A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations

Angeliki Lazaridou
University of Trento
angeliki.lazaridou@unitn.it

Ivan Titov
Saarland University
titov@mmci.uni-saarland.de

Caroline Sporleder
Trier University
csporled@coli.uni-sb.de

Abstract

We propose a joint model for unsupervised induction of sentiment, aspect and discourse information and show that by incorporating a notion of latent discourse relations in the model, we improve the prediction accuracy for aspect and sentiment polarity on the sub-sentential level. We deviate from the traditional view of discourse, as we induce types of discourse relations and associated discourse cues relevant to the considered opinion analysis task; consequently, the induced discourse relations play the role of opinion and aspect shifters. The quantitative analysis that we conducted indicated that the integration of a discourse model increased the prediction accuracy results with respect to the discourse-agnostic approach and the qualitative analysis suggests that the induced representations encode a meaningful discourse structure.

1 Introduction

With the rapid growth of the Web, it is becoming increasingly difficult to discern useful from irrelevant information, particularly in user-generated content, such as product reviews. To make it easier for the reader to separate the wheat from the chaff, it is necessary to structure the available information. In the review domain, this is done in *aspect-based sentiment analysis* which aims at identifying text fragments in which *opinions* are expressed about ratable *aspects of products*, such as 'room quality' or 'service quality'. Such fine-grained analysis can serve as the first step in *aspect-based sentiment summarization* (Hu and Liu, 2004), a task with many practical applications.

Aspect-based summarization is an active research area for which various techniques have been developed, both statistical (Mei et al., 2007; Titov and McDonald, 2008b) and not (Hu and Liu, 2004), and relying on different types of supervision sources, such as sentiment-annotated texts or polarity lexica (Turney and Littman, 2002). Most methods rely on local information (bag-of-words, short ngrams or elementary syntactic fragments) and do not attempt to account for more complex interactions. However, these local lexical representations by themselves are often not sufficient to infer a sentiment or aspect for a fragment of text. For instance, in the following example taken from a TripAdvisor[1] review:

Example 1. *The room was nice **but** let's not talk about the view.*

it is difficult to deduce on the basis of local lexical features alone that the opinion about the view is negative. The clause *let's not talk about the view* could by itself be neutral or even positive given the right context (e.g., *I've never seen such a fancy hotel room, my living room doesn't look that cool... and let's not talk about the view*). However, the contrast relation signaled by the connective *but* makes it clear that the second clause has a negative polarity. The same observations can be made about transitions between aspects: changes in aspect are often clearly marked by discourse connectives. Importantly, some of these cues are not discourse connectives in the strict linguistic sense and are specific to the review domain (e.g., the phrase *I would also* in a review indicates that the topic is likely to be changed). In order to accurately predict sentiment and topic,[2] a model needs to ac-

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

[2] In what follows, we use the terms *aspect* and *topic*, inter-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1630–1639,
Sofia, Bulgaria, August 4-9 2013.

count for these discourse phenomena and cannot rely solely on local lexical information.

These issues have not gone unnoticed to the research community. Consequently, there has recently been an increased interest in models that leverage content and discourse structure in sentiment analysis tasks. However, discourse-level information is typically incorporated in a pipeline architecture, either in the form of sentiment polarity shifters (Polanyi and Zaenen, 2006; Nakagawa et al., 2010) that operate on the lexical level or by using discourse relations (Taboada et al., 2008; Zhou et al., 2011) that comply with discourse theories like Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). Such approaches have a number of disadvantages. First, they require additional resources, such as lists of polarity shifters or discourse connectives which signal specific relations. These resources are available only for a handful of languages. Second, relying on a generic discourse analysis step that is carried out before sentiment analysis may introduce additional noise and lead to error propagation. Furthermore, these techniques will not necessarily be able to induce discourse relations informative for the sentiment analysis domain (Voll and Taboada, 2007).

An alternative approach is to define a task-specific scheme of discourse relations (Somasundaran et al., 2009). This previous work showed that task-specific discourse relations are helpful in predicting sentiment, however, in doing so they relied on gold-standard discourse annotation at test time rather than predicting it automatically or inducing it jointly with sentiment polarity.

