Hierarchical Multi-label Conditional Random Fields for Aspect-Oriented Opinion Mining

Diego Marcheggiani[1], Oscar Täckström[2,⋆], Andrea Esuli[1], and Fabrizio Sebastiani[1]

[1] Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
firstname.lastname@isti.cnr.it
[2] Swedish Institute of Computer Science
164 29 Kista, Sweden
oscar@sics.se

Abstract. A common feature of many online review sites is the use of an overall rating that summarizes the opinions expressed in a review. Unfortunately, these document-level ratings do not provide any information about the opinions contained in the review that concern a specific aspect (e.g., cleanliness) of the product being reviewed (e.g., a hotel). In this paper we study the finer-grained problem of *aspect-oriented* opinion mining *at the sentence level*, which consists of predicting, for all sentences in the review, whether the sentence expresses a positive, neutral, or negative opinion (or no opinion at all) about a specific aspect of the product. For this task we propose a set of increasingly powerful models based on conditional random fields (CRFs), including a hierarchical multi-label CRFs scheme that jointly models the overall opinion expressed in the review and the set of aspect-specific opinions expressed in each of its sentences. We evaluate the proposed models against a dataset of hotel reviews (which we here make publicly available) in which the set of aspects and the opinions expressed concerning them are manually annotated at the sentence level. We find that both hierarchical and multi-label factors lead to improved predictions of aspect-oriented opinions.

1 Introduction

Sharing textual reviews of products and services is a popular social activity on the Web. Some websites (e.g., Amazon, TripAdvisor[1]) act as hubs that gather reviews on competing products, thus allowing consumers to compare them. While an overall rating (e.g., a number of "stars") is commonly attached to each such review, only a few of these websites (e.g., TripAdvisor) allow reviewers to include *aspect-specific* ratings, such as distinct ratings for the Value and Service provided by a hotel.

The overall and the aspect-specific ratings may help the user to perform a first screening of the product, but they are of little use if the user wants to actually read the *comments* about specific aspects of the product. For example, a low rating for the Rooms aspect of a hotel may be due to the small size of the room or to the quality of the furniture; different issues may be of different importance to different persons. In this case the user may have to read a lot of text in order to retrieve the relevant information.

⋆ Currently employed by Google Research. Contact: oscart@google.com

[1] http://www.amazon.com/, http://www.tripadvisor.com/

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 273–285, 2014.

Overall rating: ★★★½☆	*Aspect-specific opinions*	
Title: Good vlue [sic], terrible service	Value: Positive	Service: Negative
OK the value is good and the hotel is reasonably priced, but the service is terrible.	Value: Positive	Service: Negative
I was waiting 10 min at the erception [sic] desk for the guy to figure out whether there was a clean room available or not.	Checkin: Negative	Service: Negative
That place is a mess.	Service: Negative	
Rooms are clean and nice, but bear in mind you just pay for lodging, service does not seem to be included.	Cleanliness: Positive	Service: Negative

Fig. 1. An example hotel review annotated with aspect-specific opinions at the sentence level

Opinion mining research [9] has frequently considered the problem of predicting the overall rating of a review [14] or the ratings of its individual aspects [5]. While these are interesting research challenges, their practical utility is somewhat limited, since this information is often already made explicit by the reviewers in the form of an ordinal score. Our goal is instead to build an automatic system that, given a sentence in a review and one of the predefined aspects of interest, (a) predicts if an opinion concerning that aspect is expressed in the sentence, and (b) if so, predicts the polarity of the opinion (i.e., positive, neutral/mixed, or negative). This is a *multi-label* problem: a sentence may be relevant for (i.e., contain opinions concerning) zero, one, or several aspects at the same time, and the opinions contained in the same sentence and pertaining to different aspects may have different polarities. For example, *the room was spacious but the location was horrible* expresses a positive opinion for the Rooms aspect and a negative opinion for the Location aspect, while the remaining aspects are not touched upon.

