Identifying Perspectives at the Document and Sentence Levels Using Statistical Models

Wei-Hao Lin*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
whlin@cs.cmu.edu

Abstract

In this paper we investigate the problem of identifying the perspective from which a document was written. By perspective we mean a point of view, for example, from the perspective of Democrats or Republicans. Can computers learn to identify the perspective of a document? Furthermore, can computers identify which sentences in a document strongly convey a particular perspective? We develop statistical models to capture how perspectives are expressed at the document and sentence levels, and evaluate the proposed models on a collection of articles on the Israeli-Palestinian conflict. The results show that the statistical models can successfully learn how perspectives are reflected in word usage and identify the perspective of a document with very high accuracy.

1 Introduction

In this paper we investigate the problem of automatically identifying the *perspective* from which a document was written. By perspective, we mean "subjective evaluation of relative significance, a point-of-view." For example, documents about the Palestinian-Israeli conflict may appear to be about the same topic, but reveal different perspectives:

(1) The inadvertent killing by Israeli forces of Palestinian civilians – usually in the course of shooting at Palestinian terrorists – is considered no different at the moral and ethical level than the deliberate targeting of Israeli civilians by Palestinian suicide bombers.

(2) In the first weeks of the Intifada, for example, Palestinian public protests and civilian demonstrations were answered brutally by Israel, which killed tens of unarmed protesters.

Example 1 is written from a Israeli perspective; Example 2 is written from a Palestinian perspective . We aim to address a research question: can computers learn to identify the perspective of a document given a training corpus of documents that are written from different perspectives?

When an issue is discussed from different perspectives, not every sentence in a document strongly reflects the perspective the author possesses. For example, the following sentences are written by one Palestinian and one Israeli:

(3) The Rhodes agreements of 1949 set them as the ceasefire lines between Israel and the Arab states.

(4) The green line was drawn up at the Rhodes Armistice talks in 1948-49.

Example 3 and 4 both factually introduce the background of the issue of the "green line" without expressing explicit perspectives. Can computers automatically discriminate between sentences that strongly express a perspective and sentences that only reflect shared background information?

*This is joint work with Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann, and supported by the Advanced Research and Development Activity (ARDA) under contract number NBCHC040037.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 227–230,
New York, June 2006.

A system that can automatically identify the perspective from which a document written will be a highly desirable tool for people analyzing huge collections of documents from different perspectives. An intelligence analyst regularly monitors the positions that foreign countries take on political and diplomatic issues. A media analyst frequently surveys broadcast news, newspapers, and web blogs for different viewpoints. What these analysts need in common is that they would like to find evidence of strong statements of differing perspectives, while ignoring statements without strong perspectives as less interesting.

In this paper we approach the problem of learning perspectives in a statistical learning framework. We develop statistical models to learn how perspectives are reflected in word usage, and evaluate the models by measuring how accurately they can predict the perspectives of unseen documents. Lacking annotation on how strongly individual sentences convey a particular perspective in our corpus poses a challenge on learning sentence-level perspectives. We propose a novel statistical model, Latent Sentence Perspective Model, to address the problem.

2 Related Work

Identifying the perspective from which a document is written is a subtask in the growing area of automatic opinion recognition and extraction. Subjective language is used to express opinions, emotions, and sentiments. So far research in automatic opinion recognition has primarily addressed learning subjective language (Wiebe et al., 2004; Riloff et al., 2003; Riloff and Wiebe, 2003), identifying opinionated documents (Yu and Hatzivassiloglou, 2003) and sentences (Yu and Hatzivassiloglou, 2003; Riloff et al., 2003; Riloff and Wiebe, 2003), and discriminating between positive and negative language (Yu and Hatzivassiloglou, 2003; Turney and Littman, 2003; Pang et al., 2002; Dave et al., 2003; Nasukawa and Yi, 2003; Morinaga et al., 2002).

Although by its very nature we expect much of the language of presenting a perspective or point-of-view to be subjective, labeling a document or a sentence as subjective is not enough to identify the perspective from which it is written. Moreover, the ideology and beliefs authors possess are often expressed in ways more than conspicuous positive or negative language toward specific targets.

3 Corpus

Our corpus consists of articles published on the `bitterlemons` website[1]. The website is set up to "contribute to mutual understanding [between Palestinians and Israels] through the open exchange of ideas". Every week an issue about Israeli-Palestinian conflict is selected for discussion, for example, "Disengagement: unilateral or coordinated?", and a Palestinian editor and an Israeli editor contribute a article addressing the issue. In addition, the Israeli and Palestinian editors invite or interview one Israeli and one Palestinian to express their views, resulting in a total of four articles in a weekly edition.

We evaluate the subjectivity of each sentence using the patterns automatically extracted from foreign news documents (Riloff and Wiebe, 2003), and find that 65.6% of Palestinian sentences and 66.2% of Israeli sentences are classified as subjective. The high but almost equivalent percentages of subjective sentences from two perspectives supports our observation in Section 2 that perspective is largely expressed in subjective language but subjectivity ratio is not necessarily indicative of the perspective of a document.