We take a different approach and induce discourse and sentiment information jointly in an unsupervised (or weakly supervised) manner. This has the advantage of not having to pre-specify a mapping from discourse cues to discourse relations; our model induces this automatically, which makes it portable to new domains and languages. Joint induction of discourse and sentiment structure also has the added benefit that the model is able to learn exactly those aspects of discourse structure that are relevant for sentiment analysis.

We start with a relatively standard joint model of sentiment and topic, which can be regarded as a cross-breed between the JST model (Lin and He, 2009) and the ASUM model (Jo and Oh, 2011), both state-of-the-art techniques. This model is weakly supervised, as it relies solely on document-level (i.e. not aspect-specific) opinion polarity labels to induce topics and sentiment on the sub-sentential level. In order to test our hypothesis that discourse information is beneficial, we add a discourse modeling component. Note that in modeling discourse we do not exploit any kind of supervision. We demonstrate that the resulting model outperforms the baseline on a product review dataset (see Section 5).

To the best of our knowledge, unsupervised joint induction of discourse structure, sentiment and topic information has not been considered before, particularly not in the context of the aspect-based sentiment analysis task. Importantly, our method for discourse modeling is a general method which can be integrated in virtually any LDA-style model of aspect and sentiment.

2 Modeling Discourse Structure

Discourse cues typically do not directly indicate sentiment polarity (or aspect). However, they can indicate how polarity (or aspect) changes as the text unfolds. As we have seen in the examples above, changes in polarity can happen on a sub-sentential level, i.e., between adjacent clauses or, from a discourse-theoretic point of view, between adjacent *elementary discourse units (EDUs)*. To model these changes we need a strong linguistic signal, for example, in the form of discourse connectives or other discourse cues. We hypothesize that these are more likely to occur at the beginning of an EDU than in the middle or at the end. This is certainly true for most of the traditional discourse relation cues (particularly connectives).

Changes in polarity or aspect are often correlated with specific discourse relations, such as 'contrast'. However, not all relations are relevant and there is no one-to-one correspondence between relations and sentiment changes.[3] Furthermore, if a discourse relation signals a change, it is typically ambiguous whether this change occurs with the polarity (example 1) or the aspect (*the room was nice but the breakfast was even better*) or both (*the room was nice but the breakfast was awful*). Therefore, we do not explicitly model

changeably as well as *sentiment levels* and *opinion polarity*.

[3] The 'explanation' relation, for example, can occur with a polarity change (*We were upgraded to a really nice room because the hotel made a terrible blunder with our booking)* but does not have to (*The room was really nice because the hotel was newly renovated*).

Name	Description
AltSame	**different** polarity, **same** aspect
SameAlt	**same** polarity, **different** aspect
AltAlt	**different** polarity and aspect

Table 1: Discourse relations

generic discourse relations; instead, inspired by the work of Somasundaran et al. (2008), we define three very general relations which encode how polarity and aspect change (Table 1). Note that we do not have a discourse relation *SameSame* since we do not expect to have strong linguistic evidence which states that an EDU contains the same sentiment information as the previous one.[4] However, we assume that the sentiment and topic flow is fairly smooth in general. In other words, for two adjacent EDUs not connected by any of the above three relations, the prior probability of staying at the same topic and sentiment level is higher than picking a new topic and sentiment level (i.e. we use "sticky states" (Fox et al., 2008)).

3 Model

In this section we describe our Bayesian model, first the discourse-agnostic model and then an extension needed to encode discourse information. The formal generative story is presented in Figure 1: the red fragments correspond to the discourse modeling component. In order to obtain the generative story for the discourse-agnostic model, they simply need to be ignored.

3.1 Discourse-agnostic model

In our approach we make an assumption that all the words in an EDU correspond to the same topic and sentiment level. We also assume that an overall sentiment of the document is defined, this is the only supervision we use in inducing the model. Unlike some of the previous work (e.g., (Titov and McDonald, 2008a)), we do not constrain aspect-specific sentiment to be the same across the document. We describe our discourse-agnostic model by first describing the set of corpus-level and document-level parameters, and then explain how the content of each document is generated.

Drawing model parameters On the corpus level, for every topic $z \in \{1, \ldots, K\}$ and every sentiment polarity level $y \in \{-1, 0, +1\}$, we start by drawing a unigram language model from a Dirichlet prior. For example, the language model of the aspect *service* may indicate that the word *friendly* is used to express a positive opinion, whereas the word *rude* expresses a negative one.