The contribution of this study is twofold. First, inspired by the "coarse-to-fine" opinion model of [11] we develop an increasingly powerful set of multi-label conditional random field (CRF) schemes [6] that jointly model the overall, document-level opinion expressed by a review together with the aspect-specific opinions expressed at the sentence level. Our models are thus able to also predict the document-level ratings. However, as already pointed out, these ratings are of smaller practical interest, because they are often explicitly provided by the reviewers, whereas the aspect-level predictions are often not available and the sentence-level annotations (i.e., the indication of which sentences justify the aspect-level ratings) are never available. The use of a conditional model for this task is in contrast with previous work in this area, which has focused on generative models, mostly based on Latent Dirichlet Allocation, with strong independence assumptions [7,12,17,18]. This problem has also been tackled via supervised learning methods in [8]; like ours, this work relies on CRFs to model the structure of the reviews, but is unable to cater for sentences that are relevant to more than one aspect at the same time, which is a strong limitation. Two works that are close in spirit to ours are [7,18], and they may be considered the "generative counterparts" of our approach.

Second, we present (and make publicly available) a new dataset of hotel reviews that we have annotated with aspect-specific opinions at the sentence level. A previous dataset annotated by opinion at the sentence level exists [16], but the dataset introduced here also adds the aspect dimension and has a multi-label nature. Only very recently, and after we created our dataset, a dataset similar to ours has been presented [7], in which elementary discourse units (EDUs), which can be sub-sentence entities, are annotated using a *single-label* model. The dataset of [7] is composed of 65 reviews, with a total of 1541 EDUs. Our dataset annotates 442 reviews, with a total of 5799 sentences.

The evaluation of generative models is often based on unannotated datasets [12,18], and thus only on a qualitative analysis of the generated output. We believe that our dataset will be a valuable resource to fuel further research in the area by enabling a quantitative evaluation, and thus a rigorous comparison of different models.

1.1 Problem Definition

Before describing our approach, let us define the task just introduced more formally. Let $\mathbb{A}$ be a discrete set of aspect labels and let $\mathbb{Y}$ be a discrete set of opinion labels. Given a review $\mathbf{x} \in \mathbf{X}$ composed of T consecutive segments, we seek to infer the values of the following variables: first, the overall opinion $y_o \in \mathbb{Y}$ expressed in $\mathbf{x}$; second, the opinion $y_t^a \in \mathbb{Y} \cup \{\textsf{No-op}\}$ expressed concerning aspect a in segment t, for each segment $t \in \{1, ..., T\}$ and each aspect $a \in \mathbb{A}$ (where No-op stands for "no opinion"). This is a *multi-label* problem, since each segment t can be assigned up to $|\mathbb{A}|$ different opinions.

To model these variables we assume a feature vector $\mathbf{x}_t$ representing review segment t and a feature vector $\mathbf{x}_o$ representing the full review. For our experiments, reported in Section 3, we use a dataset of hotel reviews; we take segments to correspond to sentences, and we take $\mathbb{Y} = \{\textsf{Positive}, \textsf{Negative}, \textsf{Neutral}\}$ and $\mathbb{A} = \{\textsf{Rooms}, \textsf{Cleanliness}, \textsf{Value}, \textsf{Service}, \textsf{Location}, \textsf{Check-in}, \textsf{Business}, \textsf{Food}, \textsf{Building}, \textsf{Other}\}$. However, we want to stress that the proposed models are flexible enough to incorporate arbitrary sets of aspects and opinion labels, and to use a different type of segmentation.

2 Models, Inference and Learning

Previous work on aspect-oriented opinion mining has focused on generative probabilistic models [7,17,18]. Thanks to their generative nature, these models can be learnt without any explicit supervision. However, at the same time they make strong independence assumptions on the variables to be inferred, which is known to limit their performance in the supervised scenario considered in this study. Instead, we turn to CRFs — a general and flexible class of structured conditional probabilistic models. Specifically, we propose a hierarchical multi-label CRF model that jointly models the overall opinion of a review together with aspect-specific opinions at the segment level. This model is inspired by the *fine-to-coarse* opinion model [11], which was recently extended to a partially supervised setting [16].[2] However, while previous work only takes opinion into

[2] While we only consider the supervised scenario in this study, our model is readily extensible to the partially supervised setting by treating a subset of the fine-grained variables as latent.

account, we jointly model both sentence-level opinion and aspect, as well as overall review opinion. Below, we introduce a sequence of increasingly powerful CRF models (that we implemented using Factorie [10].) for aspect-specific opinion mining, leading up to the full hierarchical sequential multi-label model.

