A Shallow Approach to Subjectivity Classification

Stephan Raaijmakers & Wessel Kraaij
TNO Information and Communication Technology
Delft
The Netherlands
{stephan.raaijmakers,wessel.kraaij}@tno.nl

Abstract

We present a shallow linguistic approach to subjectivity classification. Using multinomial kernel machines, we demonstrate that a data representation based on counting character n-grams is able to improve on results previously attained on the MPQA corpus using word-based n-grams and syntactic information. We compare two types of string-based representations: key substring groups and character n-grams. We find that word-spanning character n-grams significantly reduce the bias of a classifier, and boost its accuracy.[1]

Introduction

Subjectivity classification involves the discrimination between subjective and objective utterances, like sentences, or even phrases. Subjective utterances reflect a private point of view, emotion or belief. Recognition of subjectivity is important from several points of view. Pang and Lee (2004) have shown that removing objective sentences from text prior to applying sentiment classification yields higher classification accuracy. Subjectivity classification is important for product review mining (Kim and Hovy (2006)). Both summarization and information extraction (Stoyanov and Cardie (2006); Riloff et al. (2005)) benefit from an adequate discrimination between subjective and objective content. Stamatos (2006) has shown that shallow linguistic representations capture important linguistic aspects of utterances. The LingPipe suite[2] uses character n-grams for sentiment classification with good results. We investigate character n-grams for subjectivity classification.

Character n-grams

For the sentence 'This car really rocks' subword character bigrams and trigrams ('subgrams') are th, hi, is, ca, ar, re, ea, al, ll, ly , ro, oc, ck, ks, thi, his, car, rea, eal, all, lly, roc, ock, cks. A simple way of introducing sequentiality to a character-based n-grams approach is to employ n-grams on the subword level that span word boundaries and therefore reach the superword level. For instance, a bigram and trigram representation that spans word boundaries produces th, hi, is, s<sp>, <sp>c, ca, ar, r<sp>, <sp>r, re, ea, al, ll, ly, y<sp>, <sp>r, ro, oc, ck, ks, thi, his, is<sp>, s<sp>c, <sp>ca, car, ar<sp>, r<sp>r, <sp>re, rea, eal, all, lly, ly<sp>, y<sp>r, <sp>ro, roc, ock, cks with <sp> a whitespace indicator. These word-spanning n-grams ('supergrams') capture transitions between consecutive words, and thus encode phrasal effects on character level. [3] [4]

Data and experiments

Our data consists of the MPQA ((2002)) corpus, consisting of 535 news articles from 187 foreign and US sources, annotated for sentiment according to the Multi- Perspective Question Answering annotation scheme (Wiebe at al. (2005); Riloff et al. (2006)[5] We used the following shallow representations: subword character n-grams (bi-, tri- and quadrigrams); superword character n-grams (bi-, tri- and quadrigrams); key substring groups (Zhang and Lee (2006)); a mixture of key substring groups and superword character n-grams (bi-, tri- and quadrigrams). Word-internal character n-grams were used as a baseline. Key substring group features were generated with standard parameter settings of the software made available by Zhang (2006). [6] In our experiments, we use a simple, hyperparameter-free multinomial kernel: the negative geodesic kernel NGD (Zhang et al. (2005)): $K_{NGD}(x,y) = -2\arccos\left(\sum_{i=1}^{n}\sqrt{x_i y_i}\right)$

[1]This work is supported by the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

[2]Available from http://www.alias-i.com/lingpipe/

[3]Notice that the amount of string data increases significantly (39 vs. 24 character n-grams). For w words, the expansion factor for bigram and trigram superword character n-grams is $2(w-1)+3(w-1)=5w-5$ extra strings.

[4]As character n-grams do not encode positional information, attenuation arises naturally from string overlap, since similar n-grams coming from different words are considered to be the same.

[5]These authors kindly made available their exact data fragment from the MPQA corpus.

[6]Due to an inherent memory restriction of this software, we had to (uniformly) limit training set size over all runs and data representations to 5,000 datapoints, which amounts to 81% of the original training data (6,172 training points).