Similarly, for every topic z and every overall sentiment polarity $\hat{y}$, we draw a distribution $\psi_{\hat{y},z}$ over opinion polarity in this topic z. Intuitively, one would expect the sentiment of an aspect to more often agree with the overall sentiment $\hat{y}$ than not. This intuition is encoded in an asymmetric Dirichlet prior $Dir(\boldsymbol{\gamma}_{\hat{y}})$ for $\psi_{\hat{y},z} : \boldsymbol{\gamma}_{\hat{y}} = (\gamma_{\hat{y},1}, \ldots, \gamma_{\hat{y},M})$, $\gamma_{\hat{y},y} = \beta + \tau\delta_{y,\hat{y}}$, where $\delta_{y,\hat{y}}$ is the Kronecker symbol, β and τ are nonnegative scalar parameters. Using these "heavy-diagonal" priors is crucial, as this is the way to ensure that the overall sentiment level is tied to the aspect-specific sentiment level. Otherwise, sentiment levels will be specific to individual aspects (e.g., the "+1" sentiment for one topic may correspond to a "-1" sentiment for another one). Without this property we would not be able to encode soft constraints imposed by the discourse relations.

Drawing documents On the document level, as in the standard LDA model, we choose the distribution over topics for the document from a symmetric Dirichlet prior parametrized by α, which is used to control sparsity of topic assignments. Furthermore, we draw the global sentiment $\hat{y}_d$ from a uniform distribution.

The generation of a document is done on the EDU-by-EDU basis. In this work, we assume that EDU segmentation is provided by the preprocessing step. First, we generate the aspect $z_{d,s}$ for EDU s according to the distribution of topics θ_d. Then, we choose a sentiment level $y_{d,s}$ for the considered EDU from the categorical distribution $\psi_{\hat{y}_d,z_{d,s}}$, conditioned on the aspect $z_{d,s}$, as well as on the global sentiment of the document $\hat{y}_d$. Finally, we generate the bag of words for the EDU by drawing the words from the aspect- and sentiment-specific language model.

This model can be seen as a variant of a state-of-the-art model for jointly inducing sentiment and aspect at the sentence level (Jo and Oh, 2011), or, more precisely, as its combination with the JST model (Lin and He, 2009), adapted to the specifics of our setting. Both these models have been shown to perform well on sentiment and topic prediction tasks, outperforming earlier models, such as the TSM model (Mei et al., 2007). Consequently, it can be considered as a strong baseline.

[4] The typical connective in this situation would be *and* which is highly ambiguous and can signal several traditional discourse relations.

3.2 Discourse-informed model

In order to integrate discourse information into the discourse-agnostic model, we need to define a set of extra parameters and random variables.

Drawing model parameters First, at the corpus level, we draw a distribution $\tilde{\varphi}$ over four discourse relations: three relations as defined in Table 1 and an additional dummy relation 4 to indicate that there is no relation between two adjacent EDUs ($NoRelation$). This distribution is drawn from an asymmetric Dirichlet prior parametrized by a vector of hyperparameters $\boldsymbol{\nu}$. These parameters encode the intuition that most pairs of EDUs do not exhibit a discourse relation relevant for the task (i.e. favor $NoRelation$), that is ν_4 has a distinct and larger value than other parameters $\nu_{\bar{4}}$.

Every discourse relation c (including $NoRelation$ which is treated here as *Same-Same*) is associated with two groups of transition distributions, one governing transitions of sentiment ($\tilde{\psi}_c$) and another one controlling topic transitions ($\tilde{\theta}_c$). The parameter $\tilde{\psi}_{c,y_s}$, defines a distribution over sentiment polarity for the EDU $s+1$ given the sentiment for the sth EDU y_s and the discourse relation c. This distribution encodes our beliefs about sentiment transitions between EDUs s and $s+1$ related through c. For example, the distribution $\tilde{\psi}_{SameAlt,+1}$ would assign higher probability mass to the positive sentiment polarity ($+1$) than to the other 2 sentiment levels (0, -1). Similarly, the parameter $\tilde{\theta}_{c,z_s}$, defines a distribution over K aspects.

