Bidirectional Inter-dependencies of Subjective Expressions and Targets and their Value for a Joint Model

Roman Klinger and **Philipp Cimiano**
Semantic Computing Group
Cognitive Interaction Technology – Center of Excellence (CIT-EC)
Bielefeld University
33615 Bielefeld, Germany
{rklinger,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Opinion mining is often regarded as a classification or segmentation task, involving the prediction of i) subjective expressions, ii) their target and iii) their polarity. Intuitively, these three variables are bidirectionally interdependent, but most work has either attempted to predict them in isolation or proposing pipeline-based approaches that cannot model the bidirectional interaction between these variables. Towards better understanding the interaction between these variables, we propose a model that allows for analyzing the relation of target and subjective phrases in both directions, thus providing an upper bound for the impact of a joint model in comparison to a pipeline model. We report results on two public datasets (cameras and cars), showing that our model outperforms state-of-the-art models, as well as on a new dataset consisting of Twitter posts.

1 Introduction

Sentiment analysis or opinion mining is the task of identifying subjective statements about products, their polarity (*e. g.* positive, negative or neutral) in addition to the particular aspect or feature of the entity that is under discussion, *i. e.*, the so-called *target*. Opinion analysis is thus typically approached as a classification (Täckström and McDonald, 2011; Sayeed et al., 2012; Pang and Lee, 2004) or segmentation (Choi et al., 2010; Johansson and Moschitti, 2011; Yang and Cardie, 2012) task by which fragments of the input are classified or labelled as representing a subjective phrase (Yang and Cardie, 2012), a polarity or a target (Hu and Liu, 2004; Li et al., 2010; Popescu and Etzioni, 2005; Jakob and Gurevych, 2010). As an example, the sentence "I like the low weight of the camera." contains a subjective term "like", and the target "low weight", which can be classified as a positive statement.

While the three key variables (subjective phrase, polarity and target) intuitively influence each other bidirectionally, most work in the area of opinion mining has concentrated on either predicting one of these variables in isolation (*e. g.* subjective expressions by Yang and Cardie (2012)) or modeling the dependencies uni-directionally in a pipeline architecture, *e. g.* predicting targets on the basis of perfect and complete knowledge about subjective terms (Jakob and Gurevych, 2010). However, such pipeline models do not allow for inclusion of bidirectional interactions between the key variables. In this paper, we propose a model that can include bidirectional dependencies, attempting to answer the following questions which so far have not been addressed but provide the basis for a joint model:

- What is the impact of the performance loss of a non-perfect subjective term extraction in comparison to perfect knowledge?
- Further, how does perfect knowledge about targets influence the prediction of subjective terms?
- How is the latter affected if the knowledge about targets is imperfect, *i. e.* predicted by a learned model?

We study these questions using imperatively defined factor graphs (IDFs, McCallum et al. (2008), McCallum et al. (2009)) to show how these bidirectional dependencies can be modeled in an architecture which allows for further steps towards joint inference. IDFs are a convenient way to define probabilistic graphical models that make structured predictions based on complex dependencies.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 848–854,
Sofia, Bulgaria, August 4-9 2013.

2 A Model for the Extraction of Target Phrases and Subjective Expressions

This section gives a brief introduction to imperatively defined factor graphs and then introduces our model.

2.1 Imperatively Defined Factor Graphs

A factor graph (Kschischang et al., 2001) is a bipartite graph over factors and variables. Let factor graph G define a probability distribution over a set of output variables $\mathbf{y}$ conditioned on input variables $\mathbf{x}$. A factor Ψ_i computes a scalar value over the subset of variables $\mathbf{x}_i$ and $\mathbf{y}_i$ that are neighbors of Ψ_i in the graph. Often this real-valued function is defined as the exponential of an inner product over sufficient statistics $\{f_{ik}(\mathbf{x}_i, \mathbf{y}_i)\}$ and parameters $\{\theta_{ik}\}$, where $k \in [1, K_i]$ and K_i is the number of parameters for factor Ψ_i.