2.1 CRF Models of Aspect-Oriented Opinion

A CRF models the conditional distribution $p(\mathbf{y}|\mathbf{x})$ over a collection of output variables $\mathbf{y} \in \mathbf{Y}$, given an input $\mathbf{x} \in \mathbf{X}$, as a globally normalized log-linear distribution [6]:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{\Psi_c \in \mathbf{F}} \Psi_c(\mathbf{y}_c, \mathbf{x}_c) \propto \prod_{\Psi_c \in \mathbf{F}} \Psi_c(\mathbf{y}_c, \mathbf{x}_c)\,, \tag{1}$$

where $\mathbf{F}$ is the set of factors and $Z(\mathbf{x}) = \sum_{\mathbf{y} \in \mathbf{Y}} p(\mathbf{y}|\mathbf{x})$ is a normalization constant. In this study, $\mathbf{y} = \{y_o\} \cup \{y_t^a : t \in [1, T], a \in \mathbb{A}\}$. Each factor $\Psi_c(\mathbf{y}_c, \mathbf{x}_c) = \exp(\mathbf{w} \cdot \mathbf{f}(\mathbf{y}_c, \mathbf{x}_c))$ scores a set of variables $\mathbf{y}_c \subset \mathbf{y}$ by means of the parameter vector $\mathbf{w}$ and the feature vector $\mathbf{f}(\mathbf{y}_c, \mathbf{x}_c)$. The models described in what follows differ in terms of their factorization of Equation (1) and in the features employed.

Linear-Chain Baseline Model. As a baseline model, we take a simple first-order linear-chain CRF (LC) in which a separate linear chain over opinions at the segment level is defined for each aspect. This model is able to take into account sequential dependencies between segment opinions [11,13] specific to the same aspect, whereas opinions related to different aspects are assumed to be independent. Formally, the LC model factors as

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{a \in \mathbb{A}} \prod_{t=1}^{T} \Psi_s(y_t^a, \mathbf{x}_t) \prod_{t=1}^{T-1} \Psi_{\curvearrowright}(y_t^a, y_{t+1}^a)\,, \tag{2}$$

where $\Psi_s(y_t^a, \mathbf{x}_t)$ models the aspect-specific opinion of the segment at position t and $\Psi_{\curvearrowright}(y_t^a, y_{t+1}^a)$ models the transition between the aspect-specific opinion variables at position t and $t+1$ in the linear chain corresponding to aspect a.

Multi-label Models. The assumption of the LC model that the aspect-specific opinions expressed in each segment are independent of each other may be overly strong for two reasons. First, only a limited number of aspects are generally addressed in each segment. Second, when several aspects are mentioned, it is likely that there are dependencies between them based on discourse structure considerations. To address these shortcomings, we propose to model the dependencies between aspect-specific opinion variables within each segment, by adopting the multi-label pairwise CRF formulation of [4].

We first consider the *Independent Multi-Label* (IML) model, in which there are factors between the opinion variables within a segment, while each segment is independent from each other. In terms of Equation (1), the IML model factors as

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{T} \prod_{a \in \mathbb{A}} \Psi_s(y_t^a, \mathbf{x}_t) \prod_{b \in \mathbb{A} \setminus \{a\}} \Psi_m(y_t^a, y_t^b)\,, \tag{3}$$

where $\Psi_m(y_t^a, y_t^b)$ is the pairwise multi-label factor, which models the interdependence of the opinion variables corresponding to aspects a and b at position t. Note that this factor ignores the input, considering only the interaction of the opinion variables.

To allow for sequential dependencies between segments, the IML model can naturally be combined with the LC model. This yields the *Chain Multi-Label* (CML) model:

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{T} \prod_{a \in \mathbb{A}} \Psi_s(y_t^a, \mathbf{x}_t) \prod_{b \in \mathbb{A} \setminus \{a\}} \Psi_m(y_t^a, y_t^b) \prod_{t=1}^{T-1} \Psi_{\frown}(y_t^a, y_{t+1}^a)\,. \tag{4}$$

Hierarchical (Multi-label) Models. Thus far, we have only modeled the aspect-specific opinions expressed at the segment level. However, many online review sites ask users to provide an overall opinion in the form of a numerical rating as part of their review. As shown by [11,16], jointly modeling the overall opinion and the segment-level opinions in a hierarchical fashion can be beneficial to prediction at both levels.