These two families of transition distributions are each defined in the following way. For the distribution $\tilde{\theta}$, for relations that favor changing the aspect (*SameAlt* and *AltAlt*), the probability of the preferred (K-1) transitions is proportional to ω_θ and for the remaining transitions it is proportional to 1. On the other hand, for the relations that favor keeping the same aspect (*NoRelation* and *Alt-Same*), the probability of the preferred transition is proportional to ω'_θ, whereas the probability of the (K-1) remaining transitions is again proportional to 1. For the sentiment transitions, the distribution $\tilde{\psi}_{c,y_s}$ is defined in the analogous way (but depends on ω_ψ and ω'_ψ). These scalars are hand-coded and define soft constraints that discourse relations impose on the local flow of sentiment and aspects.

The parameter $\tilde{\phi}_c$ is a language model over discourse cues $\tilde{w}$, which are not restricted to unigrams but can generate phrases of arbitrary (and variable) size. For this reason, we draw them from a Dirichlet process (DP) (i.e. one for each discourse relation, except for $NoRelation$). The base measure G_0 provides the probability of an n-word sequence calculated with the bigram probability model estimated from the corpus.[5] This model component bears strong similarities to the Bayesian model of word segmentation (Goldwater et al., 2009), though we use the DP process to generate only the prefix of the EDU, whereas the rest of the EDU is generated from the bag-of-words model.

Drawing documents As pointed out above, the content generation is broken into two steps, where first we draw the discourse cue $\tilde{w}_{d,s}$ from $\tilde{\phi}_c$ and then we generate the remaining words.

The second difference at the data generation step (Figure 1) is in the way the aspect and sentiment labels are drawn. As the discourse relation between the EDUs has already been chosen, we have some expectations about the values of the sentiment and aspect of the following EDU, which are encoded by the distributions $\tilde{\psi}$ and $\tilde{\theta}$. These are only soft constraints that have to be taken into consideration along with the information provided by the aspect-sentiment model. This coupling of information naturally translates into the *product-of-experts* (PoE) (Hinton, 1999) approach, where two sources of information jointly contribute to the final result. The PoE model seems to be more appropriate here than a mixture model, as we do not want the discourse transition to overpower the sentiment-topic model. In the PoE model, in order for an outcome to be chosen, it needs to have a non-negligible probability under both models.

4 Inference

Since exact inference of our model is intractable, we use collapsed Gibbs sampling. The variables that need to be inferred are the topic assignments $\mathbf{z}$, the sentiment assignments $\mathbf{y}$, the discourse relations $\mathbf{c}$ and the discourse cue $\tilde{\boldsymbol{w}}$ (or, more precisely, its length) and are all sampled jointly (for each EDU) since we expect them to be highly dependent. All other variables (i.e. unknown distributions) could be marginalized out to obtain a collapsed Gibbs sampler (Griffiths and Steyvers, 2004).

[5] This measure is improper but it serves the purpose of favoring long cues, the behavior arguably desirable for our application.

Global parameters:

$\tilde{\varphi} \sim Dir(\nu)$ **[distrib of disc rel]**
for each discourse relation $c = 1, .., 4$:
$\tilde{\phi}_c \sim \text{DP}(\eta, G_o)$ **[distrib of disc rel specific disc cues]**
$\tilde{\theta}_{c,k}$ **- fixed [distrib of rel specific aspect transitions]**
$\tilde{\phi}_{c,y}$ **- fixed [distrib of rel specific sent transitions]**
for each aspect $k = 1, 2...K$:
for each sentiment $y = -1, 0, +1$:
$\phi_{k,y} \sim Dir(\boldsymbol{\lambda}_k)$ [unigram language models]
for each global sentiment $\hat{y} = -1, 0, +1$:
$\psi_{\hat{y},k} \sim Dir(\boldsymbol{\gamma})$ [sent distrib given overall sentiment]