A *factor template* T_j consists of parameters $\{\theta_{jk}\}$, sufficient statistic functions $\{f_{jk}\}$, and a description of an arbitrary relationship between variables, yielding a set of tuples $\{(\mathbf{x}_j, \mathbf{y}_j)\}$. For each of these tuples, the factor template instantiates a factor that shares $\{\theta_{jk}\}$ and $\{f_{jk}\}$ with all other instantiations of T_j. Let $\mathcal{T}$ be the set of factor templates and $Z(\mathbf{x})$ be the partition function for normalization. The probability distribution can then be written as $p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{T_j \in \mathcal{T}} \prod_{(\mathbf{x}_i, \mathbf{y}_i) \in T_j} \exp\left(\sum_{k=1}^{K_j} \theta_{jk} f_{jk}(\mathbf{x}_i, \mathbf{y}_i)\right)$.

FACTORIE[1] (McCallum et al., 2008; McCallum et al., 2009) is an implementation of imperatively defined factor graphs in the context of Markov chain Monte Carlo (MCMC) inference, a common approach for inference in very large graph structures (Culotta and McCallum, 2006; Richardson and Domingos, 2006; Milch et al., 2006). The term imperative is used to denote that actual code in an imperative programming language is written to describe templates and the relationship of tuples they yield. This flexibility is beneficial for modeling inter-dependencies as well as designing information flow in joint models.

[1] http://factorie.cs.umass.edu

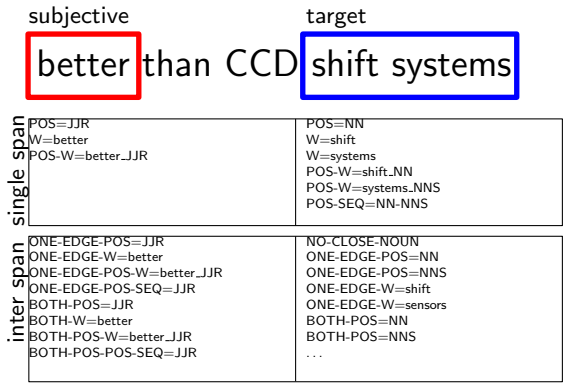


Figure 1: Example for features extracted for target and subjective expressions (text snippet taken from the camera data set (Kessler et al., 2010)). IOB-like features are merged for simplicity in this depiction.

2.2 Model

Our model is similar to a semi-Markov conditional random field (Sarawagi and Cohen, 2004). It predicts the offsets for target mentions and subjective phrases and can use the information of each other during inference. In contrast to a linear chain conditional random field (Lafferty et al., 2001), this allows for taking distant dependencies of unobserved variables into account and simplifies the design of features measuring characteristics of multi-token phrases. The relevant variables, *i. e.* target and subjective phrase, are modelled via complex span variables of the form $s = (l, r, c)$ with a left and right offset l and r, and a class $c \in \{\text{target}, \text{subjective}\}$. These offsets denote the span on a token sequence $\mathbf{t} = (t_1, \ldots, t_n)$.

We use two different templates to define factors between variables: a *single span* template and an *inter-span* template. The *single span* template defines factors with scores based on features of the tokens in the span and its vicinity. In our model, all features are boolean. As token-based features we use the POS tag, the lower-case representation of the token as well as both in combination. The actual span representation consists of these features prefixed with "I" for all tokens in the span, with "B" for the token at the beginning of the span, and with "E" for the token at the end of the span. In addition, the sequence of POS tags of all tokens in the span is included as a feature.

The inter-span template takes three characteristics of spans into account: Firstly, we measure if a potential target span contains a noun which is the closest noun to a subjective expression. Secondly, we measure for each span if a span of the other class is in the same sentence. A third feature indicates whether there is only one edge in the dependency graph between the tokens contained in spans of a different class. These features are to a great extent inspired by Jakob and Gurevych

(2010). For parsing, we use the Stanford parser (Klein and Manning, 2003).

