CONTINUOUS MODELS OF AFFECT FROM TEXT USING N-GRAMS

Nikolaos Malandrakis[1], *Alexandros Potamianos*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
[2]Dept. of ECE, Technical Univ. of Crete, 73100 Chania, Greece

malandra@usc.edu, potam@telecom.tuc.gr, shri@sipi.usc.edu

ABSTRACT

We propose a method of affective text analysis and modeling that is capable of generating continuous valence ratings at the sentence level starting from word and multi-word term valence ratings. Motivated from the language modeling literature, a back-off algorithm is employed to efficiently fuse the valence of single-word and multi-word terms. Specifically, a term detection criterion is used to select the appropriate n-gram terms, starting with bigrams and potentially backing off to unigrams. Term affective ratings are generated by a lexicon expansion method, using semantic similarity estimates computed on a large web corpus. The proposed framework provides state-of-the art results in the sentence level SemEval'07 task of news headline polarity detection, reaching an accuracy of 75%.

Index Terms— emotion, affect, affective lexicon, polarity detection, language understanding.

1. INTRODUCTION

The analysis of language affect, the emotional content of lexical information, is a significant sub-task of many applications in a variety of fields including sentiment analysis and emotion recognition from multimedia streams (audio, video, text) [1, 2], including multimedia content analysis through subtitles [3] and news headlines analysis [4].

Analysis can happen at various levels, targeting different lexical units: words, phrases, sentences etc. The most popular target in research have been sentences. For the most part research has revolved around hierarchical models of affect, usually combining affective ratings of words into affective ratings for sentences. Word ratings are provided by affective lexica, either manually annotated ones, like the General Inquirer [5] and Affective norms for English Words (ANEW) [6] or, more typically, automatically expanded ones like SentiWordNet [7] and WORDNET AFFECT [8].

The methods for combining word ratings into sentence ratings vary significantly. Usually a word selection step is involved, which removes non-relevant words from the sentence and keeps only those considered to carry affective significance. Word selection is typically done using part-of-speech tags [9] (content words), but also affective ratings and/or sentence structure has been used [10]. The actual combination of ratings is usually done by some simple numeric method, like taking the arithmetic mean. There have been attempts at incorporating more complex fusion rules into the process [9, 10], although such rules are usually specified manually. In [11], a supervised method is used to train the parameters of multiple hand-coded rules of (affective score) composition. However, up to now more complex fusion methods have shown little improvement over simpler distributional approaches and compositional fusion. There have been no attempts to use the ratings of multi-word terms [12], such as names, the meaning of which can not be directly expressed as a combination of the meaning of their parts, though there have been non-hierarchical approaches that take into account n-grams [13]. This is, in part, due to the lack of affective lexica that would provide the ratings for these terms.

We propose an affective model inspired by language modeling techniques, aiming to incorporate non-compositional aspects into a compositional framework. At the top level this is a compositional approach that combines ratings of terms into ratings for sentences using simple numerical methods; however these terms do not have to be simple words: they may also be multi-word terms. The method resembles a back-off bigram language model in that we use overlapping bigrams and (using term selection criteria) fall back to unigrams (words). Thus, phrases with non-compositional semantics are implicitly modeled via the expansion of the affective lexicon with n-grams. The proposed approach also serves as a workaround for the modeling of the affective content of simple syntactic rules, e.g., handling of negations. Affective ratings for both single- and multi-word terms are produced by a generalization of the affective lexicon expansion method presented in [14]. The ratings produced at every level (word, n-gram, sentence) are continuous valence/polarity scores.

Compared to prior research, our work differs in the way we handle the affective compositionality assumption. We neither assume that the affective ratings of a phrase or sentence can be composed from the affective ratings of all words included, nor do we focus on identifying non-compositional rules. Instead, we capture non-compositionality by explicitly estimating the affective scores of n-grams and then merge

these scores with a compositional lexical affective model.