Data Generation:

for each document d:
$\hat{y}_d \sim Unif(-1, 0, +1)$ [global sentiment]
$\theta_d \sim Dir(\alpha)$ [distr over aspects]
for every EDU s:
$\boldsymbol{c_{d,s}} \sim \tilde{\varphi}$ **[draw disc relation]**
if $c_{d,s} \neq NoRelation$
$\tilde{\boldsymbol{w}}_{\boldsymbol{d,s}} \sim \tilde{\phi}_{c_{d,s}}$ **[draw disc cue]**
$z_{d,s} \sim \theta_d * \tilde{\theta}_{c_{d,s}, z_{d,s-1}}$ [draw aspect]
$y_{d,s} \sim \psi_{\hat{y}_d, z_{d,s}} * \tilde{\psi}_{c_{d,s}, y_{d,s-1}}$ [draw sentiment level]
for each word after disc cue:
$w_{d,s} \sim \phi_{z_{d,s}, y_{d,s}}$ [draw words]

Figure 1: The generative story for the joint model. The components responsible for modeling discourse information are emphasized in red: when dropped, one is left with the discourse-agnostic model.

Unfortunately, the use of the PoE model prevents us from marginalizing the parameters exactly. Instead, as in Naseem et al. (2009), we resort to an approximation. We assume that $z_{d,s}$ and $y_{d,s}$ are drawn twice; once from the document specific distribution and once from the discourse transition distributions. Under this simplification, we can easily derive the conditional probabilities for the collapsed Gibbs sampling.

5 Experiments

To the best of our knowledge, this is the first work that aims at evaluating directly the joint information of the sentiment and aspect assignment at the sub-sentential level of full reviews; most existing studies either focus on indirect evaluation of the produced models (e.g., classifying the overall sentiment of sentences (Titov and McDonald, 2008a; Brody and Elhadad, 2010) or even reviews (Nakagawa et al., 2010; Jo and Oh, 2011)) or evaluated solely at the sentential or even document level. Consequently, in order to evaluate our methods, we created a new dataset which will be publicly released.

Aspects	**Frequency**
service	246
value	55
location	121
rooms	316
sleep quality	56
cleanliness	59
amenities	180
food	81
recommendation	121
rest	306
Total	1541

Table 2: Distribution of aspects in the data.

Dataset and Annotation The dataset we created consists of 13559 hotel reviews from TripAdvisor.com.[6] Since our modeling is performed on the EDU level, all sentences where segmented using the SLSEG software package.[7] As a result, our dataset consists of 322,935 EDUs.

For creating the gold standard, 9 annotators annotated a random subset of our dataset (65 reviews, 1541 EDUs). The annotators were presented with the whole review partitioned in EDUs and were asked to annotate every EDU with the aspect and sentiment (i.e. $+1$, 0 or -1) it expresses. Table 2 presents the distribution of aspects in the dataset. The distribution of the sentiments is uniform. The label *rest* captures cases where EDUs do not refer to any aspect or to a very rare aspect. The inter-annotator agreement (IAA), as measured in terms of Cohen's kappa score, was 66% for the aspect labeling, 70% for the sentiment annotation and 61% for the joint task of sentiment and aspect annotation. Though these scores may not seem very high, they are similar to the ones reported in related sentiment annotation efforts (see e.g., Ganu et al. (2009)).

Experimental setup In order to quantitatively evaluate the model predictions, we run two sets of experiments. In the first, we treat the task as an unsupervised classification problem and evaluate the output of the models directly against the gold standard annotation. This is a very challenging set-up, as the model has no prior information about the aspects defined (Table 2). In the second set of experiments, we show that aspects and sentiments induced by our model can be used to construct informative features for supervised classification. In

[6] Downloadable from `http://clic.cimec.unitn.it/~angeliki.lazaridou/datasets/ACL2013Sentiment.tar.gz`

[7] `www.sfu.ca/~mtaboada/research/SLSeg.html`

Model	Precision	Recall	F1
Random	3.9	3.8	3.8
SentAsp	15.0	10.2	9.2
Discourse	**16.5**	**13.8**	**10.8**

Table 3: Results in terms of macro-averaged precision, recall and F1.

Model	Unmarked	Marked
SentAsp	9.2	5.4
Discourse	**9.3**	**11.5**

Table 4: Separate evaluation (F1) of the "marked" and the "unmarked" EDUs.

all the cases, we compare the discourse-agnostic and the discourse-informed models.

