Extracting Opinion Expressions and Their Polarities – Exploration of Pipelines and Joint Models

Richard Johansson and **Alessandro Moschitti**
DISI, University of Trento
Via Sommarive 14, 38123 Trento (TN), Italy
{johansson, moschitti}@disi.unitn.it

Abstract

We investigate systems that identify opinion expressions and assigns polarities to the extracted expressions. In particular, we demonstrate the benefit of integrating opinion extraction and polarity classification into a joint model using features reflecting the global polarity structure. The model is trained using large-margin structured prediction methods.

The system is evaluated on the MPQA opinion corpus, where we compare it to the only previously published end-to-end system for opinion expression extraction and polarity classification. The results show an improvement of between 10 and 15 absolute points in F-measure.

1 Introduction

Automatic systems for the analysis of opinions expressed in text on the web have been studied extensively. Initially, this was formulated as a coarse-grained task – locating opinionated documents – and tackled using methods derived from standard retrieval or categorization. However, in recent years there has been a shift towards a more detailed task: not only finding the text expressing the opinion, but also analysing it: *who* holds the opinion and to *what* is addressed; it is positive or negative (*polarity*); what its *intensity* is. This more complex formulation leads us deep into NLP territory; the methods employed here have been inspired by information extraction and semantic role labeling, combinatorial optimization and structured machine learning.

A crucial step in the automatic analysis of opinion is to mark up the *opinion expressions*: the pieces of text allowing us to infer that someone has a particular feeling about some topic. Then, opinions can be assigned a *polarity* describing whether the feeling is positive, neutral or negative. These two tasks have generally been tackled in isolation. Breck et al. (2007) introduced a sequence model to extract opinions and we took this one step further by adding a reranker on top of the sequence labeler to take the global sentence structure into account in (Johansson and Moschitti, 2010b); later we also added holder extraction (Johansson and Moschitti, 2010a). For the task of classifiying the polarity of a given expression, there has been fairly extensive work on suitable classification features (Wilson et al., 2009).

While the tasks of expression detection and polarity classification have mostly been studied in isolation, Choi and Cardie (2010) developed a sequence labeler that simultaneously extracted opinion expressions and assigned polarities. This is so far the only published result on joint opinion segmentation and polarity classification. However, their experiment lacked the obvious baseline: a standard pipeline consisting of an expression identifier followed by a polarity classifier.

In addition, while theirs is the first end-to-end system for expression extraction with polarities, it is still a sequence labeler, which, by construction, is restricted to use simple local features. In contrast, in (Johansson and Moschitti, 2010b), we showed that global structure matters: opinions interact to a large extent, and we can learn about their interactions on the opinion level by means of their interactions on the syntactic and semantic levels. It is intuitive that this should also be valid when polarities enter the

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 101–106,
Portland, Oregon, June 19-24, 2011.

picture – this was also noted by Choi and Cardie (2008). Evaluative adjectives referring to the same evaluee may cluster together in the same clause or be dominated by a verb of categorization; opinions with opposite polarities may be conjoined through a contrastive discourse connective such as *but*.

In this paper, we first implement two strong baselines consisting of pipelines of opinion expression segmentation and polarity labeling and compare them to the joint opinion extractor and polarity classifier by Choi and Cardie (2010). Secondly, we extend the global structure approach and add features reflecting the polarity structure of the sentence. Our systems were superior by between 8 and 14 absolute F-measure points.

2 The MPQA Opinion Corpus

Our system was developed using version 2.0 of the MPQA corpus (Wiebe et al., 2005). The central building block in the MPQA annotation is the *opinion expression*. Opinion expressions belong to two categories: Direct subjective expressions (DSEs) are explicit mentions of opinion whereas expressive subjective elements (ESEs) signal the attitude of the speaker by the choice of words. Opinions have two features: *polarity* and *intensity*, and most expressions are also associated with a *holder*, also called *source*. In this work, we only consider polarities, not intensities or holders. The polarity takes the values POSITIVE, NEUTRAL, NEGATIVE, and BOTH; for compatibility with Choi and Cardie (2010), we mapped BOTH to NEUTRAL.

3 The Baselines

In order to test our hypothesis against strong baselines, we developed two pipeline systems. The first part of each pipeline extracts opinion expressions, and this is followed by a multiclass classifier assigning a polarity to a given opinion expression, similar to that described by Wilson et al. (2009).

The first of the two baselines extracts opinion expressions using a sequence labeler similar to that by Breck et al. (2007) and Choi et al. (2006). Sequence labeling techniques such as HMMs and CRFs are widely used for segmentation problems such as named entity recognition and noun chunk extraction. We trained a first-order labeler with the discriminative training method by Collins (2002) and used common features: words, POS, lemmas in a sliding window. In addition, we used *subjectivity clues* extracted from the lexicon by Wilson et al. (2005).