The features described so far, however, cannot differentiate between a possible aspect mention which is a target of a subjective expression and one which is not. Therefore, the features of the inter-span template are actually built by taking the cross-product of the three described characteristics with all single-span features. Spans which are not in the context of a span of a different class are represented by a 'negated' feature (namely No-Close-Noun, No-Single-Edge, and Not-Both-In-Sentence). The example in Figure 1 shows features for two spans which are in context of each other. All of these features representing the text are taken into account for each class, *i. e.*, target and subjective expression.

Inference is performed via Markov Chain Monte Carlo (MCMC) sampling. In each sampling step, only the variables which actually change need to be evaluated, and therefore the sampler directs the process of unrolling the templates to factors. These world changes are necessary to find the maximum a posteriori (MAP) configuration as well as learning the parameters of the model. For each token in the sequence, a span of length one of each class is proposed if no span containing the token exists. For each existing span, it is proposed to change its label, shorten or extend it by one token if possible (all at the beginning and at the end of the span, respectively). Finally, a span can be removed completely.

In order to learn the parameters of our model, we apply SampleRank (Wick et al., 2011). A crucial component in the framework is the objective function which gives feedback about the quality of a sample proposal during training. We use the following objective function $f(\mathbf{t})$ to evaluate a proposed span $\mathbf{t}$:

$$f(\mathbf{t}) = \max_{\mathbf{g} \in \mathbf{s}} \frac{o(\mathbf{t}, \mathbf{g})}{|\mathbf{g}|} - \alpha \cdot p(\mathbf{t}, \mathbf{g}) \,,$$

where $\mathbf{s}$ is the set of all spans in the gold standard. Further, the function o calculates the overlap in terms of tokens of two spans and the function p returns the number of tokens in $\mathbf{t}$ that are not contained in $\mathbf{g}$, *i. e.*, those which are outside the overlap (both functions taking into account the class of the span). Thus, the first part of the objective function represents the fraction of correctly proposed contiguous tokens, while the second part penalizes a span for containing too many tokens that are outside the best span. Here, α is a parameter which controls the penalty.

3 Results and Discussion

3.1 Experimental Setting

We report results on the J.D. Power and Associates Sentiment Corpora[2], an annotated data set of blog posts in the car and in the camera domain (Kessler et al., 2010). From the rich annotation set, we use subjective terms and entity mentions which are in relation to them as targets. We do not consider `comitter`, `negator`, `neutralizer`, `comparison`, `opo`, or `descriptor` annotations to be subjective expressions. Results on these data sets are compared to Jakob and Gurevych (2010).

In addition, we report results on a Twitter data set[3] for the first time (Spina et al., 2012). Here, we use a Twitter-specific tokenizer and POS tagger[4] (Owoputi et al., 2013) instead of the Stanford parser. Hence, the single-edge-based feature described in Section 2.2 is not used for this dataset. A short summary of the datasets is given in Table 1.

As evaluation metric we use the F_1 measure, the harmonic mean between precision and recall. True positive spans are evaluated in a perfect match and approximate match mode, where the latter regards a span as positive if one token within it is included in a corresponding span in the gold standard. In this case, other predicted spans matching *the same* gold span do not count as false positives. In the objective function, α is set to 0.01 to prefer spans which are longer than the gold phrase over predicting no span.

Four different experiments are performed (all via 10-fold cross validation): First, predicting subjectivity expressions followed by predicting targets while making use of the previous prediction. Sec-

[2] `http://verbs.colorado.edu/jdpacorpus/`

[3] `http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html`

[4] In version 0.3, `http://www.ark.cs.cmu.edu/TweetNLP/`

	Car	Camera	Twitter
Texts	457	178	9238
Targets	11966	4516	1418
Subjectives	15056	5128	1519

Table 1: Statistics of the data sets.