The LC, IML and CML models can be adapted to include the overall rating variable in a hierarchical model structure analogous to that of the "coarse-to-fine" opinion model of [11]. This is accomplished by adding the following two factors to the three models above: the overall opinion factor $\Psi_o(y_o, \mathbf{x}_o)$, which models the overall opinion with respect to the input; and the pairwise factor $\Psi_h(y_t^a, y_o)$, which connects the two levels of the hierarchy by modeling the interaction of the aspect-specific opinion variable at position t and the overall opinion variable.

By combining the shared product of factors $\Phi(y_o, y_t^a, \mathbf{x}) = \Psi_s(y_t^a, \mathbf{x}_t) \cdot \Psi_o(y_o, \mathbf{x}_o) \cdot \Psi_h(y_t^a, y_o)$ with the LC, IML and CML models, we get the *Linear-Chain Overall* (LCO) model:

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{T} \prod_{a \in \mathbb{A}} \Phi(y_o, y_t^a, \mathbf{x}) \prod_{t=1}^{T-1} \Psi_{\frown}(y_t^a, y_{t+1}^a)\,, \tag{5}$$

the *Independent Multi-Label Overall* (IMLO) model:

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{T} \prod_{a \in \mathbb{A}} \Phi(y_o, y_t^a, \mathbf{x}) \prod_{b \in \mathbb{A} \setminus \{a\}} \Psi_m(y_t^a, y_t^b)\,, \tag{6}$$

and the *Chain Multi-Label Overall* (CMLO) model:

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{T} \prod_{a \in \mathbb{A}} \Phi(y_o, y_t^a, \mathbf{x}) \prod_{b \in \mathbb{A} \setminus \{a\}} \Psi_m(y_t^a, y_t^b) \prod_{t=1}^{T-1} \Psi_{\frown}(y_t^a, y_{t+1}^a)\,. \tag{7}$$

2.2 Model Features

The joint problem of aspect-oriented opinion prediction requires model features that help to discriminate opinions and aspects, as well as opinions specific to a particular aspect. In the experiments of Section 3 we use both word and word bigram identity features, as well as a set of polarity lexicon features based on the General Inquirer (GI) [15], MPQA [20], and SentiWordNet (SWN) [2] lexicons. The numerical polarity values of these lexicons are mapped into the set {Positive, Negative, Neutral}. The mapped lexicon values are used to generalize word bigram features by substituting the matching words of the bigram with the correspondent polarity. For example, the bigram *nice hotel* is generalized to the bigram *SWN:positive hotel*, from looking up *nice* in SentiWordNet. These features are used both with segment-level and review-level factors; see Table 1.

Table 1. The collection of model factors and their corresponding features, see Section 2.1 for details on notation. Feature vectors: $\mathbf{x}_t$: {words, bigrams, SWN/MPQA/GI bigrams, χ^2 lexicon matches} in the t:th segment in review $\mathbf{x}$; $\mathbf{x}_o$: {words, bigrams, SWN/MPQA/GI bigrams} in $\mathbf{x}$.

Factor	Description	Features
$\Psi_s(y_t^a, \mathbf{x}_t)$	Segment aspect-opinion	$\mathbf{x}_t \otimes y_t^a \otimes a$
$\Psi_o(y_o, \mathbf{x}_o)$	Overall opinion	$\mathbf{x}_o \otimes y_o$
$\Psi_\frown(y_t^a, y_{t+1}^a)$	Segment aspect-opinion transition	$y_t^a \otimes y_{t+1}^a \otimes a$
$\Psi_m(y_t^a, y_t^b)$	Multi-label segment aspect-opinion	$y_t^a \otimes y_t^b \otimes a \otimes b$
$\Psi_h(y_t^a, y_o)$	Hierarchical overall / segment aspect-opinion	$y_t^a \otimes y_o \otimes a$

In addition to these features we use an aspect-specific lexicon obtained via the algorithm proposed in [18]; this is an algorithm that iteratively builds a set of aspect-specific words by adding to it words that co-occur with the words already present in it, and where co-occurrence is detected via the χ^2 measure. We use the output of this algorithm to create what we call the χ^2 *lexicon*, in which each word is associated with the (normalized) frequency with which the word is used to describe a certain aspect.