2. CREATING N-GRAM RATINGS

The first part of the method is the generation of continuous valence ratings for all n-grams contained in each sentence. The method we use is a refinement of the one presented in [14], which in turn builds on [15]. We start from an existing, annotated lexicon. A subset of it is used as seed words and the affective ratings of new words/terms are all expressed as a weighted linear combination of their semantic similarities to these seed words multiplied with the affective ratings of the seeds:

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i \; v(w_i) \; d(w_i, w_j), \tag{1}$$

where w_j is the word whose affect we aim to characterize, $w_1 ... w_N$ are the N seed words, $v(w_i)$ is the valence rating for seed word w_i, a_i is the (trainable) weight corresponding to word w_i and $d(w_i, w_j)$ is a measure of semantic similarity between words w_i and w_j. Given an annotated corpus of K words and a set of $N < M$ seed words we can use (1) to create a system of M linear equations with $N+1$ unknown variables; the N weights $a_1 ... a_N$ and the extra weight a_0 which is the shift (bias). The values of these coefficients can be estimated using the least mean square (LMS) algorithm. Once the weights of the seed words are estimated the valence of an unseen word w_j can be computed using (1). For details see [14].

A context-based similarity metric $d()$ is used for the experiments presented in this work. Context-based similarity metrics compute similarity between feature vectors extracted from term context, i.e., using a "bag-of-words" context model. The metric we use computes cosine similarity between the context vectors of w_1 and w_2 as

$$S^K(w_1, w_2) = \frac{\sum_{i=1}^{L} t_{w_1,i} t_{w_2,i}}{\sqrt{\sum_{i=1}^{L} (t_{w_1,i})^2} \sqrt{\sum_{i=1}^{L} (t_{w_2,i})^2}} \tag{2}$$

where K the context window length, L the vocabulary size and $t_{w,i}$ a value representing the frequency of occurrence of vocabulary word v_i within the left or right context window K of term w. We use a binary weighting scheme, so all $t_{w,i}$ are zero or one depending on whether a word occurs within the context of w. For more details on the semantic similarity metrics and their performance on semantic similarity tasks see [16, 17].

The method has no requirement that limiting it to estimating word ratings or even limiting it to any specific language. To apply to bigrams, only the semantic similarity metric has to be extended to handle both unigrams and bigrams. The generalization is straightforward for context-based metrics and indeed such metrics have been successfully used to estimate the semantic similarity between multi-word terms [17].

3. CREATING SENTENCE RATINGS

Given a sentence $s = w_1 w_2 ... w_N$ we assume that its affective content can be expressed as a composition of the affective contents of the terms (words or otherwise) it contains [18]. In [14], we used three simple numerical methods to combine word ratings into sentence ratings. In this work, we use a modified version that includes trainable weights:

1. Linear fusion (average)

$$v_a(s) = b_0 + b_1 \; \frac{1}{N} \sum_{i=1}^{N} v(w_i) \tag{3}$$

2. Weighed linear fusion (valence weighted average)

$$v_w(s) = b_0 + \frac{b_1}{\sum_{i=1}^{N} |v(w_i)|} \sum_{i=1}^{N} v(w_i)^2 \cdot \text{sign}(v(w_i)) \tag{4}$$

3. Non-linear min-max fusion (max)

$$\begin{array}{rcl} v_m(s) & = & b_0 + b_1 \; v(w_z) \\ z & = & \arg\max_i (|v(w_i)|) \end{array} \tag{5}$$

Constants b_0 and b_1 are trainable weights corresponding to an offset and unigrams respectively. Such linear fusion methods implicitly make a compositionality assumption that we wish to relax: next, we propose expanding their definition to incorporate terms (of length n) instead of just words.

Our model, much like an n-gram language model, takes into account the partial ratings of all *overlapping* n-grams within a sentence. For a sentence $s = w_1 w_2 ... w_N$, we create both a unigram and a bigram affective model, λ_1 and λ_2 respectively, that estimate the sentence level affective score as follows[1]:

$$\begin{array}{rcl} v(s|\lambda_1) & = & \frac{1}{N} \sum_{i=1}^{N} v(w_i), \\ v(s|\lambda_2) & = & \frac{1}{N-1} \sum_{i=1}^{N-1} v(w_i w_{i+1}), \end{array} \tag{6}$$

where the valence $v(w_i)$ of word w_i and the valence $v(w_i w_{i+1})$ of bigram $w_i w_{i+1}$ are estimated using Eq. (1). Then we use a criterion for selecting between unigrams and bigrams. We define $c(i,j)$ as a selection criterion for the bigram $w_i w_j$ and use the bigram $w_i w_j$ if $c(i,j)$ is larger than some threshold t or back-off to the unigrams w_i, w_j otherwise.