Results

Results in table 1 show that superword character n-grams perform the best. Even on the basis of 81% of the training data, this representation also leads to improvement of results reported on this data by Riloff et al. (2006). For a bias-variance decomposition of the classification error, we used the definition of bias and variance proposed by Kohavi and Wolpert (1996). Following Webb (2000) we applied 10 runs of 3-fold cross-validation, in order to get better estimates of bias and variance. From table 1, we see that supergrams yield lower bias than subgrams and key substring groups. The combination of key substring groups and superword character n-grams has a slightly lower bias than superword character n-grams alone, but it is unclear if this difference is significant. Overall performance of this combination is lower than the performance of superword character n-grams alone. The variance produced by supergrams is lower than subgrams, but slightly higher than for KSG or the combination of KSG and supergrams. The combined advantage of lower bias and lower or comparable variance of supergrams compared to subgrams and KSG-based representations is quite clear.

Bias decomposition				
	SUB	SUPER	KSG	KSG + SUPER
	39.8	35.9	37.9	35.8
Variance decomposition				
	SUB	SUPER	KSG	KSG + SUPER
	27.9	27.7	27.5	27.6
Average accuracy, recall, precision and F_1				
	SUB	SUPER	KSG	KSG + SUPER
Acc	74.57	**82.5**	77.9	81.9
Rec	77.6	**84.9**	80.7	84.4
Prec	75.9	**83.2**	78.9	82.5
F_1	76.7	**84**	79.8	83.5

Table 1: Results (best scores in bold).

Related work

Riloff et al. (2006) address subjectivity classification for MPQA. They report a best accuracy of 74.9% (three-fold cross-validation) using either a combination of unigrams and bigrams, or unigrams, bigrams together with syntactic extraction patterns. Li et al. (2007) report an F-measure of 77.9 on a single training-test split of the MPQA corpus, using support vector machines and raw token unigrams, lemma unigrams and part of speech information.

Conclusions

We compared character n-gram representations with key substring group representations, and provided empirical evidence for the superior performance of superword character n-grams: character n-grams that surpass word boundaries. We found that these n-grams significantly reduce the bias of a classifier, and also outperform the combination of shallow and deep linguistic features used by Riloff et al. (2006).

Acknowledgments

We are indebted to Janyce Wiebe, Theresa Wilson, Dell Zhang and the anonymous reviewers for valuable feedback and suggestions.

References

Kim, S.-M., and Hovy, E. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 483–490. Association for Computational Linguistics.

Kohavi, R., and Wolpert, D. H. 1996. Bias plus variance decomposition for zero-one loss functions. In Saitta, L., ed., *Machine Learning: Proceedings of the Thirteenth International Conference*, 275–283. Morgan Kaufmann.

Li, Y.; Bontcheva, K.; and Cunnigham, H. 2007. Experiments of opinion analysis on the corpora MPQA and NTCIR-6. In *Proceedings of NTCIR-6 Workshop Meeting, May 15-18, Tokyo, Japan*, 323–329.

MPQA. 2002. The MPQA corpus, version 1.2, available from http://www.cs.pitt.edu/~wiebe/pubs/pub1.html.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 271–278.

Riloff, E.; Patwardhan, S.; and Wiebe, J. 2006. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 440–448. Sydney, Australia: Association for Computational Linguistics.

Riloff, E.; Wiebe, J.; and Phillips, W. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings 20th National Conference on Artificial Intelligence (AAAI-2005)*, 1106–1111.

Stamatatos, E. 2006. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd Int. Workshop on Text-based Information Retrieval (TIR'06)*, 41–46.

Stoyanov, V., and Cardie, C. 2006. Toward opinion summarization: Linking the sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 9–14. Association for Computational Linguistics.

Webb, G. I. 2000. MultiBoosting: A technique for combining Boosting and Wagging. *Machine Learning* 40(2):159–196.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2-3):165–210.

Zhang, D., and Lee, W. S. 2006. Extracting key-substring-group features for text classification. In *Proceedings KDD'06*, 474–483.

Zhang, D.; Chen, X.; and Lee, W. S. 2005. Text classification with kernels on the multinomial manifold. In *Proceedings SIGIR'05*, 266–273.

Zhang, D. 2006. Key substring group software; available from http://www.dcs.bbk.ac.uk/~dell/publications/dellzhang_kdd2006_supplement.html.