2.3 Inference and Learning

While the maximum a posteriori (MAP) assignment $\mathbf{y}^* \in \mathbf{Y}$ and factor marginals can be inferred exactly in the LC and LCO models by means of variants of the Viterbi and forward-backward algorithms [11], exact inference is not tractable in the remaining models due to a combinatorial explosion and to the presence of loops in the graph structure. Instead, we revert to approximate inference via Gibbs sampling (see, e.g., [3]).

All models are trained to approximately minimize the Hamming loss over the training set $\mathbf{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$ using the SampleRank algorithm [19], which is a natural fit to sampling-based inference.[3] Briefly put, with SampleRank the model parameters $\mathbf{w}$

[3] While inference and learning algorithms are likely to impact results, this decision brings about no substantial loss of generality, since the focus of this study is on comparing model structures.

are updated locally after each draw from the Gibbs sampler by taking an atomic gradient step with respect to the local Hamming loss incurred by the sampled variable setting. This procedure is repeated for a number of epochs until the ℓ_2-norm of the sum of the atomic gradients from the epoch is below a threshold ϵ; in each epoch every variable in the training set is sampled in turn. For the experiments in Section 3, the SampleRank learning rate was fixed to $\alpha = 1$ and the gradient threshold to $\epsilon = 10^{-5}$.

After fitting the model parameters to the training data, at test time we perform 100 Gibbs sampling epochs to find an approximate MAP assignment $\mathbf{y}^*$ for input $\mathbf{x}$ with respect to the distribution $p(\mathbf{y}|\mathbf{x})$.

3 Experiments

In this section we study the proposed models empirically. After discussing our evaluation strategy, we describe and discuss the creation of a new dataset of hotel reviews, which has been manually annotated with aspect-specific opinion at the sentence level. Finally, we compare the proposed models quantitatively by their performance on this dataset.

Evaluation Measures. When evaluating system output and comparing human annotations below, we view the task as composed of the following two subtasks:

Aspect identification: for each segment and for each aspect, predict if there is any opinion expressed towards the aspect in the segment. Since each of these aspect-specific tasks is a binary problem, for this subtask we adopt the standard F_1 evaluation measure.

Opinion prediction: for each segment and each applicable (true positive) aspect for the segment, predict the opinion expressed towards the aspect in the segment. Since opinions are placed on an ordinal scale, as an evaluation measure we adopt *macro-averaged mean absolute error* (MAE^M) [1], a measure for evaluating ordinal classification that is also robust to the presence of imbalance in the dataset.

Let $\mathbf{T}$ be the correct label assignments and let $\widehat{\mathbf{T}}$ be the corresponding model predictions. Let $\mathbf{T}_j = \{y_i : y_i \in \mathbf{T}, y_i = j\}$ and let n be the number of unique labels in $\mathbf{T}$. The macro-averaged mean absolute error is defined as

$$\mathrm{MAE}^M(\mathbf{T}, \widehat{\mathbf{T}}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{|\mathbf{T}_j|} \sum_{y_i \in \mathbf{T}_j} |y_i - \hat{y}_i| \tag{8}$$

This is suitable for evaluating the overall review-level opinion predictions. However, when evaluating the aspect-specific opinions at the segment level, we instead report $\mathrm{MAE}^M(\mathbf{T}_{\widehat{\mathbf{I}}_a}, \widehat{\mathbf{T}}_{\widehat{\mathbf{I}}_a})$, where $\widehat{\mathbf{I}}_a$ is the sequence of indices of segment opinion labels that were predicted as true positive for aspect a and $\mathbf{T}_{\widehat{\mathbf{I}}_a}$ is the set of true positive opinion labels for aspect a.

Inter-annotator Agreement Measures. We also use F_1 and MAE^M to assess inter-annotator agreement, by computing the average of these measures over all pairs of annotators. While F_1 and the micro-averaged version of MAE are both symmetric, the use of macro-averaging makes MAE^M asymmetric, i.e., switching the predicted labels with the gold standard labels may change the outcome. This is problematic when

Table 2. Number of opinion expressions at the sentence level, broken down by aspect and opinion. Out of 5799 annotated sentences, 4810 sentences contain at least one opinion-laden expression.

