UNIVERSITÄT
DES
SAARLANDES

Master's Thesis

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

submitted in partial fulfillment of the requirements
for the degree of Master of Science in Language Science and Technology

Angeliki Lazaridou

Supervisors:
Dr. Ivan Titov
Dr. Caroline Sporleder

October 29, 2012

Acknowledgments

First of all, I would like to express my gratitude to my supervisors Ivan Titov and Caroline Sporleder. Ivan's door had been literally always open and he was there to answer all questions, no matter how difficult or naive they were. But most importantly, he gave me material to think, be creative and critical of my work and of others. Caroline provided me with very useful insights and ideas concerning the discourse part. I would also like to thank Cristophe Cerisara who commented on earlier versions of this thesis.

I would like to thank all my colleagues in Saarland and Nancy for the funny moments we had together during these two very "international" years. Special thanks to my 8 annotators Evi, Lea, Iliana, Antonia, Mikhail, Milos, Jesus and Richard for volunteering to help me with the annotations.

I would like to thank LCT for the financial support. Especially, I would like Dr. Kordoni, Ms. Pernice and Dr. Pogodalla for taking care of all the organizational stuff and letting time for us the students to focus on our studies.

I wouldn't have made it so far without the real and genuine support of my family and especially my mother's, grandmother's and grandfather's love and guidance.

Finally, I would like to thank Nikos for being with me throughout the happy but also the difficult moments of this master, for believing in me more than I do and for making me laugh.

Declaration

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 29 October 2012

Signature:

Abstract

In this work, we develop a generative framework for jointly inducing information related to sentiment analysis of opinionated texts. The joint induction of sentiment and aspect is done on the sub-sentential level, thus yielding a fine-grained analysis. We argue that by incorporating discourse information, we can achieve more accurate estimations. In particular, we deviate from the "traditional" view of discourse, and we model a discourse structure appropriate for the particular task. This is achieved by designing a Bayesian model, where priors encode our beliefs about the different discourse classes as well as the constraints they impose to the local structure. Our model is thus able to induce discriminative cue phrases which indicate that a change of discourse is about to happen. While the quantitative analysis that we conducted indicated that learning a discourse model suitable for this task significantly increased the results of the aspect-based sentiment analysis over a discourse-agnostic approach, the qualitative analysis confirmed that the induced representation is a meaningful discourse structure.

Contents

CHAPTER 1

Introduction

Sentiment analysis, also known as opinion mining (two interchangeable terms as pointed out in Pang and Lee [1]), has been a very vivid area of research during the past few years. Its main goal lies in capturing the emotion that is expressed in texts. The type of texts usually ranges from semi-structured product reviews which are richer in sentiment information, to highly structured news articles, where opinions are usually more difficult to be inferred due to the objective goal and factual nature of these documents. More recently, research has also moved to revealing opinions in social media platforms and blogs, where texts often exhibit a looser structure.

1.1 Aspect-based sentiment analysis

One particular task of sentiment analysis that has gained a lot of interest is the one of aspect-based sentiment analysis. With a fast growing web, the amount of information provided online is constantly growing, thus making it difficult to navigate through all the reviews and draw inferences. The structuring of the available information is explored by aspect-based sentiment analysis, where the problem focuses on identifying opinions exhibited in opinionated texts about ratable aspects of products. Furthemore, such fined-graned analysis can serve as the basis for sentiment summarization, a task very popular, especially in the industry. Algorithms for this task have been incorporated in platforms like Google products[1] and Bing Shopping[2]; by exploiting a mass of user reviews, these platforms aim at producing summarization templates consisting of sentiment information (usually in the form of ratings) for all the ratable aspects identified in the individual reviews (Figure 1.1) and further aggregate the results of reviews to produce overall ratings for individual products (Figure 1.2).

[1] http://www.google.com/shopping

[2] http://www.bing.com/shopping

Figure 1.1: One of the many available reviews for the hotel "Hotel des Prelats"

Figure 1.2: The aggregated results for all the reviews of the particular hotel resulted in the overall rating of 5

Techniques for dealing with the problem of aspect-based sentiment analysis vary from fully probabilistic supervised or unsupervised (Jo and Oh [2]) frameworks, which exploit a set of features often consisting of lexical features to knowledge-rich techniques that exploit various source of information like polarity lexica (Turney and Littman [3]) or grammar patterns (Hatzivassiloglou and McKeown [4]). However, in some cases neither probabilistic nor lexical information is enough to infer the opinion of a text; in the sentence extracted by the review in Figure 1.1

Example 1. *The bathroom was also spacious,* ***but*** *the toilet is separated from the room by a half-wall ... so if you're traveling with a friend, make sure it's a good friend.*

it is very difficult to infer that the opinion about the bathroom is both positive and negative. Even though the first part of the sentence clearly reveals the positive opinion of the writer towards the bathroom, at the second part of the sentence, the information is more implicit. However, for the reader it is easier to identify this information since there is some specific linguistic structure predictive for the sentiment flow between the parts of the sentence. This linguistic structure in our case is the discourse information.

