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

Diego Marcheggiani[+], Oscar Täckström[*],
Andrea Esuli[+], Fabrizio Sebastiani[+]

+ Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
firstname.lastname@isti.cnr.it

* Swedish Institute of Computer Science
SE-164 29 Kista, Sweden
oscar@sics.se

ECIR 2014
Amsterdam, NL, 15 April 2014

Outline

Outline

Motivations

- Overall rating is commonly attached to product reviews.
- Some websites (e.g., TripAdvisor) allow reviewers to include *aspect-specific* ratings.
- Both of them are of little use if the reader is interested in the *comments* about specific aspects of the product.

Motivations (cont'd)

- E.g., we want to see at first glance why the reviewer gave a negative rating for aspect `Check-in`.
  Was the receptionist impolite? Was the waiting too long?

Overall rating: ★★★★★	*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

- The goal of this task is predicting, for each sentence 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.

Problem Definition

Overall rating:	*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

- $\mathbb{A}$: set of aspect labels (`Rooms`, `Cleanliness`, `Value`, `Service`, `Location`, `Check-in`, `Business`, `Food`, `Building`, `Other`);
- $\mathbb{Y}$: set of opinion labels (`Positive`, `Negative`, `Neutral`);
- $\mathbf{x}$: review composed of T consecutive sentences;
- For each sentence $t \in \{1, ..., T\}$ and each aspect $a \in \mathbb{A}$, we seek to infer the values of the opinion $y_t^a \in \mathbb{Y} \cup \{\text{No-op}\}$ (where No-op stands for "no opinion");

Outline

Linear-Chain CRFs Baseline

We adopt CRFs as a learning algorithm.

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

- Ψ_c: a factor;
- $\mathbf{F}$: the set of factors that model the distribution $p(\mathbf{y}|\mathbf{x})$;
- $Z(\mathbf{x})$: a normalization function.

Baseline: the traditional Linear-Chain (**LC**) CRF:

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

Multi-Label Models

In order to model the dependencies between the opinion related to different aspects we introduce the multi-label factor $\Psi_m(y_t^a, y_t^b)$

We first consider the *Independent Multi-Label* (**IML**) model:

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

IML can be combined with **LC** to obtain the *Chain Multi-Label* (**CML**) model:

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

Hierarchical (Multi-Label) Models

Jointly modeling the overall opinion y_o and the sentence-level opinions y_t^a in a hierarchical fashion can be beneficial to prediction at both levels:

$$\Phi(y_o, y_t^a, \mathbf{x}) = \Psi_o(y_o, \mathbf{x}_o) \cdot \Psi_h(y_t^a, y_o)$$

Hierarchical (Multi-Label) Models cont'd

LC, **IML**, **CML** can be adapted to include the overall rating variable into a hierarchical model structure; this produces:

1. the *Linear-Chain Overall* (**LCO**) model:

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

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

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

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

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

Features

We represent the sentence $\mathbf{x}_t$ via the following features:

- word unigrams and bigrams;
- polarity lexicon features:
  - General Inquirer;
  - MPQA;
  - SentiWordNet;
- aspect-specific lexicon features:
  - the lexicon gives the likelihood of the co-occurrence between words and aspects using the χ^2 measure.

We represent the entire review $\mathbf{x}_o$ via the following features:

- word unigrams and bigrams;
- polarity lexicon features:
  - General Inquirer;
  - MPQA;
  - SentiWordNet.

Inference and Learning

Problem:

- presence of loops in the graphs;
- exact inference is not tractable.

We revert to approximate inference via Gibbs sampling

- by adopting SampleRank as the learning algorithm: this is a natural fit for sampling-based inference;
- by using Gibbs sampling to obtain the MAP assignment.

Outline

Dataset

We have produced a new dataset[1] of manually annotated hotel reviews.

- three annotators annotated 442 randomly selected reviews from a publicly available TripAdvisor dataset for a total of 5799 sentences;
- the annotations are related to 9 aspects often present in hotel reviews (`Rooms`, `Cleanliness`, `Value`, `Service`, `Location`, `Check-in`, `Business`, `Food`, `Building`) plus the "catch-all" aspect `Other`;

[1] Available at http://nemis.isti.cnr.it/~marcheggiani/datasets/.

Dataset (cont'd)

- the annotation distinguishes between `Positive`, `Negative` and `Neutral/Mixed` opinions;
- out of the 442 reviews, 73 reviews were independently annotated by all three annotators (inter-annotator agreement);
- the remaining reviews were then partitioned into a training set (70%) and a test set (30%).

Evaluation Measures

In the evaluation phase we view the task as composed of the following two subtasks:

- *Aspect identification:*
  - standard F_1 measure

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

- *Opinion prediction:*
  - *macro-averaged mean absolute error* (MAE^M) to each applicable (true positive) aspect for the sentence

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

where $\mathbf{T}$ is the correct label assignments and $\widehat{\mathbf{T}}$ is the corresponding model predictions.

Evaluation Scenario

We perform two separate evaluations:

1. we compare the different models by their accuracy on the test set;
2. we compare the top-performing model to the human annotators on the set of 73 reviews independently annotated by all 3 annotators.

Since training is non-deterministic due to the use of sampling-based inference, we report the average over five trials with different random seeds.

Models Comparison Results

Table : Sentence-level aspect identification results in terms of F_1 (higher 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*

Table : Sentence-level opinion prediction results (restricted to the true positive aspects for each sentence) in terms of MAE^M (lower is better).

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	*Avg*
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*

Human Comparison Results

Table : F_1 results of the best-performing model (**IMLO**) and the human annotators (higher is better).

	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*

Table : MAE^M results of the best-performing model (**IMLO**) and the human annotators (lower is better).

	Other	Service	Rooms	Clean.	Food	Location	Check-in	Value	Building	Business	*Avg*
Human	.308	.219	.191	.259	.150	.202	.234	.003	.114	.029	**.171**
IMLO	.676	.498	.445	.142	.451	.704	.212	.387	.025	.415	*.396*

Outline

Conclusions

- We have devised a sequence of increasingly powerful CRF models:
  - Multi-label CRF models;
  - Hierarchical CRF models;
- We have produced and made available a manually annotated dataset of hotel reviews.
- Model comparison results:
  - **IML** and **CML** significantly outperform the **LC** baseline;
  - the hierarchical models improve the opinion prediction.
- Comparison with human performance:
  - much work remains to be done.

That's it!

Thanks for your attention!
Questions?