapped onto 5 affect segments (see Fig. 1).

We extracted a subcorpus from SAL by using a majority vote strategy. A turn is only considered for the subcorpus if the FEELTRACE data of at least two judges correspond to the same affect segment at the end of the turn. Hence, 35 out of 672 utterances had to be discarded due to missing agreement between annotators. For

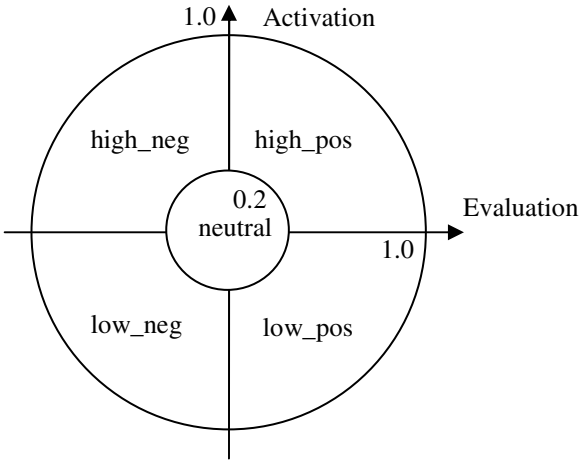

Fig. 1. Affect segmentation in the E/A space

remaining 637 utterances a majority vote could be obtained resulting in the following affect distribution – *high_neg*: 182 utterances, *low_neg*: 112 utterances, *neutral*: 139 utterances, *low_pos*: 24 utterances, *high_pos*: 180 utterances. Example utterances from the SAL corpus with the majority vote are shown in Fig. 2.

```
[1 - Affect segment: neutral] (Breath intake) Well,
I'll be able to have fun when I've done all the work,
but you see I have a very, er, heavy, difficult couple
of months ahead of me.

[2 - Affect segment: high_pos] (Laugh) I'm damn awful.
How are you (laugh)?

[3 - Affect segment: low_pos] Yup.

[4 - Affect segment: high_neg] Yes, that's not very
pleasant, is it?

[5 - Affect segment: low_neg] Erm, that's probably
true.
```

Fig. 2. Examples of SAL utterances

Note that utterance 1 is annotated as *neutral* even though it contains the words *fun, heavy, difficult* that usually indicate highly affective utterances, utterance 2 is annotated as *high_pos* despite of the words *damn* and *awful*, utterance 3 is annotated as *low_pos* despite of no affect words at all, utterance 4 is annotated as *high_neg* despite of the word *pleasant*, utterance 5 as *low_neg* despite of the word *true*. Furthermore, utterances 4 and 5 are very similar regarding their grammatical and lexical form, but still got different annotations (*high_neg* vs. *low_neg*). Summing up, the properties of utterances in the SAL corpus may be characterized as follows:

1. *Genre.* The SAL corpus consists of transcripted spontaneous spoken utterances. The spontaneous utterances in SAL may be grammatically incorrect and often contain repairs, repetitions and inexact words.
2. *Length of emotional text*. The length of a dialogue turn in SAL is variable and may consist of one word only.
3. *Annotations.* An annotation in SAL indicates the affect experienced by several test persons conversing with one of the SAL characters. Test persons and annotators in SAL are different people.

Due to the properties above, the SAL corpus presents a great challenge to computer-aided affect sensing if it is based on lexical features only. In contrast, human annotators relied in addition to linguistic features also on visual and acoustic features in order to annotate dialogue turns.

4 Feature Extraction and Evaluation

Features are computed automatically for dialogue turns making use of frequency lists as well as the affective qualities of words extracted from various dictionaries. In particular, we concentrate on the following features:

- *WORD FEATURES:* Each word in the dictionary represents a word feature to which a frequency value is assigned. The frequency value represents how many times the word occurs in a dialogue turn. The use of word features is based on the assumption that the frequency of certain words is characteristic of the affective tone of the dialogue turn. The number of word features depends on the size of the dictionary. For example, when using LIWC, we obtain 2,251 word features.
- *LIWC FEATURES:* For each LIWC category, we compute how frequently it occurs in the dialogue turn by counting the number of words that correspond to the category. Depending on whether we rely on all categories or just the affect-related categories, we obtain sets with 68 (CAT-68) or 8 features (CAT-8) respectively.
- *DAL FEATURES:* We compute the averaged scores for the emotive connotations (evaluation, activation, imagery). In sum, we get a set with three features (EA-AVG).