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	NotRelated	*Total*
Pos	893	513	484	180	287	435	93	188	185	23	63	*3344*
Neg	353	248	287	66	127	51	56	87	62	3	40	*1377*
Neu	167	40	111	5	82	38	12	35	22	4	350	*866*
Total	1413	801	882	251	496	524	161	310	269	30	453	*5134*

used to measure inter-annotator agreement, since no annotator can be given precedence over the others (unless they have different levels of expertise). We thus symmetrize the measure by treating each annotator in turn as the gold standard and by averaging the corresponding results. This yields the *symmetrized macro-averaged mean absolute error*:

$$\mathrm{sMAE}^M(\mathbf{T}, \widehat{\mathbf{T}}) = \frac{1}{2}\left(\mathrm{MAE}^M(\mathbf{T}, \widehat{\mathbf{T}}) + \mathrm{MAE}^M(\widehat{\mathbf{T}}, \mathbf{T})\right) \tag{9}$$

3.1 Annotated Dataset

We have produced a new dataset of manually annotated hotel reviews[4]. Three equally experienced annotators provided sentence-level annotations of a subset of 442 randomly selected reviews from the publicly available TripAdvisor dataset [18]. Each review comes with an overall rating on a discrete ordinal scale from 1 to 5 "stars".

The annotations are related to 9 aspects often present in hotel reviews. In addition to the 7 aspects explicitly present (at the review level) in the TripAdvisor dataset (Rooms, Cleanliness, Value, Service, Location, Check-in, and Business), we decided to add 2 other aspects (Food and Building), since many comments in the reviews refer to them. Furthermore, the "catch-all" aspects Other and NotRelated were added, for a total of 11 aspects. Other captures those opinion-related aspects that cannot be assigned to any of the first 9 aspects, but which are still about the hotel under review. The NotRelated aspect captures those opinion-related aspects that are not relevant to the hotel under review. In what follows, segments marked as NotRelated are treated as non-opinionated.

The annotation distinguishes between Positive, Negative and Neutral/Mixed opinions. The Neutral/Mixed label is assigned to opinions that are about an aspect without expressing a polarized opinion, and to opinions of contrasting polarities, such as *The room was average size* (neutral) and *Pricey but worth it!* (mixed). The annotations also distinguish between explicit and implicit opinion expressions, i.e., between expressions that refer directly to an aspect and expressions that refer indirectly to an aspect by referring to some other property/entity related to the aspect. For example, *Fine rooms* is an explicitly expressed positive opinion concerning the Rooms aspect, while *We had great views over the East River* is an implicitly expressed positive opinion concerning the Location aspect.

[4] At `http://nemis.isti.cnr.it/~marcheggiani/datasets/` the interested reader may find both the dataset and a more detailed explanation of it.

Table 3. Inter-annotator agreement results. Top 3 rows: segment-level aspect agreement, expressed in terms of F_1 (higher is better). Bottom 3 rows: segment-level opinion agreement (restricted to the true positive aspects for each segment), expressed in terms of sMAE^M (lower is better).

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	*Avg*
Overall	.607	.719	.793	.733	.794	.795	.464	.575	.553	.631	*.675*
Implicit	.167	.123	.263	.111	.306	.286	.061	.131	.095	.333	*.188*
Explicit	.479	.684	.706	.739	.741	.710	.481	.560	.521	.624	*.625*
Overall	.308	.219	.191	.114	.234	.259	.003	.202	.150	.029	*.171*
Implicit	.167	.000	.000	.000	.074	.061	.000	.000	.000	.000	*.030*
Explicit	.262	.167	.147	.064	.190	.119	.000	.179	.092	.000	*.122*

Out of the 442 reviews, 73 reviews were independently annotated by all three annotators so as to facilitate the measurement of inter-annotator agreement, while the remaining 369 reviews were subdivided equally among the annotators. These 369 reviews were then partitioned into a training set (258 reviews, 70% of the total) and a test set (111 reviews, 30% of the total). The data were split by selecting reviews for each subset in an interleaved fashion, so that each subset constitutes a minimally biased sample both with respect to the full dataset and with respect to annotator experience.

Table 2 shows, for each aspect and for each opinion type, the number of segments annotated with a given aspect and a given opinion type (across the unique reviews and averaged across the shared reviews). Both opinions and aspects show a markedly imbalanced distribution. As expected, the imbalance with respect to opinion is towards the Positive label. In terms of aspects, the Rooms, Service and Other aspects dominate.