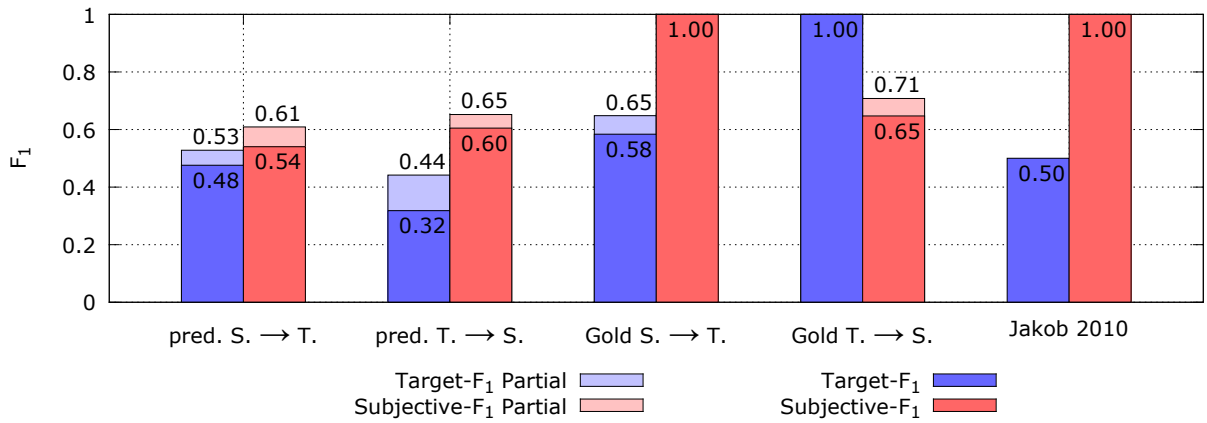


Figure 2: Results for the workflow of first predicting subjective phrases, then targets (pred. S. → T.), and vice versa (pred. T. → S.), as well as in comparison to having perfect information for the first step for the camera data set.

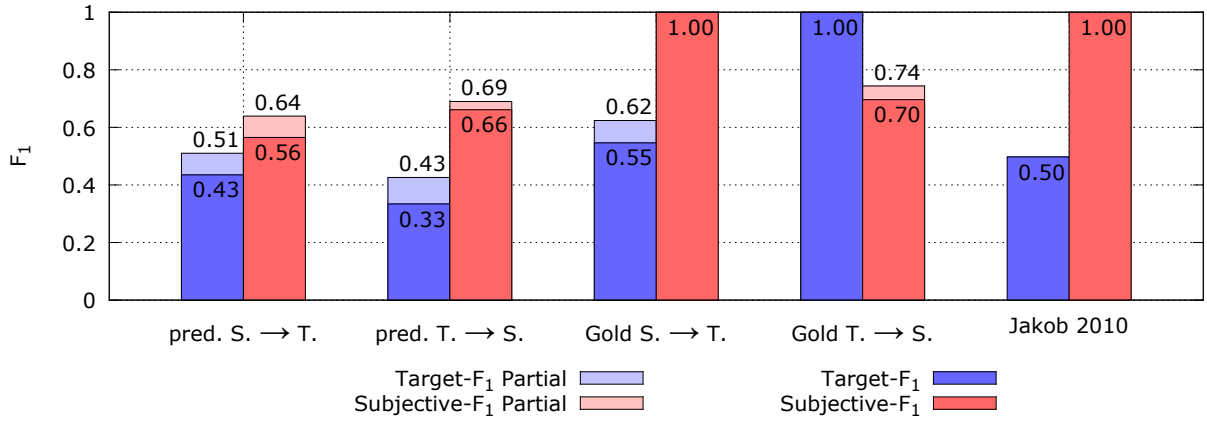


Figure 3: Results for the car data set.

ond, predicting targets followed by predicting subjective expressions. Third, assuming perfect knowledge of subjective expressions when predicting targets, and fourth, assuming perfect knowledge of targets in predicting subjective expressions. This provides us with the information how good a prediction can be with perfect knowledge of the other variable as well as an estimate of how good the prediction can be without any previous knowledge.

3.2 Results

Figures 2, 3 and 4 show the results for the four different settings compared to the results by Jakob and Gurevych (2010) for cars and cameras. The darker bars correspond to perfect match, the lighter ones to the increase when taking partial matches into account. In the following we only discuss the perfect match.