We did not normalize word or category counts with respect to text length since an earlier experiment did not show a significant effect of normalization on the recognition rate.

To find the most relevant features for emotion recognition, we conducted several experiments with subsets of features. In particular, we explored the following options for feature reduction.

Selection of the most frequent words as features

This option is based on the assumption that more common words are more suitable to discriminate between different emotional states expressed by a dialogue turn. Since the available dictionaries differ very much in size, we apply different selection strategies for them:

- For BNC, we extract 56 sets of word features: BNC-650, BNC-1400, …, BNC-72800 where *BNC-n* contains words with at least *n* occurrences in the BNC corpus. The values of *n* are selected in the manner that facilitates the comparison with the number of features in DAL. The frequency word list for the BNC corpus is freely available under [http://www.kilgarriff.co.uk/BNC_lists/all.al.gz].
- For SAL, we extract 95 sets of word features *SAL-n* (*n*=1…95) where *SAL-n* contains the *m/n* most frequent words from the SAL corpus and *m* = 2,051 is the number of words in SAL.

Selection of words with higher emotional expressivity
This option is based on the assumption that words that have higher activation and evaluation values or that can be categorized as emotion words are more discriminative than other words. We employ different strategies for the two available dictionaries:

- For DAL, we extract 40 sets of word features DAL-1, DAL-2, …, DAL-40 and corresponding DAL features (EA-AVG) where *DAL-n* contains all words from DAL with $\sqrt{activation^2 + evaluation^2} \geq \frac{n-1}{n}$. The distribution of the E/A space is illustrated in Fig. 3. For example, the complete set of DAL word features lies within the outmost circle while the dotted area corresponds to word features with higher emotional expressivity. Fig. 3 shows circle 3 (DAL-3) containing the word *brutality*.

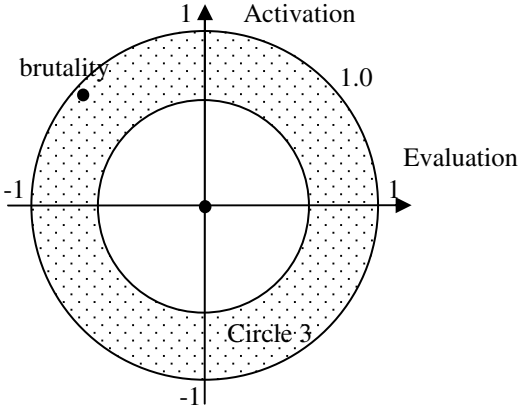

Fig. 3. Extraction of words from DAL

- For LIWC, we extract 2 sets of word features LIWC-68 and LIWC-8 where LIWC is the original set of 2,251 word features and LIWC-8 only contains 617 word features that correspond to the affect-related categories.

5 Results

Classification is done using SVM from the WEKA data mining toolkit [18]. Figures 4-6 and Table 1 below show the 10-fold cross-validation results that are