In order to induce the model, we let the sampler run for 2000 iterations. We use the last sample to define the labeling. The number of topics K was set to 10 in order to match the number of aspects defined in our annotation scheme (see Table 2). The hyperpriors were chosen in a qualitative experiment over a subset of our dataset by manually inspecting the produced languages models. The resulting values are: $\alpha = 10^{-3}$, $\beta = 5 * 10^{-4}$, $\tau = 5 * 10^{-4}$, $\eta = 10^{-3}$, $\nu_4 = 10^3$, $\nu_{\bar{4}} = 10^{-4}$, $\omega_\theta = 85$ and $\omega'_\theta = \omega_\psi = \omega'_\psi = 5$.

5.1 Direct clustering evaluation

Our labels encoding aspect and sentiment level can be regarded as clusters. Consequently we can apply techniques developed in the context of clustering evaluation. We use a version of the standard metrics considered for the word sense induction task (Agirre and Soroa, 2007) where a clustering is converted to a classification problem. This is achieved by splitting the gold standard into two subsets; the training portion is used to choose one-to-one correspondence from the gold classes to the induced clusters and then the chosen mapping is applied to the testing portion. We perform 10-fold cross validation and report precision, recall and F1 score. Our dataset is very skewed and the majority class (*rest*) is arguably the least important, so we use macro-averaging over labels and then average those across folds to arrive to the reported numbers. We compare the discourse-informed model (*Discourse*) against two baselines; the discourse-agnostic *SentAsp* model and *Random* which assigns a random label to an EDU while respecting the distribution of labels in the training set.

Table 3 presents the first analysis conducted on the full set of EDUs. We observe that by incorporating latent discourse relation we improve performance over the discourse-agnostic model *SentAsp* (statistically significant according to paired t-test with $p < 0.01$). Note that fairly low scores in this evaluation setting are expected for any unsupervised model of sentiment and topics, as models are unsupervised both in the aspect-specific sentiment and in topic labels and the total number of labels is 28 (all aspects can be associated with the 3 sentiment levels except for *rest* which can only be used with neutral (0) sentiment). Consequently, induced topics, though informative (as we confirm in Section 5.3), may not correspond to the topics defined in the gold standard. For example, one well-known property of LDA-style topic models is their tendency to induce topics which account for similar fraction of words in the dataset (Jagarlamudi et al., 2012), thus, over-splitting 'heavy' topics (e.g. *rooms* in our case). The same, though to lesser degree, is true for sentiment levels where the border between neutral and positive (or negative) is also vaguely defined.

	Content	Aspect	Polarity
1	*but* certainly off **its** greatness	value	neg
2	*and* while small **they** are nice	rooms	pos
3	*but* **it** is not free for all guests	amenities	neg
4	*and* the water was brown	clean	neg
5	*and* no tea making facilities	rooms	neg
6	*when* i checked out	service	pos
7	*and* if you do not	service	neg
8	*when* we got home	clean	neu

Table 5: Examples of EDUs where local information is not sufficiently informative.

To gain insight into our model, we conducted an experiment similar to the one presented in Somasundaran et al. (2009). We divide the dataset in two subsets; one containing all EDUs starting with a discourse cue ("marked") and one containing the remaining EDUs ("unmarked"). We hypothesize that the effect of the discourse-aware model should be stronger on the first subset, since the presence of the connective indicates the possibility of a discourse relation with the previous EDU. The set of discourse connectives is taken from the Penn Discourse Treebank (Prasad et al., 2008), thus creating a list of 240 potential connectives.

Table 5 presents a subset of "marked" EDUs for which trying to assign the sentiment and aspect out of context (i.e. without the previous EDU) is a difficult task. In examples 1-3 there is no explicit mention of the aspect. However, there is an anaphoric expression (marked in bold) which

refers to a mention of the aspect in some previous EDU. On the other hand, in examples 4 and 5 there is an ambiguity in the choice of aspect; in example 5, *tea making facilities* can refer to a breakfast at the hotel (label *food*) or to facilities in the room (label *rooms*). Finally, examples 6-8 are too short and not informative at all which indicates that the segmentation tool does not always predict a desirable segmentation. Consequently, automatic induction of segmentation may be a better option.