For the second baseline, we added our opinion expression reranker (Johansson and Moschitti, 2010b) on top of the expression sequence labeler.

Given an expression, we use a classifier to assign a polarity value: positive, neutral, or negative. We trained linear support vector machines to carry out this classification. The problem of polarity classification has been studied in detail by Wilson et al. (2009), who used a set of carefully devised linguistic features. Our classifier is simpler and is based on fairly shallow features: words, POS, subjectivity clues, and bigrams inside and around the expression.

4 The Joint Model

We formulate the opinion extraction task as a structured prediction problem $\hat{y} = \arg\max_y \ w \cdot \Phi(x, y)$. where w is a weight vector and Φ a feature extractor representing a sentence x and a set y of polarity-labeled opinions. This is a high-level formulation – we still need an inference procedure for the $\arg\max$ and a learner to estimate w on a training set.

4.1 Approximate Inference

Since there is a combinatorial number of ways to segment a sentence and label the segments with polarities, the tractability of the $\arg\max$ operation will obviously depend on whether we can factorize the problem for a particular Φ.

Choi and Cardie (2010) used a Markov factorization and could thus apply standard sequence labeling with a Viterbi $\arg\max$. However, in (Johansson and Moschitti, 2010b), we showed that a large improvement can be achieved if *relations* between possible expressions are considered; these relations can be syntactic or semantic in nature, for instance. This representation breaks the Markov assumption and the $\arg\max$ becomes intractable. We instead used a reranking approximation: a Viterbi-based sequence tagger following Breck et al. (2007) generated a manageable hypothesis set of complete segmentations, from which the reranking classifier picked one hypothesis as its final output. Since the set is small, no particular structure assumption (such

as Markovization) needs to be made, so the reranker can in principle use features of arbitrary complexity.

We now adapt that approach to the problem of joint opinion expression segmentation and polarity classification. In that case, we not only need hypotheses generated by a sequence labeler, but also the polarity labelings output by a polarity classifier. The hypothesis generation thus proceeds as follows:

- For a given sentence, let the base sequence labeler generate up to k_s sequences of unlabeled opinion expressions;
- for every sequence, apply the base polarity classifier to generate up to k_p polarity labelings.

Thus, the hypothesis set size is at most $k_s \cdot k_p$. We used a k_s of 64 and a k_p of 4 in all experiments.

To illustrate this process we give a hypothetical example, assuming $k_s = k_p = 2$ and the sentence *The appeasement emboldened the terrorists*. We first generate the opinion expression sequence candidates:

The [appeasement] emboldened the [terrorists]
The [appeasement] [emboldened] the [terrorists]

and in the second step we add polarity values:

The [appeasement]$_-$ emboldened the [terrorists]$_-$
The [appeasement]$_-$ [emboldened]$_+$ the [terrorists]$_-$
The [appeasement]$_0$ emboldened the [terrorists]$_-$
The [appeasement]$_-$ [emboldened]$_0$ the [terrorists]$_-$

4.2 Features of the Joint Model

The features used by the joint opinion segmenter and polarity classifier are based on pairs of opinions: basic features extracted from each expression such as polarities and words, and relational features describing their interaction. To extract relations we used the parser by Johansson and Nugues (2008) to annotate sentences with dependencies and shallow semantics in the PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004) frameworks.

Figure 1 shows the sentence *the appeasement emboldened the terrorists*, where *appeasement* and *terrorists* are opinions with negative polarity, with dependency syntax (above the text) and a predicate–argument structure (below). The predicate *emboldened*, an instance of the PropBank frame `embolden.01`, has two semantic arguments: the Agent (A0) and the Theme (A1), realized syntactically as a subject and a direct object, respectively.

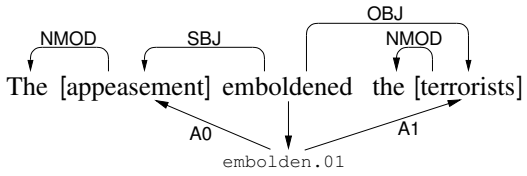

Figure 1: Syntactic and shallow semantic structure.

The model used the following novel features that take the polarities of the expressions into account. The examples are given with respect to the two expressions (*appeasement* and *terrorists*) in Figure 1.

Base polarity classifier score. Sum of the scores from the polarity classifier for every opinion.

Polarity pair. For every pair of opinions in the sentence, we add the pair of polarities: NEGATIVE+NEGATIVE.

Polarity pair and syntactic path. For a pair of opinions, we use the polarities and a representation of the path through the syntax tree between the expressions, following standard practice from dependency-based SRL (Johansson and Nugues, 2008): NEGATIVE+SBJ↑OBJ↓+NEGATIVE.