This observation has motivated the research community to investigate ways to

make available discourse information of any form in various sentiment analysis tasks (Sauper et al. [5], Somasundaran et al. [6], Taboada et al. [7]). Usually, research that aims at leveraging discourse structure, has to integrate it as an external resource in the architecture pipeline. This external information might come as an output from a discourse parser or from annotated data. However, in these techniques, we can identify two major disadvantages. First, techniques relying on the output of discourse parsers have to come across a natural error propagation; the results that have been reported for discourse parsers do not go beyond 50% in F-score (Soricut and Marcu [8]). Thus, it becomes clear that having the full output of these parsers is a very demanding process and one can argue that sentiment analysis can also benefit from something simpler than a full discourse parse. Second, and most important, current theories of discourse(Mann and Thompson [9]) enumerate a great deal of discourse classes. These theories were not built having in mind some specific task and therefore many of the fine-grained classes are not relevant for the sentiment analysis task. As a result, these discourse classes have to go through a post-processing phase in order to represent more coarse-grained information.

Thus, the questions that this work tries to answer can be formulated in the following way:

1. Can we learn a model of discourse appropriate for the sentiment analysis domain?

2. Can this model of discourse be useful for other tasks?

1.2 Joint modeling of sentiment, aspect and discourse information on the sub-sentential level

In order to create summarization templates like the one in Figure 1.1, there is a need of analysis of information on a fined-grained level; the sub-sentential level. Since this detailed level of representation requires a lot of fine-grained annotation in order to train a supervised system, we thus propose a fully unsupervised framework. We develop a generative model where we jointly induce three kinds information that appear in semi-structured opinionated reviews; the sentiment, the aspect and the discourse structure. Since we expect that the discourse information will favour the sentiment analysis task, we decide not to restrict our model working on the sentence-level, but rather go deeper into the structure of the sentence and induce information for a meaningful segmentation of the sentence. We thus have these three types of information as latent variables

and we try to learn through a Bayesian model the values of these variables for every sub-sentential part.

At the core of our approach is the hypothesis that given a segmentation, the first few words of every segment indicate some predisposition for the sentiment and the aspect of the current segment. In other words, the aspect and sentiment of every segment should adhere to some intra-sentential soft and flexible constraints suggested by the discourse structure. This is motivated by the Example 1, where the word **but** acts as a signal that the information about the sentiment and/or the aspect will be affected. Describing the latter in a generative process, we assume that every segment is being generated by some discourse class which poses some constraints on the next segment under a Markovian assumption. These classes are signaled by the discourse-class-specific discourse cues, which are not known beforehand and therefore are induced by the generative process. It is very important to note that since all this information is combined in an unsupervised framework, we do not expect to induce discourse information in the same form as the one established in current theories. Therefore, these classes define the local structure that we expect to find in opinionated texts and are the ones that we believe will have a positive impact on the performance of our system.

1.3 Contributions

To the best of our knowledge, this is the only research that tries to leverage all three kinds of information in a joint framework for discovering finally the sentiment, aspect and discourse structure in a sub-sentential level. At the same time, since our algorithm uses no pre-compiled list of discourse cues, we are also discovering those cues that are most discriminant for the sentiment domain, which can afterwards be incorporated in platforms that automatically generate sentiment summaries in order to generate summaries that are better structured and exhibit more naturalness. In this work, we expect that the induction of discourse structure has more to gain from the sentiment analysis task than the other way around, since our target domain contains texts which are sentiment-oriented but exhibit looser structure. However, our method can serve as a good starting point for structured texts where discourse signal tends to be more significant and where sentiment information is given implicitly (e.g. news articles).

Finally, even though this work is not treating discourse segmentation and assumes that some oracle provides us with this information, the simplicity of the model, as we will see in the next session, provides a natural framework for the incorporation of other

tasks, including segmentation.

1.4 Thesis outline

The thesis is structured in the following way: Section 2 presents previous work done in aspect-based sentiment analysis, as well as work in sentiment analysis that leverages discourse information. Section 3 discusses necessary technical information concerning Bayesian Modeling. Section 4 describes our generative model of aspect, sentiment and discourse information and gives details of the inference. Section 5 provides information concerning the experimental setup, including information about the dataset, the manual annotation, the preprocessing as well as the metrics used to evaluate our model against the baseline. Section 6 presents the results of our method and finally 7 summarizes the work done in this thesis as well as providing our future research directions.

CHAPTER 2

Related Work

In this thesis, we draw on previous work in two main areas. First, we are focusing on the problem of sentiment analysis and more specifically on the aspect-based sentiment analysis and we provide literature of research that treats the problem both in a pipeline (see Section 2.1.1) (i.e aspect identification and the sentiment classification) and in a joint framework (see Section 2.1.2). Furthermore, we review different methods for using discourse information in sentiment analysis tasks that make use of the RST theory (see Section 2.2.3), polarity shifters (see Section 2.2.2) as well as methods that leverage local structure (see Section 2.2.4).