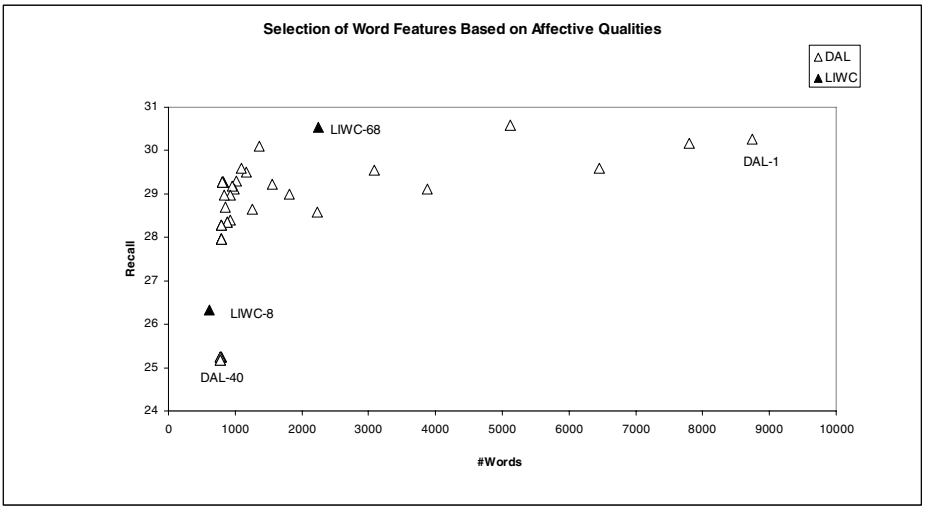

Fig. 4. Selection of Word Features Based on Affective Qualities

averaged over affect classes. The recall, precision, and fMeasure measures are expressed in percent.

We examine the aforementioned subsets of features with respect to their relevance to affect recognition in dialogue turns. In Fig. 4, we reduce the number of word features originating from the DAL and the LIWC dictionary based on their affective qualities. In Fig. 5, we reduce the number of word features based on their frequency. Here, we start from the general-purpose BNC corpus and frequency lists generated from the SAL corpus. In Fig. 6, we compare word features lists with and without features based on affective annotations/categories: word features from DAL that are selected according to their evaluation and activation values, the complete set of LIWC words with and without affective categories and a reduction of the LIWC words to affective LIWC words with and without affective categories. In Table 1, we list the affect sensing results based on features from affective annotations from DAL (EA-AVG), from LIWC (CAT-8, CAT-68) and their combinations. The first column indicates the number of features, the second column the names of the feature sets.

The recognition rates for the feature sets examined in this paper range from 21.70% recall (EA-AVG) to 36.20% recall (SAL-19). Compared to choice by chance for a 5-class problem (20%), the recognition rates are rather low. But as we have seen in Section 4, the corpus presents a great challenge to affect recognition since speakers do not always use expressive words that can be directly mapped onto emotions.

We obtained the best results for SAL-19 (recall 36.20%) – one of the dictionaries which was directly extracted from our corpus. We combined SAL-19 with the best LIWC categories and part of speech (POS) features from the BNC tagset, but it didn't improve recognition rates (36.20% vs. 32.84%).