Polarity pair and syntactic dominance. In addition to the detailed syntactic path, we use a simpler feature based on dominance, i.e. that one expression is above the other in the syntax tree. In the example, no such feature is extracted since neither of the expressions dominates the other.

Polarity pair and word pair. The polarity pair concatenated with the words of the closest nodes of the two expressions: NEGATIVE+NEGATIVE+*appeasement*+*terrorists*.

Polarity pair and types and syntactic path. From the opinion sequence labeler, we get the expression type as in MPQA (DSE or ESE): ESE-NEGATIVE:+SBJ↑OBJ↓+ESE-NEGATIVE.

Polarity pair and semantic relation. When two opinions are directly connected through a link in the semantic structure, we add the role label as a feature.

Polarity pair and words along syntactic path. We follow the path between the expressions and add a feature for every word we pass: NEGATIVE:+emboldened+NEGATIVE.

We also used the features we developed in (Johansson and Moschitti, 2010b) to represent relations between expressions without taking polarity into account.

4.3 Training the Model

To train the model – find w – we applied max-margin estimation for structured outputs, a generalization of the well-known support vector machine from binary classification to prediction of structured objects. Formally, for a training set $\mathcal{T} = \{\langle x_i, y_i \rangle\}$, where the output space for the input x_i is $\mathcal{Y}_i$, we state the learning problem as a quadratic program:

$$\begin{array}{ll} \text{minimize}_{\,w} & \|w\|^2 \\ \text{subject to} & w(\Phi(x_i, y_i) - \Phi(x_i, y_{ij})) \geq \Delta(y_i, y_{ij}), \\ & \forall \langle x_i, y_i \rangle \in \mathcal{T}, y_{ij} \in \mathcal{Y}_i \end{array}$$

Since real-world data tends to be noisy, we may regularize to reduce overfitting and introduce a parameter C as in regular SVMs (Taskar et al., 2004). The quadratic program is usually not solved directly since the number of constraints precludes a direct solution. Instead, an approximation is needed in practice; we used SVM$^{\text{struct}}$ (Tsochantaridis et al., 2005; Joachims et al., 2009), which finds a solution by successively finding the most violated constraints and adding them to a working set. The loss Δ was defined as 1 minus a weighted combination of polarity-labeled and unlabeled intersection F-measure as described in Section 5.

5 Experiments

Opinion expression boundaries are hard to define rigorously (Wiebe et al., 2005), so evaluations of their quality typically use soft metrics. The MPQA annotators used the *overlap* metric: an expression is counted as correct if it overlaps with one in the gold standard. This has also been used to evaluate opinion extractors (Choi et al., 2006; Breck et al., 2007). However, this metric has a number of problems: 1) it is possible to "fool" the metric by creating expressions that cover the whole sentence; 2) it does not give higher credit to output that is "almost perfect" rather than "almost incorrect". Therefore, in (Johansson and Moschitti, 2010b), we measured the *intersection* between the system output and the gold standard: every compared segment is assigned a score between 0 and 1, as opposed to strict or overlap scoring that only assigns 0 or 1. For compatibility we present results in both metrics.

5.1 Evaluation of Segmentation with Polarity

We first compared the two baselines to the new integrated segmentation/polarity system. Table 1 shows the performance according to the intersection metric. Our first baseline consists of an expression segmenter and a polarity classifier (ES+PC), while in the second baseline we also add the expression reranker (ER) as we did in (Johansson and Moschitti, 2010b). The new reranker described in this paper is referred to as the expression/polarity reranker (EPR). We carried out the evaluation using the same partition of the MPQA dataset as in our previous work (Johansson and Moschitti, 2010b), with 541 documents in the training set and 150 in the test set.

System	P	R	F
ES+PC	56.5	38.4	45.7
ES+ER+PC	53.8	44.5	48.8
ES+PC+EPR	54.7	45.6	49.7

Table 1: Results with intersection metric.

The result shows that the reranking-based models give us significant boosts in recall, following our previous results in (Johansson and Moschitti, 2010b), which also mainly improved the recall. The precision shows a slight drop but much lower than the recall improvement.

In addition, we see the benefit of the new reranker with polarity interaction features. The system using this reranker (ES+PC+EPR) outperforms the expression reranker (ES+ER+PC). The performance differences are statistically significant according to a permutation test: precision $p < 0.02$, recall and F-measure $p < 0.005$.

5.2 Comparison with Previous Results

Since the results by Choi and Cardie (2010) are the only ones that we are aware of, we carried out an

evaluation in their setting.[1] Table 2 shows our figures (for the two baselines and the new reranker) along with theirs, referred to as C & C (2010). The table shows the scores for every polarity value. For compatibility with their evaluation, we used the overlap metric and carried out the evaluation using a 10-fold cross-validation procedure on a 400-document subset of the MPQA corpus.