$$v(w_i w_j) = \left\{ \begin{array}{ll} b_1 \; v(w_i w_j | \lambda_1), & \text{if } c(i,j) \leq t \\ b_2 \; v(w_i w_j | \lambda_2), & \text{if } c(i,j) > t \end{array} \right\}, \tag{7}$$

[1]The following equations correspond to the "average" model; however the same expansion can be used for the weighted average and max fusion models in (3)-(5).

where b_1 and b_2 are trainable weights of the unigram and bigram models. After performing term selection, we combine the scores:

$$v_b(s) = b_0 + \frac{1}{N-1}\sum_{i=1}^{N-1} v(w_i w_{i+1}) \quad (8)$$

where b_0 is an additional trainable weight of the equation and acts as an offset. An issue with this equation is that each word is counted twice, apart from the first and last one. We rectify it by explicitly adding back (with the appropriate weight) the unigram of the first and last word, leading to:

$$v_{bo}(s) = b_0 + \frac{1}{N}\left[\frac{b_1}{2}(v(w_1)+v(w_N)) + \sum_{i=1}^{N-1} v(w_i w_{i+1})\right] \quad (9)$$

The weighting factors b_0, b_1, b_2 are trained using LMS using a similar approach to the word-level model.

The criterion $c(i,j)$ used to select the appropriate n-gram model should in some way reflect the semantic or affective non-compositionality of the term in question. We present results for the following two criteria:

1. A semantic (non-)compositionality metric, operating on the co-occurrence probability $p(w_i, w_j)$ of w_i and w_j in text:

$$c_s(i,j) = p(w_i w_j)\ \log\frac{p(w_i w_j)}{p(w_i)p(w_j)}. \quad (10)$$

2. An affective (non-)compositionality metric, operating on the valence ratings generated by our model for the bigram and the unigrams, as follows:

$$c_a(i,j) = \left|v(w_i w_j) - 0.5[v(w_i) + v(w_j)]\right|. \quad (11)$$

4. EXPERIMENTAL PROCEDURE

The main word corpus we use to train the lexicon creation algorithm is the *Affective Norms for English Words* (ANEW) dataset. ANEW consists of 1034 words, rated in 3 continuous dimensions of arousal, valence and dominance (we only use valence). To train sentence-level models and evaluate their performance we use the *SemEval 2007: Task 14* corpus [4]. This corpus contains news headlines, split into a development set of 250 and a testing set of 1000. We use both subsets as intended: the development set to estimate the fusion parameters (b_0, b_1, b_2) and the testing set for evaluation. The headlines are manually annotated on a valence scale of $[-100, 100]$.

Computing the values of the similarity metric used in (1) requires a text corpus. The one we use is an accumulation of web data, gathered by submitting queries to the Yahoo! search engine. Specifically, we use the vocabulary of English packaged in the aspell spellchecker for English, containing 135433 words. For each of these words we pose an individual (IND) query to the Yahoo! search engine and from the response we collect the snippets (short representative excerpts of the document shown under each result) of the top 500 results. Each snippet is usually composed of two sentences: title and content. The corpus contains 116 million sentences. All contextual similarities between words or words and terms are calculated over this corpus.

Also important to the lexicon creation process is seed selection, the method of selecting the seed words from the candidates (the training set). Our selection method is unsupervised and based on two criteria: good seeds must have extreme valence ratings (positive or negative) and a good seed set should be balanced, the sum of valence ratings should be zero. So we start by sorting the positive and negative seeds separately by their valence rating. Then we add positive and negative seeds iteratively to the seed set so as to minimize the absolute value of the sum of their valence ratings and maximize their absolute valence ratings, until the required number of seed words N is reached.

The term affect model is trained using the ANEW corpus, using all $M = 1034$ words in it as training samples and a subset as seed words. That model is used to generate ratings for all unigrams and bigrams contained in the SemEval testing and development sets, creating a pool of candidate terms for each sentence, a bag-of-terms. It should be noted that with regards to unigram terms, we apply content word selection: part-of-speech tagging is performed using *TreeTagger* [19] and any unigrams that are not nouns, adjectives, verbs or adverbs are removed from the selection pools.