Comparing the results (for the car and camera data sets, Figure 2 and 3) for subjectivity prediction, one can observe a limited performance when targets are not known (0.54 F_1 for the camera set, 0.56 F_1 for the car set), an upper bound with perfect target information is much higher (0.65 F_1, 0.7 F_1). When first predicting targets followed by subjective term prediction, we obtain results of 0.6 F_1 and 0.66 F_1. The results for target prediction are much lower when not knowing subjective expressions in advance (0.32 F_1, 0.33 F_1), and clearly increase with predicted subjective expressions (0.48 F_1, 0.43 F_1) and outperform previous results when compared to Jakob and Gurevych (2010) (0.58 F_1, 0.55 F_1 in comparison to their 0.5 F_1 on both sets).

The results for the Twitter data set show the same characteristics (in Figure 4). However, they are generally much lower. In addition, the difference between exact and partial match evaluation modes

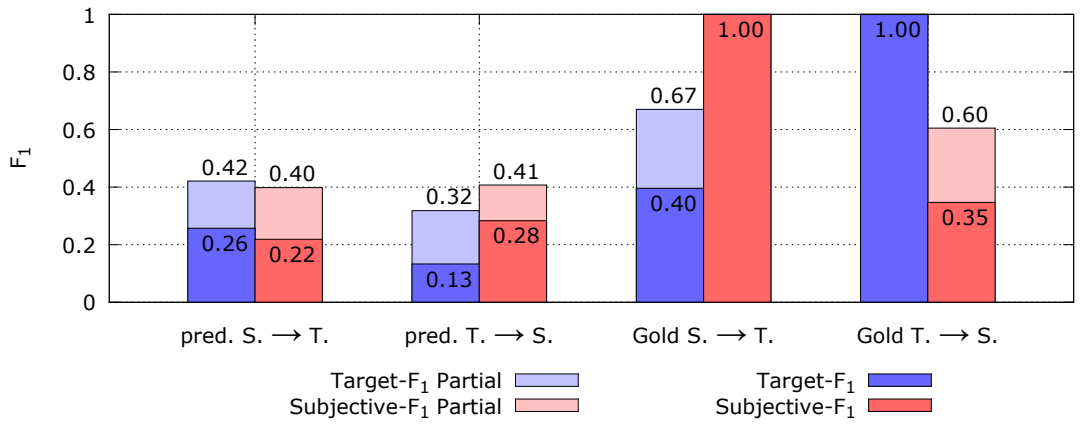
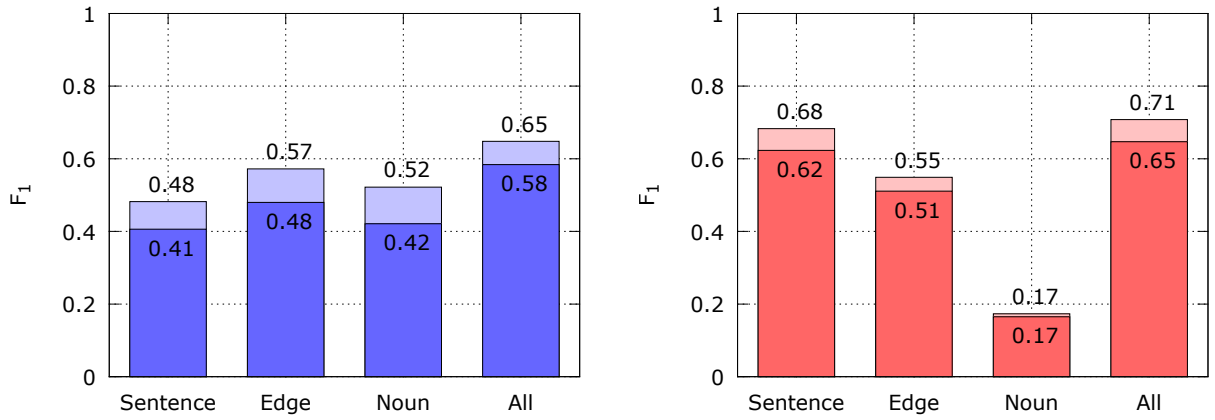


Figure 4: Results for the Twitter data set.