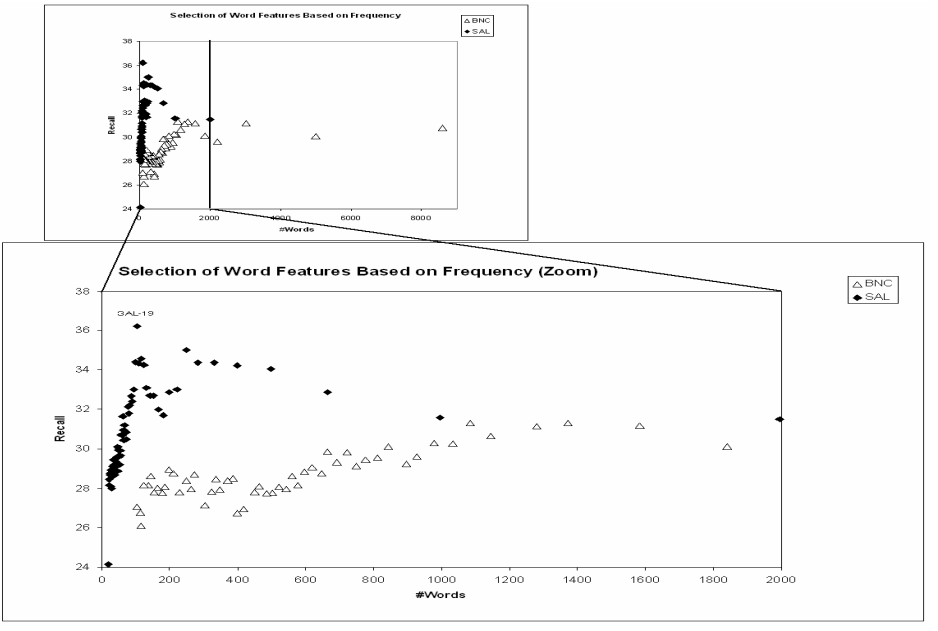

Fig. 5. Selection of Word Features Based on Frequency

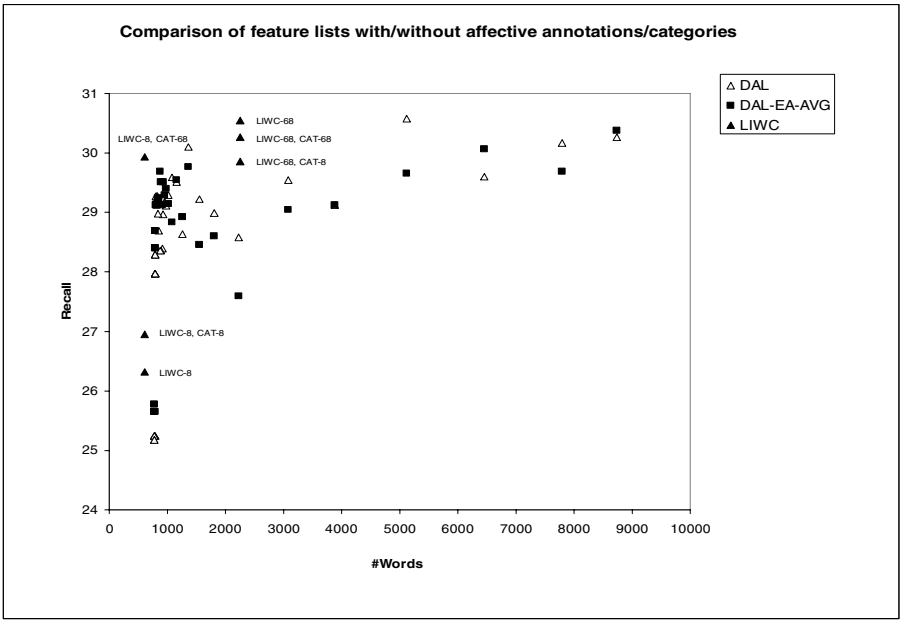

Fig. 6. Comparison of feature lists with/without affective annotation/categories

Table 1. Results of affect sensing using exclusively affective annotations/categories

#Features	Features	Recall	Precision	fMeasure
A: 71	CAT-68 + EA-AVG	28.54	28.89	26.42
A: 68	**CAT-68**	**28.90**	**30.37**	**26.83**
A: 3	EA-AVG	21.70	13.60	14.61
A: 11	CAT-8 + EA-AVG	23.92	14.57	16.72
A: 8	CAT-8	24.46	25.40	17.76
POS: 48	POS	24.20	22.01	19.76
W: 104, A: 68, POS: 48	**SAL-19, CAT-68, POS**	**32.84**	**34.46**	**32.66**

We assumed that classification results depend on the coverage – the ratio of words found both in the employed dictionary (or subdictionary) and the corpus divided by the number of words in the corpus. However, we did not find any distinct influence of the coverage on the classification rate.

In the following, we examine the recognition results in more detail to shed light on the questions addressed in Section 1.

6 Discussion

Results in the previous section allow for answering the questions in the introduction:

1. **Do we get higher recognition rates if we restrict ourselves to word features that convey emotional content?** Fig. 4 shows that a reduction of word features to the most expressive ones does not lead to a significant change in recognition results. For instance, we achieve a recall rate of 30.27 % for a set with nearly 9,000 word features (DAL-1) which slowly degrades to 25.18% (DAL-40) when restricting ourselves to a much smaller set of 779 more expressive word features that are selected based on their activation and evaluation values in the Whissell dictionary. Obviously, the choice of word features based on their emotional expressivity does not have a great impact on recognition rate, but may provide a useful criterion of feature reduction without risking a severe degradation of recognition rates. For instance, we get nearly the same recognition rates for sets with about 9000 word features and for sets with between 1300 and 1500 more expressive features.
2. **Do emotive annotations in dictionaries improve affect sensing for dialogue turns?** As Fig. 6 shows, the classification results do not change significantly when augmenting word features by features that are based on affective annotations from the DAL or affective categories of the LIWC dictionary. Obviously, affect-related features do not include discriminative information that is not yet included in the word counts. On the other hand, the results do not degrade dramatically when relying exclusively on affective annotations (see Table 1). For example, when using the affective annotations from the LIWC only (CAT-68) we get similar results, but just need to consider 68 features as opposed to several hundreds of features. A reduction of features is of major importance when analysing affect in a real-time application.