Inter-annotator Agreement. We use the 73 shared reviews (943 sentences) to measure the agreement between the 3 annotators with respect to both aspects and opinions, using F_1 and symmetrized MAE^M. For each aspect we separately measure the agreement on implicit and explicit opinionated mentions, and the agreement on mentions of both types.

From the agreement results in Table 3 (top) we see a large disagreement with respect to implicit opinions. However, the agreement overall (disregarding the explicit/implicit distinction) is higher than the agreement on explicit opinions in isolation. This suggest that, while it is difficult for annotators to separate implicit from explicit opinions, separating opinionated mentions from non-opinionated mentions is easier. In what follows we thus ignore the distinction between implicit and explicit opinions.

Table 3 (bottom 3 rows) shows the agreement on the true positive opinion annotations, that is, the agreement on the opinions with respect to those aspects on which the two annotators agree. Closer inspection of the data shows that, as could be expected, the disagreement mainly affects the pairs Neutral–Positive and Neutral–Negative.

3.2 Results

All models were trained on the training set described in Section 3.1, for a total of 258 reviews. Below we describe two separate evaluations. First, we compare the different

Table 4. Aspect-oriented opinion prediction results for different CRF models averaged across five experiments with 5 different random seeds. Top 6 rows: segment-level aspect prediction results in terms of F_1 (higher is better). Bottom 6 rows: segment-level opinion prediction results (restricted to the true positive aspects for each segment) in terms of MAE^M (lower is better).

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	*Avg*
LC	.499	.606	.662	.700	.579	.623	.329	**.395**	.298	.000	*.469*
IML	**.542**	.597	**.664**	**.732**	**.605**	.668	**.371**	.373	**.363**	.000	*.491*
CML	.489	**.645**	.655	.708	.605	**.673**	.327	.408	.358	**.076**	**.494**
LCO	.515	.586	.661	.697	.582	.611	.301	.384	**.368**	**.173**	**.488**
IMLO	.513	.621	**.685**	.702	.593	.614	**.370**	.363	.348	.040	*.485*
CMLO	**.531**	**.629**	.663	**.706**	**.602**	**.618**	.271	**.393**	.350	.081	*.485*
LC	.526	.721	.572	1.000	.566	.932	**.644**	**.616**	.693	.000	*.627*
IML	.520	**.659**	**.494**	**.956**	**.377**	.939	.670	.700	.668	.000	*.598*
CML	**.492**	.681	.613	.978	.482	**.906**	.735	.691	**.377**	.000	**.595**
LCO	.482	.626	**.398**	1.000	.633	**.903**	.690	.490	.233	.000	*.546*
IMLO	**.473**	**.615**	.398	1.000	**.457**	.970	**.343**	**.469**	.269	.000	**500**
CMLO	.499	.626	.428	1.000	.711	.906	.536	.552	**.232**	.000	*.549*

models by their accuracy on the test set (111 reviews). Since training is non-deterministic due to the use of sampling-based inference, we report the average over five trials with different random seeds. Second, we compare the best-performing model to the human annotators on the set of 73 reviews independently annotated by all three annotators.

Comparison among Systems. As shown in Table 4, the multi-label and hierarchical models outperform the LC baseline in both aspect identification and opinion prediction. In particular, the multi-label models (IML, CML) significantly outperform the baseline on both subtasks, which shows the importance of modeling the interdependence of different aspects and their opinions within a segment. On the other hand, combining both multi-label and transition factors in the hierarchical model (CMLO) leads to worse predictions compared to only including the multi-label factors (IMLO) or the transition factors. We hypothesize that this is due to inference errors, where the more complex graph structure causes the Gibbs sampler to converge more slowly. Furthermore, while the hierarchical models provide a significant improvement compared to their non-hierarchical counterparts in terms of opinion prediction, modeling both the overall and segment-level opinions is not helpful for aspect identification. This is not too surprising, given that the overall opinion contains no information about aspect-specific opinions.[5]

[5] In addition to the reported experiments, we performed initial experiments with models that also included variables for overall opinions with respect to specific aspects. However, including these variables hurts performance at the segment level. We hypothesize that this is because reviewers often rate multiple aspects while only discussing a subset of them in the review text.