POSITIVE	P	R	F
ES+PC	59.3	46.2	51.8
ES+ER+PC	53.1	50.9	52.0
ES+PC+EPR	58.2	49.3	53.4
C & C (2010)	67.1	31.8	43.1

NEUTRAL	P	R	F
ES+PC	61.0	49.3	54.3
ES+ER+PC	55.1	57.7	56.4
ES+PC+EPR	60.3	55.8	58.0
C & C (2010)	66.6	31.9	43.1

NEGATIVE	P	R	F
ES+PC	71.6	52.2	60.3
ES+ER+PC	65.4	58.2	61.6
ES+PC+EPR	67.6	59.9	63.5
C & C (2010)	76.2	40.4	52.8

Table 2: Results with overlap metric.

The C & C system shows a large precision bias despite being optimized with respect to the recall-promoting overlap metric. In recall and F-measure, their system scores much lower than our simplest baseline, which is in turn clearly outperformed by the stronger baseline and the polarity-based reranker. The precision is lower than for C & C overall, but this is offset by recall boosts for all polarities that are much larger than the precision drops. The polarity-based reranker (ES+PC+EPR) soundly outperforms all other systems.

6 Conclusion

We have studied the implementation of end-to-end systems for opinion expression extraction and polarity labeling. We first showed that it was easy to improve over previous results simply by combining an opinion extractor and a polarity classifier; the improvements were between 7.5 and 11 points in overlap F-measure.

However, our most interesting result is that a joint model of expression extraction and polarity labeling significantly improves over the sequential approach. This model uses features describing the interaction of opinions through linguistic structures. This precludes exact inference, but we resorted to a reranker. The model was trained using approximate max-margin learning. The final system improved over the baseline by 4 points in intersection F-measure and 7 points in recall. The improvements over Choi and Cardie (2010) ranged between 10 and 15 in overlap F-measure and between 17 and 24 in recall.

This is not only of practical value but also confirms our linguistic intuitions that surface phenomena such as syntax and semantic roles are used in encoding the rhetorical organization of the sentence, and that we can thus extract useful information from those structures. This would also suggest that we should leave the surface and instead process the *discourse structure*, and this has indeed been proposed (Somasundaran et al., 2009). However, automatic discourse structure analysis is still in its infancy while syntactic and shallow semantic parsing are relatively mature.

Interesting future work should be devoted to address the use of structural kernels for the proposed reranker. This would allow to better exploit syntactic and shallow semantic structures, e.g. as in (Moschitti, 2008), also applying lexical similarity and syntactic kernels (Bloehdorn et al., 2006; Bloehdorn and Moschitti, 2007a; Bloehdorn and Moschitti, 2007b; Moschitti, 2009).

Acknowledgements

The research described in this paper has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant 231126: LivingKnowledge – Facts, Opinions and Bias in Time, and under grant 247758: Trustworthy Eternal Systems via Evolving Software, Data and Knowledge (EternalS).

[1] In addition to polarity, their system also assigned opinion intensity which we do not consider here.

References

Stephan Bloehdorn and Alessandro Moschitti. 2007a. Combined syntactic and semantic kernels for text classification. In *Proceedings of ECIR 2007*, Rome, Italy.

Stephan Bloehdorn and Alessandro Moschitti. 2007b. Structure and semantics for expressive text kernels. In *In Proceedings of CIKM '07*.

Stephan Bloehdorn, Roberto Basili, Marco Cammisa, and Alessandro Moschitti. 2006. Semantic kernels for text classification based on topological measures of feature similarity. In *Proceedings of ICDM 06*, Hong Kong, 2006.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2683–2688, Hyderabad, India.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 793–801, Honolulu, United States.

Yejin Choi and Claire Cardie. 2010. Hierarchical sequential learning for extracting opinions and their attributes. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 269–274, Uppsala, Sweden.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8.

Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59.

Richard Johansson and Alessandro Moschitti. 2010a. Reranking models in fine-grained opinion analysis. In *Proceedings of the 23rd International Conference of Computational Linguistics (Coling 2010)*, pages 519–527, Beijing, China.

Richard Johansson and Alessandro Moschitti. 2010b. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 67–76, Uppsala, Sweden.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 183–187, Manchester, United Kingdom.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, United States.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceeding of CIKM '08*, NY, USA.

Alessandro Moschitti. 2009. Syntactic and Semantic Kernels for Short Text Pair Categorization. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 576–584, Athens, Greece, March. Association for Computational Linguistics.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of EMNLP 2009: conference on Empirical Methods in Natural Language Processing*.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin Markov networks. In *Advances in Neural Information Processing Systems 16*, Vancouver, Canada.

Iannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.