4 Statistical Modeling of Perspectives

We approach the problem of learning perspectives in a statistical learning framework. Denote a training corpus as pairs of documents W_n and their perspectives labels D_n, $n = 1, \ldots, N$, N is the total number of documents in the corpus. Given a new document $\tilde{W}$ with a unknown document perspective $\tilde{D}$, identifying its perspective is to calculate the following conditional probability,

$$P(\tilde{D}|\tilde{W}, \{D_n, W_n\}_{n=1}^{N}) \quad (5)$$

We are interested in how strongly each sentence in the document convey perspective. Denote the intensity of the m-th sentence of the n-th document as a binary random variable $S_{m,n}$, $m = 1, \ldots, M_n$, M_n is the total number of sentences of the n-th document. Evaluating how strongly a sentence conveys

[1] `http://www.bitterlemons.org`

a particular perspective is to calculate the following conditional probability,

$$P(S_{m,n}|\{D_n, W_n\}_{n=1}^{N}) \quad (6)$$

4.1 Document Perspective Models

The process of generating documents from a particular perspective is modeled as follows,

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ W_n &\sim \text{Multinomial}(L_n, \theta_d) \end{aligned}$$

The model is known as naïve Bayes models (NB), which has been widely used for NLP tasks such as text categorization (Lewis, 1998). To calculate (5) under NB in a full Bayesian manner is, however, complicated, and alternatively we employ Markov Chain Monte Carlo (MCMC) methods to simulate samples from the posterior distributions.

4.2 Latent Sentence Perspective Models

We introduce a new binary random variables, S, to model how strongly a perspective is expressed at the sentence level. The value of S is either s^1 or s^0, where s^1 means the sentence is written strongly from a perspective, and s^0 is not. The whole generative process is modeled as follows,

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \tau &\sim \text{Beta}(\alpha_\tau, \beta_\tau) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ S_{m,n} &\sim \text{Binomial}(1, \tau) \\ W_{m,n} &\sim \text{Multinomial}(L_{m,n}, \theta) \end{aligned}$$

π and θ carry the same semantics as those in NB. S is naturally modeled as a binary variable, where τ is the parameter of S and represents how likely a perspective is strongly expressed at the sentence given on the overall document perspective. We call this model **Latent Sentence Perspective Models** (LSPM), because S is never directly observed in either training or testing documents and need to be inferred. To calculate (6) under LSPM is difficult. We again resort to MCMC methods to simulate samples from the posterior distributions.

5 Experiments

5.1 Identifying Perspectives at the Document Level

To objectively evaluate how well naïve Bayes models (NB) learn to identify perspectives expressed at the document level, we train NB against on the `bitterlemons` corpus, and evaluate how accurately NB predicts the perspective of a unseen document as either Palestinian or Israeli in ten-fold cross-validation manner. The average classification accuracy over 10 folds is reported. We compare three different models, including NB with two different inference methods and Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000). NB-B uses full Bayesian inference and NB-M uses Maximum a posteriori (MAP).

Model	Data Set	Accuracy	Reduction
Baseline		0.5	
SVM	Editors	0.9724	
NB-M	Editors	0.9895	61%
NB-B	Editors	0.9909	67%
SVM	Guests	0.8621	
NB-M	Guests	0.8789	12%
NB-B	Guests	0.8859	17%

Table 1: Results of Identifying Perspectives at the Document Level

The results in Table 1 show that both NB and SVM perform surprisingly well on both Editors and Guests subsets of the `bitterlemons` corpus. We also see that NBs further reduce classification errors even though SVM already achieves high accuracy. By considering the full posterior distribution NB-B further improves on NB-M, which performs only point estimation. The results strongly suggest that the word choices made by authors, either consciously or subconsciously, reflect much of their political perspectives.

5.2 Identifying Perspectives at the Sentence Level

In addition to identify the perspectives of a document, we are interested in which sentences in the document strongly convey perspectives. Although the posterior probability that a sentence

covey strongly perspectives in (6) is of our interest, we can not directly evaluate their quality due to the lack of golden truth at the sentence level. Alternatively we evaluate how accurately LSPM predicts the perspective of a document, in the same way of evaluating SVM and NB in the previous section. If LSPM does not achieve similar identification accuracy after modeling sentence-level information, we will doubt the quality of predictions on how strongly a sentence convey perspective made by LSPM.

Model	Training	Testing	Accuracy
Baseline			0.5
NB-M	Guest	Editor	0.9327
NB-B	Guest	Editor	0.9346
LSPM	Guest	Editor	0.9493
NB-M	Editors	Guests	0.8485
NB-B	Editors	Guests	0.8585
LSPM	Guest	Editor	0.8699

Table 2: Results of Perspective Identification at the Sentence Level

The experimental results in Table 2 show that the LSPM achieves similarly or even slightly better accuracy than those of NBs, which is very encouraging and suggests that the proposed LSPM closely match how perspectives are expressed at the document and sentence levels. If one does not explicitly model the uncertainty at the sentence level, one can train NB directly against the sentences to classify a sentence into Palestinian or Israeli perspective. We obtain the accuracy of 0.7529, which is much lower than the accuracy previously achieved at the document level. Therefore identifying perspective at the sentence level is much harder than at that the document level, and the high accuracy of identifying document-level perspectives suggests that LPSM closely captures the perspectives expressed at the document and sentence levels, given individual sentences are very short and much less informative about overall perspective.

6 Summary of Contributions

In this paper we study the problem of learning to identify the perspective from which a text was written at the document and sentence levels. We show that perspectives are expressed in word usage, and statistical learning algorithms such as SVM and naïve Bayes models can successfully uncover the word patterns chosen by authors from different perspectives. Furthermore, we develop a novel statistical model to infer how strongly a sentence convey perspective without any labels. By introducing latent variables, Latent Sentence Perspective Models are shown to capture well how perspectives are reflected at the document and sentence levels. The proposed statistical models can help analysts sift through a large collection of documents written from different perspectives. The unique sentence-level perspective modeling can automatically identify sentences that are strongly representative of the perspective of interest, and we plan to manually evaluate their quality in the future work.

References

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*.

David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the European Conference on Machine Learning (ECML)*.

S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. 2002. Mining product reputations on the web. In *Proceedings of the 2002 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

T. Nasukawa and J. Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003).*

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).*

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003).*

Peter Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3).

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003).*