The final part of each experiment is the term selection and training of n-gram fusion parameters. Term selection criteria are calculated on the web data corpus (for c_s) and the automatically generated affective ratings (for c_a). Then we select a back-off threshold t, a value under which we fall back to unigrams. Given a criterion and a threshold we can select the relevant terms for the development and testing data. Finally, the n-gram fusion parameters are trained using LMS on the development set and the model is used to generate ratings for the evaluation dataset.

5. RESULTS

As a baseline, we conduct an experiment using the fusion schemes defined in equations (3), (4) and (5), utilizing only unigram terms. Polarity detection results (2-class classification accuracy) as a function of the number of seeds used by the word model are shown in Fig. 1, for the three fusion methods. The method is significantly different from the one used in [14]: the term rating method is more accurate and we are using supervised training for the sentence model. Overall that leads to an improvement of binary classification accuracy, from our best result of 70% in [14] to 72.8% here (for linear fusion). The simple average model performs better throughout our experiments. Best results are achieved over a wide

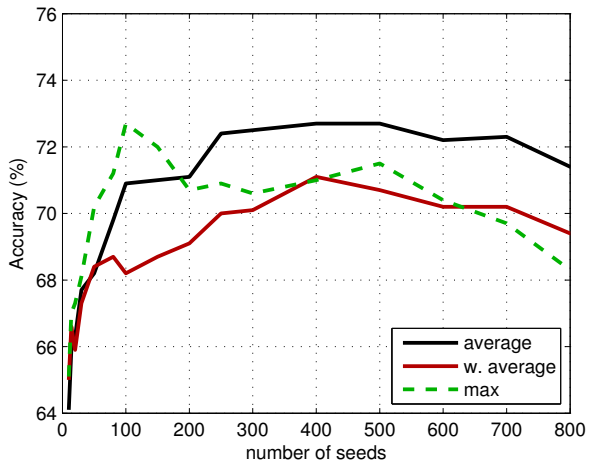

Fig. 1. Binary classification accuracy of the sentence rating algorithm as a function of the number of seed words, when using only unigram terms for linear and non-linear fusion.

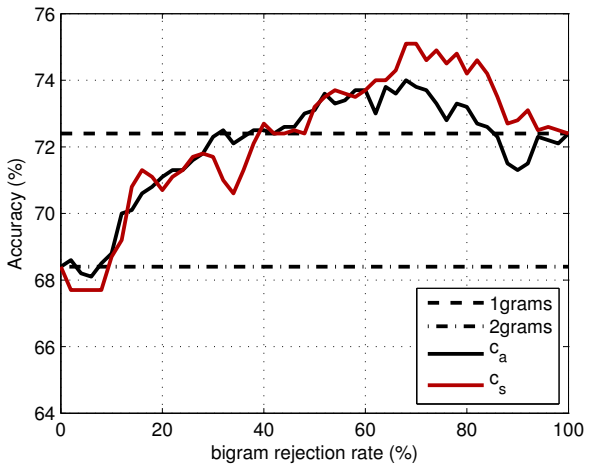

Fig. 2. Binary classification accuracy of the sentence rating algorithm as a function of the bigram selection threshold (bigram rejection rate).

range of number of seeds, from 300 to 600 seeds.

Next we present results for the n-gram model using the criteria described in Section 1 for term selection. We are interested in the performance of the model for various mixes of unigram and bigram terms, i.e., different values of the threshold t in (7). To investigate, we select a term affect model, trained on the ANEW dataset with 600 seed words, and use different term selection criteria and threshold t to select terms. For each value of the threshold t, the sentence model weights b_i are trained on the SemEval development set and the corresponding n-gram affective model is used to create ratings for the evaluation set sentences. The results are shown in Fig. 2 as a function of the bigram rejection rate (back-off rate).

The first thing to note are the two baseline scores, obtained when using all unigram or all bigram terms. They serve as an indirect comparison of the quality of unigram and bigram term ratings. Clearly the baseline performance of bigrams is lower than that of unigrams probably due to the lack of bigram seeds in the affective model and the (probably lower) performance of the semantic similarity metric when computing semantic similarities between bigrams and unigrams in (1). Despite the performance gap between unigram- and bigram-only models, combining the two using back-off significantly improves classification accuracy. As expected, the best performance is achieved for a mix that contains mainly unigrams (70% unigrams and 30% bigrams). The two term selection criteria detect terms in different ways, with c_s detecting non-compositional semantics and c_a detecting non-compositional affect. Both selection criteria beat the unigram-only baseline, however, the semantics-based criterion provides better performance (although the difference is not statistically significant). Although the affect-based criterion c_a does not perform as well, we expect its performance to improve as the performance of the term affective model (that provides the ratings for c_a) improves.