Table 4 presents quantitative results of this analysis. Although the performance over the "unmarked" example is the same for the two models, this is not the case for the "marked" instances where the discourse-informed model leverages the discourse signal and achieves better performance. This behavior agrees with our initial hypothesis, and suggests that our discourse representation, though application-specific, relies in part on the information encoded in linguistically-defined discourse cues. We will confirm this intuition in the qualitative evaluation section. The increase for the "marked" EDUs does not translate into greater differences for the overall scores (Table 3) as marked relations are considerably less frequent than unmarked ones in our gold standard (i.e. 35% of the EDUs are "marked"). Nevertheless, this clearly suggests that the discourse-informed model is in fact capable of exploiting discourse signal.

5.2 Qualitative analysis

To investigate the quality of the induced discourse structure, we present the most frequent discourse cues extracted for every discourse relation. Table 6 presents a selection of cues that best explain the discourse relation they have been associated with. A general observation is that among the cues there are not only "traditional" discourse connectives like *even though, although, and*, but also cues that are discriminative for the specific application.

In relation *SameAlt* we can mostly observe phrases that tend to introduce a new aspect, since an explicit mention of it is provided (e.g *the location is, the room was*) and more specific phrases like *in addition* are used to introduce a new aspect with the same sentiment. However, these cues reveal important information about the aspect of the EDU, and since they are associated with the language model $\tilde{\phi}$, they are not visible anymore to the language model of aspects ϕ.

Cues for the relation *AltSame* also include phrases that contain some anaphoric expressions, which might refer to previous mentions of an aspect in the discourse (i.e. previous EDU). We expect that since there is an anaphoric expression, explicit lexical features for the aspect will be missing, making thus the decision concerning aspect assignment ambiguous for any discourse-agnostic model. Interestingly, we found the expressions *unfortunately, fortunately, the only thing* in the same relation, since all indicate a change in sentiment. Finally, *AltAlt* can be viewed as a mixture of the other two relations. Furthermore, for this relation we can find expressions that tend to be used at the end of a review, since at this point we normally change the aspect and often even sentiment. Some examples of these cases are *overall, we will definitely* and even the misspelled version of the latter *i would definately*.

Discourse relation	Discourse Cues
SameAlt	**the location is** , **the room was**, the hotel has, **and the room**, and the bed, breakfast was, the staff were, **in addition**, good luck
AltSame	**but**, and, it was, and it was, and they, although, and it, but it, **but it was**, however, which was, this is, **this was**, they were, the only thing, even though, **unfortunately**, needless to say, fortunately
AltAlt	**the room was**, the staff were, **the only**, the hotel is, but the, however, also, or, **overall i**, unfortunately, **we will definitely**, on the plus, the only downside , even though, and even though, **i would definately**

Table 6: Induced cues from the discourse relations

5.3 Features in supervised learning

As an additional experiment to demonstrate informative of the output of the two models, we design a supervised learning task of predicting sentiment and topic of EDUs. In this setting, the feature vector of every EDU consists of its bag-of-word-representation to which we add two extra features; the models' predictions of topic and sentiment. We train a support vector machine with a polynomial kernel using the default parameters of Weka[8] and perform 10-fold cross-validation.

Table 7 presents results of this analysis in terms of accuracy for four classification tasks, i.e. predicting both sentiment and topic, only sentiment and only topic for all EDUs, as well as predicting sentiment and topic for the "marked" dataset. First, we observe that incorporation of the topic-

[8] http://www.cs.waikato.ac.nz/ml/weka/

Features	**aspect+sentiment** (28 classes)	**aspect** (10 classes)	**sentiment** (3 classes)	**Marked only sentiment+aspect** (28 classes)
only unigrams	36.3	49.8	57.1	26.2
unigrams + SentAsp	38.0	50.4	59.3	27.8
unigrams + Discourse	**39.1**	**52.4**	**59.4**	**29.1**

Table 7: Supervised learning at the EDU level (accuracy)

model features on a *unigram-only* model results in an improvement in classification performance across all tasks (predicting sentiment, predicting aspects, or both); as a matter of fact, our accuracy results for predicting sentiment are comparable to the sentence-level results presented by Täckström and McDonald (2011). We have to stress that accuracies for the joint task (i.e. predicting both sentiment and topic) are expected to be lower since it can also be seen as the product of the two other tasks (i.e. predicting only sentiment and only topic). We also observe that the features induced from the *Discourse* model result in higher accuracy than the ones from the discourse-agnostic model *SentAsp* both in the complete set of EDUs and the "marked" subset, results that are in line with the ones presented in Table 4. Finally, the fact that the results for the complete set of EDUs are higher than the ones for the "marked" dataset clearly suggests that the latter constitute a hard case for sentiment analysis, in which exploiting discourse signal proves to be beneficial.