3. **Are common words more useful to affect sensing than less common words?** Fig. 5 shows that a reduction of the BNC lists based on their frequency does not lead to a dramatic change in recognition rates. Based on our experiments, it is hard to say whether a reduction of features should be based rather on the frequency of words or their expressive qualities. A comparison of Fig. 4 and Fig. 5 shows that datasets that are chosen on the basis of the expressed affect may be outperformed by datasets with words that are derived from the investigated corpus and chosen on the basis of frequencies. Nevertheless, a reduction of word features based on frequencies has to be taken with care. For instance, Wiebe and colleagues [17] found that rare words, especially hapax legomena (words occurring only once in a corpus) can be successfully used for affect sensing. According to their studies, we should not entirely exclude rare words from further consideration. However, as long as we do not have concrete knowledge regarding the predictive power of specific words, a reduction of word features based on their frequency seems to be reasonable.
4. **Are dictionaries of affect more useful to affect sensing than general-purpose dictionaries?** Our experiments show that general-purpose dictionaries may provide similar results as affect dictionaries for similar numbers of features. For instance, we achieve a recall of about 30% with the BNC and the DAL datasets when using around 5000 features. The best result – a recall of 36.20% – has been obtained for SAL-19 which just contains 104 word features resp. Here, we have to consider, however, that SAL-1 to SAL-95 have been specifically tailored to our corpus.

7 Conclusion

In this paper, we investigated the potential benefits of several dictionaries for affect sensing in dialogue. Our results indicate that lexical affect sensing may also be successfully conducted with a general-purpose dictionary. Obviously, affective annotations/categories in dictionaries do not provide much more information on the affective qualities of a dialogue turn in addition to word counts. On the other hand, the affective annotations seem to provide a good means to reduce the number of features for classification tasks. For example, when just considering the 68 affective word categories of LIWC, we get similar results as when using several thousands word features taken from an affect dictionary or a general-purpose dictionary. Such a reduction is of high significance when sensing affect in real-time.

We showed classification results for the SAL corpus as a corpus that presents a great challenge to affect sensing due to its properties, but we assume that our findings are also applicable for other "less difficult" corpora. For example, we studied a corpus with 215 movie reviews [www.reelviews.net] distributed over 5 emotional classes (40 movie reviews with zero rating, 25 movie reviews with four-star-rating, and 50 movie reviews each with other three ratings) and revealed similar trends (Fig. 7).