Overall, we observe a significant improvement over the unigram baseline: binary classification accuracy improves from 72.4% when using only unigrams to 75.1% at a back-off rate around 0.7, when using the c_s selection criterion. Comparable results in the literature are 62% [20], 66% [21], 71% [11] and 72.8% (using cross-validation) [22]. The model also achieved a correlation to the ground truth of 0.60, compared to 0.50 achieved by the best system in the literature [4].

6. CONCLUSIONS

Motivated by the language modeling literature, we proposed a method for creating sentence-level affective ratings by combining the affective ratings of n-gram terms. The method builds upon an affective lexicon expansion method capable of generating continuous affective ratings for multi-word terms that utilizes semantic similarity scores estimated on a web snippet corpus. Given affective ratings for single- and multi-word terms we proposed an algorithm for term selection and a supervised model for term rating fusion. The inclusion of bigram terms into a unigram-only model improved performance significantly, and the model achieved state-of-the-art performance on the SemEval task.

7. ACKNOWLEDGMENTS

The majority of this work was performed while Nikolaos Malandrakis was with the Dept. of ECE, Technical U. of Crete. This work was partially supported by the IST Programme of the EU under contract number 296170 (PortDial project).

8. REFERENCES

[1] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[2] C. M. Lee, S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. ICSLP*, 2002, pp. 873–876.

[3] A. Purandare and D. J. Litman, "Humor: Prosody analysis and automatic recognition for f*r*i*e*n*d*s*.," in *Proc. EMNLP*, 2006, pp. 208–215.

[4] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. SemEval*, 2007, pp. 70–74.

[5] P. Stone, D. Dunphy, M. Smith, and D. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press, 1966.

[6] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1," The Center for Research in Psychophysiology, University of Florida, 1999.

[7] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. LREC*, 2006, pp. 417–422.

[8] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," in *Proc. LREC*, 2004, vol. 4, pp. 1083–1086.

[9] F.-R. Chaumartin, "UPAR7: A knowledge-based system for headline sentiment tagging," in *Proc. SemEval*, 2007, pp. 422–425.

[10] A. Andreevskaia and S. Bergler, "CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging," in *Proc. SemEval*, 2007, pp. 117–120.

[11] K. Moilanen, S. Pulman, and Y. Zhang, "Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression," in *Proc. WASSA Workshop at ECAI*, 2010, pp. 36–43.

[12] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for NLP," in *Computational Linguistics and Intelligent Text Processing*, vol. 2276 of *Lecture Notes in Computer Science*, pp. 189–206. 2002.

[13] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech and Language*, vol. 25, no. 1, pp. 29–44, 2011.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Kernel models for affective lexicon creation," in *Proc. Interspeech*, 2011, pp. 2977–2980.

[15] P. Turney and M. L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. Technical report ERC-1094 (NRC 44929)," National Research Council of Canada, 2002.

[16] E. Iosif, A. Tegos, A. Pangos, E. Fosler-Lussier, and A. Potamianos, "Combining statistical similarity measures for automatic induction of semantic classes," in *Proc. IEEE/ACL Workshop Spoken Language Technology*, 2006, pp. 86–89.

[17] E. Iosif and A. Potamianos, "Unsupervised Semantic Similarity Computation Between Terms Using Web Documents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1637–1647, 2010.

[18] F. J. Pelletier, "The principle of semantic compositionality," *Topoi*, vol. 13, pp. 11–24, 1994.

[19] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. International Conference on New Methods in Language Processing*, 1994, vol. 12, pp. 44–49.

[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 1, pp. 1–41, 2010.

[21] K. Moilanen and S. Pulman, "Sentiment Composition," in *Proc. RANLP*, 2007, pp. 378–382.

[22] J. Carrillo de Albornoz, L. Plaza, and P. Gervs, "A hybrid approach to emotional sentence polarity and intensity classification," in *Proc. CoNLL*, 2010, pp. 153–161.