6 Related Work

Recently, there has been significant interest in leveraging content structure for a number of NLP tasks (Webber et al., 2011). Sentiment analysis has not been an exception to this and discourse has been used in order to enforce constraints on the assignment of polarity labels at several granularity levels, ranging from the lexical level (Polanyi and Zaenen, 2006) to the review level (Taboada et al., 2011). One way to deal with this problem is to model the interactions by using a precompiled set of polarity shifters (Nakagawa et al., 2010; Polanyi and Zaenen, 2006; Sadamitsu et al., 2008). Socher et al. (2011) defined a recurrent neural network model, which, in essence, learns those polarity shifters relying on sentence-level sentiment labels. Though successful, this model is unlikely to capture intra-sentence non-local phenomena such as effect of discourse connectives, unless it is provided with syntactic information as an input. This may be problematic for the noisy sentiment-analysis domain and especially for poor-resource languages. Similar to our work, others have focused on modeling interactions between phrases and sentences. However, this has been achieved by either using a subset of relations that can be found in discourse theories (Zhou et al., 2011; Asher et al., 2008; Snyder and Barzilay, 2007) or by using directly (Taboada et al., 2008) the output of discourse parsers (Soricut and Marcu, 2003). Discourse cues as predictive features of topic boundaries have also been considered in Eisenstein and Barzilay (2008). This work was extended by Trivedi and Eisenstein (2013), where discourse connectors are used as features for modeling subjectivity transitions.

Another related line of research was presented in Somasundaran et al. (2009) where a domain-specific discourse scheme is considered. Similarly to our set-up, discourse relations enforce constraints on sentiment polarity of associated sentiment expressions. Somasundaran et al. (2009) show that gold-standard discourse information encoded in this way provides a useful signal for prediction of sentiment, but they leave automatic discourse relation prediction for future work. They use an integer linear programming framework to enforce agreement between classifiers and soft constraints provided by discourse annotations. This contrasts with our work; we do not rely on expert discourse annotation, but rather induce both discourse relations and cues jointly with aspect and sentiment.

7 Conclusions and Future Work

In this work, we showed that by jointly inducing discourse information in the form of discourse cues, we can achieve better predictions for aspect-specific sentiment polarity. Our contribution consists in proposing a general way of how discourse information can be integrated in any LDA-style discourse-agnostic model of aspect and sentiment. In the future, we aim at modeling more flexible sets of discourse relations and automatically inducing discourse segmentation relevant to the task.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the SemEval*, pages 7–12.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling opinion in discourse: A preliminary study. *Proceedings of Coling*, pages 5–8.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL*, pages 804–812.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. 2008. An HDP-HMM for systems with state persistence. In *Proceedings of ICML*.

Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of WebDB*.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Geoffrey E Hinton. 1999. Products of experts. In *Proceedings of ICANN*, volume 1, pages 1–6.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*, pages 168–177.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. *Proceedings of EACL*, pages 204–213.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*, pages 815–824.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceeding of CIKM*, pages 375–384.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Proceedings of NAACL*, pages 786–794.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36(1):341–385.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*.

Kugatsu Sadamitsu, Satoshi Sekine, and Mikio Yamamoto. 2008. Sentiment analysis based on probabilistic models using inter-sentence information. In *Proceedings of ACL*, pages 2892–2896.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *Proceedings of HLT-NAACL*, pages 300–307.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, pages 151–161.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of Coling*, pages 801–808.

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*, pages 170–179.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL*, pages 149–156.

Maite Taboada, Kimberly Voll, and Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser University, Tech. Rep*, 20.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of ACL*, pages 569–574.

Ivan Titov and Ryan McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316.

Ivan Titov and Ryan McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*, pages 112–120.

Rakshit Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *In Proceedings of NAACL*.

Peter D Turney and Michael L Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Kimberly Voll and Maite Taboada. 2007. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In *Proceedings of Australian Conf. on AI*.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, 1(1):1–54.

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings EMNLP*, pages 162–171.