(a) Camera Data Set, given subjective terms.

(b) Camera Data Set, given target terms.

Figure 5: Evaluation of the impact of different features.

is higher. This is due to the existence of many more phrases spanning multiple tokens.

Exemplarily, the impact of the three features in the inter-span templates for the camera data set is depicted in Figure 5 for (a) given subjective terms (b) given targets, respectively. Detecting the closest noun is mainly of importance for target identification and only to a minor extent for detecting subjective phrases. A short path in the dependency graph and detecting if both phrases are in the same sentence have a high positive impact for both subjective and target phrases.

3.3 Conclusion and Discussion

The experiments in this paper show that target phrases and subjective terms are clearly interdependent. However, the impact of knowledge about one type of entity for the prediction of the other type of entity has been shown to be asymmetric. The results clearly suggest that the impact of subjective terms on target terms is higher than the other way round. Therefore, if a pipeline architecture is chosen, this order is to be preferred. However, the results with perfect knowledge of the counterpart entity show (in both directions) that the entities influence each other positively. Therefore, the challenge of extracting subjective expressions and their targets is a great candidate for applying supervised, joint inference.

Acknowledgments

Roman Klinger has been funded by the “It’s OWL” project (“Intelligent Technical Systems Ostwestfalen-Lippe”, `http://www.its-owl.de/`), a leading-edge cluster of the German Ministry of Education and Research. We thank the information extraction and synthesis laboratory (IESL) at the University of Massachusetts Amherst for their support.

References

Yoonjung Choi, Seongchan Kim, and Sung-Hyon Myaeng. 2010. Detecting Opinions and their Opinion Targets in NTCIR-8. *Proceedings of NTCIR8 Workshop Meeting*, pages 249–254.

A. Culotta and A. McCallum. 2006. Tractable Learning and Inference with High-Order Representations. In *ICML Workshop on Open Problems in Statistical Relational Learning*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers – Volume 2*, pages 101–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 ICWSM JDPA Sentment Corpus for the Automotive Domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

D. Klein and Ch. D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems*.

F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. 2001. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Trans on Information Theory*, 47(2):498–519.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1371–1376, Atlanta, Georgia, USA.

A. McCallum, K. Rohanimanesh, M. Wick, K. Schultz, and Sameer Singh. 2008. FACTORIE: Efficient Probabilistic Programming via Imperative Declarations of Structure, Inference and Learning. In *NIPS Workshop on Probabilistic Programming*.

Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

B. Milch, B. Marthi, and S. Russell. 2006. *BLOG: Relational Modeling with Unknown Objects*. Ph.D. thesis, University of California, Berkeley.

O. Owoputi, B. OConnor, Ch. Dyer, K. Gimpely, N. Schneider, and N. A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Main Volume*, pages 271–278, Barcelona, Spain, July.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

M. Richardson and P. Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136.

Sunita Sarawagi and William W. Cohen. 2004. Semimarkov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems*.

Asad Sayeed, Jordan Boyd-Graber, Bryan Rusk, and Amy Weinberg. 2012. Grammatical structures for word-level sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 667–676, Montréal, Canada, June. Association for Computational Linguistics.

D. Spina, E. Meij, A. Oghina, M. T. Bui, M. Breuss, and M. de Rijke. 2012. A Corpus for Entity Profiling in Microblog Posts. In *LREC Workshop on Information Access Technologies for Online Reputation Management*.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages

569–574, Portland, Oregon, USA, June. Association for Computational Linguistics.

M. Wick, K. Rohanimanesh, K. Bellare, A. Culotta, and A. McCallum. 2011. SampleRank: Training factor graphs with atomic gradients. In *International Conference on Machine Learning*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345, Stroudsburg, PA, USA. Association for Computational Linguistics.