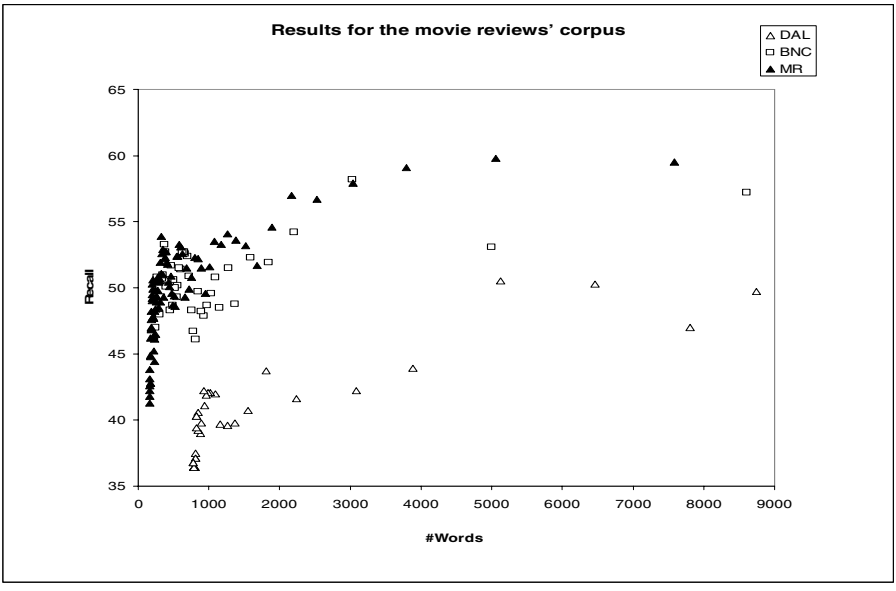

Fig. 7. Results for the movie reviews' corpus

MR in Fig. 7 represents word features' sets extracted from the frequency list of the movie reviews' corpus (an equivalence of the SAL feature sets). Noticeable is a significantly higher recognition rates compared with that in the SAL corpus (greater than 35%).

Of course, the results should be taken with care since our experiments are based on two corpora only. In our future work, we will therefore investigate to what extent the results may be generalized for other corpora. Furthermore, we will have a look at additional affect dictionaries, such as the WordNet-Affect Database.

Acknowledgments

This work was partially financed by the European Network of Excellence HUMAINE and the European CALLAS IP.

We are greatly indebted to Thurid Vogt for her useful comments and suggestions.

References

1. Coltheart, M.: The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology 33A, 497–505 (1981)
2. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: 'FEELTRACE': An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland, pp. 19–24 (2000)

3. Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Shriberg, E., Stolcke, A.: Distinguishing Deceptive from Non-Deceptive Speech, URL (2005), http://www.speech.cs.cmu.edu/awb/jss/IS051364.PDF
4. Kilgarriff, A.: Putting Frequencies in the Dictionary. International Journal of Lexicography 10(2), 135–155 (1997)
5. Kollias, S.: ERMIS Project. URL (2007), http://www.image.ntua.gr/ermis/
6. Langenmayr, A.: Sprachpsychologie. Hogrefe, Göttingen (1997)
7. Mairesse, F., Walker, M.: Words Mark the Nerds: Computational Models of Personality Recognition through Language. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006), Vancouver, pp. 543–548 (July 2006)
8. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the 2nd international Conference on Knowledge Capture, Sanibel Island, FL, USA (October 23 - 25). K-CAP '03, ACM Press, New York (2003)
9. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Analysis of Affect Expressed through the Evolving Language of Online Communication. In: Proc. of IUI 2007, pp. 278–281. ACM Press, New York (2007)
10. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count (LIWC): LIWC2001. Erlbaum Publishers, Mahwah, NJ (2001)
11. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology 54, 547–577 (2003)
12. Riloff, Ellen, Patwardhan, Siddharth, Wiebe, Janyce.: Feature Subsumption for Opinion Analysis. In: Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing. Sydney, AUS: Association for Computational Linguistics, pp. 440–448 (2006)
13. Shaikh, S., Islam, Md.T., Ishizuka, M., Prendinger, H.: Implementation of Affect Sensitive News Agent (ASNA) for affective classification of news summary. In: Proceedings International Conference on Computer and Information Technology (ICCIT-06), Dhaka, Bangladesh (2006)
14. Valitutti, A., Strapparava, C., Stock, O.: Developing Affective Lexical Resources. PsychNology Journal 2(1), 61–83 (2004)
15. Weintraub, W.: Verbal Behavior in Everyday Life. Springer, Heidelberg (1989)
16. Whissell, C.M.: The dictionary of affect in language. In: Plutchik, R., Kellerman, H. (eds.) Emotion: Theory, Research, and Experience, pp. 113–131. Academic Press, New York (1989)
17. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning Subjective Language. Computational linguistics 30(3), 277–308 (2004)
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
19. Zhang, L., Barnden, J.A., Hendley, R.J., Wallington, A.M.: Exploitation in Affect Detection in Open-Ended Improvisational Text. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text, ACM Press, New York (2006)