Table 5. Comparison between the best-performing model (IMLO) and the human annotators with IMLO results averaged over five runs (F_1 for the top two rows, sMAE^M for the bottom two rows)

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	*Avg*
Human	.607	.719	.793	.795	.553	.575	.794	.464	.733	.631	*.675*
IMLO	.479	.585	.606	.614	.536	.673	.407	.429	.208	.190	*.473*
Human	.308	.219	.191	.259	.150	.202	.234	.003	.114	.029	*.171*
IMLO	.676	.498	.445	.142	.451	.704	.212	.387	.025	.415	*.396*

The overall review-level opinion prediction results (not shown in Table 4) are in line with the segment-level results. The IMLO model (.504) outperforms the LCO baseline (.518), as measured with MAE^M. However, as with the segment-level predictions, including both multi-label and transition factors in the hierarchical model (CMLO) hurts overall opinion prediction (.544).

Comparison among Humans and System. We now turn to a comparison between the best-performing model (IMLO) and the human annotators, treating the model as a fourth annotator when computing inter-annotator agreement. This allows us to assess how far our model is from human-level performance. Table 5 clearly shows that much work remains to be done for both subtasks. The aspects Building and Business are difficult to detect for the automatic system, while a human identifies them with ease. We believe that the reasons for the poor performance may be different for the two aspects. For the Business aspect, the reason is likely the scarcity of training annotations, whereas for the Building aspect the reason may be lexical promiscuity (that is, a hotel building may be described by a multitude of features, such as interior, furniture, architecture, etc.).

Interestingly, the system identifies the Value aspect at close to human level, but performs dramatically worse on its opinion prediction. We suggest that this is because assessing the value of something coined in absolute terms (for example, that a $30 room is cheap) requires world knowledge (or feature engineering).

4 Conclusions

We have considered the problem of aspect-oriented opinion mining at the sentence level. Specifically, we have devised a sequence of increasingly powerful CRF models, culminating in a hierarchical multi-label model that jointly models both the overall opinion expressed in a review and the set of aspect-specific opinions expressed in each sentence of the review. Moreover, we have produced a manually annotated dataset of hotel reviews in which the set of relevant aspects and the opinions expressed concerning these aspects are annotated for each sentence; we make this dataset publicly available with the hope to spur further research in this area. We have evaluated the proposed models on this dataset; the empirical results show that the hierarchical multi-label model outperforms a strong comparable baseline.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA 2009), Pisa, IT, pp. 283–287 (2009)
2. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), Valletta, MT (2010)
3. Bishop, C.M.: Pattern recognition and machine learning. Springer, Heidelberg (2006)
4. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005), Bremen, DE, pp. 195–200 (2005)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), Seattle, US, pp. 168–177 (2004)
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning (ICML 2001), Williamstown, US, pp. 282–289 (2001)
7. Lazaridou, A., Titov, I., Sporleder, C.: A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, BL, pp. 1630–1639 (2013)
8. Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.J., Zhang, S., Yu, H.: Structure-aware review mining and summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Bejing, CN, pp. 653–661 (2010)
9. Liu, B.: Sentiment analysis and opinion mining. Morgan & Claypool Publishers, San Rafael (2012)
10. McCallum, A., Schultz, K., Singh, S.: Factorie: Probabilistic programming via imperatively defined factor graphs. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), Vancouver, CA, pp. 1249–1257 (2009)
11. McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J.: Structured models for fine-to-coarse sentiment analysis. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, CZ, pp. 432–439 (2007)
12. Moghaddam, S., Ester, M.: ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In: Proceedings of the 34th ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR 2011), Bejing, CN, pp. 665–674 (2011)
13. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, ES, pp. 271–278 (2004)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, US, pp. 79–86 (2002)
15. Stone, P.J., Dunphy, D.C., Smith, M.S.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
16. Täckström, O., McDonald, R.: Discovering fine-grained sentiment with latent variable structured prediction models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 368–374. Springer, Heidelberg (2011)

17. Titov, I., McDonald, R.T.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008), Columbus, US, pp. 308–316 (2008)
18. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: A rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), Washington, US, pp. 783–792 (2010)
19. Wick, M., Rohanimanesh, K., Bellare, K., Culotta, A., McCallum, A.: SampleRank: Training factor graphs with atomic gradients. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), Bellevue, US, pp. 777–784 (2011)
20. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA, pp. 347–354 (2005)