2.1 Aspect-based sentiment analysis

Aspect-based sentiment analysis has already received a considerable attention in the research community and is considered the next step after document-level polarity classification, where the goal is to assign to the whole document, referring to a product, the global opinion towards this product. However, it is usually the case that the global information about products is already provided by the users in the form of a rating, so the need moves to acquiring more fine-grained information which is not provided explicitly in blogs or even online product reviews, like the opinion information concerning specific aspects/features of the product. Figure 2.1 illustrates a hotel review from TripAdvisor. The user's global opinion concerning the hotel is indicated in the blue box. Green boxes indicate aspects of the hotel, whereas the underlined sentence refers to the aspect "location" without explicitly using this keyword. This research has mainly developed in two main paths; the methods that approach the task through a pipeline of subtasks, namely the aspect extraction and the sentiment classification and the research that deals with the problem in a joint framework based on variants of topic

models. Our work falls within the latter direction.

Figure 2.1: A review as appeared in TripAdvisor referring to the hotel *Hotel Des Prelats*.

2.1.1 Two-stage aspect-based sentiment analysis

Aspect-based sentiment analysis usually consists of two main phases. Traditionally, the task of identifying properties of a product has been cast as an information extraction problem. Research on this has developed in different ways with the main axis being in applying different filters to reduce noise from noun phrases. Hu and Liu [10] followed a frequency-based approach where noun phrases that are frequently talked about are considered salient properties of reviewed products, resulting in high recall. In a more linguistically-informed framework, Popescu and Etzioni [11] improved over the latter method by aiming at better precision. The candidate noun phrases where pruned using Wordnet and morphological cues. Such approaches work well in detecting aspects that are strongly associated with a single noun (e.g battery life in product reviews), but are less useful when aspects encompass many low frequency terms, that are difficult to be discovered. One example of such aspect is the aspect *room* which is usually found in hotel reviews and includes references to different dimensions of it, such us the view, the decoration, the size etc.

More recently, attention has been drawn to Bayesian methods. More specifically, several methods have adapted a widely-known topic modeling algorithm known as *Latent Dirichlet Allocation*. In LDA, document are viewed as mixtures of topics and a topic is defined as a probability distribution over words. It has been shown (Titov and McDonald [12]) that applying LDA in the traditional way for this task is problematic, since all reviews for a product talk about the aspect. In other words, it models topics based on document-level co-occurrences and thus these topics tend to cover the main topic of the review (i.e. the product for which the review was written). Therefore, these topics globally (i.e. on the document level) classify terms into product instances (e.g.

in the *hotel* domain, we would obtain topics about hotels in New York and hotels in Las Vegas).

Among possible solutions to remedy this problem, is to consider a sentence as a document, and therefore apply LDA on single sentences (Brody and Elhadad [13]). However, as Titov and McDonald [12] have stated, for the specific task, applying LDA on the sentence level will harm the performance since the co-occurrence domain is not large enough. For this reason, they propose a *Multi-Grain Latent Dirichlet Allocation*. In MG-LDA, a word can be sampled from two distinct types of topics: global topics and local topics. The intuition behind this model is that the global topics will discover topics that will be indicative for product types (e.g different hotels, different cities etc), whereas local ones will correspond to ratable aspects.

Having discovered the aspects, the next step is then to classify the opinions expressed. Opinion classification on the sentence-level usually proceeds with a subjectivity analysis, where the goal is to determine whether a sentence is subjective or objective. A sentence is subjective if it contains one or more opinions towards some topic; otherwise, the sentence is objective (Wilson [14]). In our setting, we define objective sentences to be those that express facts or "average" opinions (e.g. The breakfast was OK) and we model them by going beyond the binary classification and introducing a third label, *neutral*.

For the opinion classification, different approaches have been proposed, ranging from fully supervised approaches that require a heavy feature engineering phase to corpus-based but knowledge-rich approaches. One of the early methods proposed by Hatzivassiloglou and McKeown [4] focused only on classifying adjectives. For this reason, they used syntactic patterns to identify whether conjunctions of adjectives had the same or different orientation. The final step was a clustering algorithm that resulted in the formation of two groups, where the group with the highest frequency was labeled as "positive" under the assumption that positive adjectives are more frequents. Turney and Littman [3] were also based on this technique and created a Semantic Orientation Pointwise Mutual Information measure (SO-PMI) which used the IR measure of PMI adapted to measure correlation with a set of positive and negative opinion words. The latter is heavily based on opinion seed words. This need of lexica providing prior polarity have given rise to several resources like Senti-Wordnet (Esuli and Sebastiani [15]) and General Inquirer[1]. However, it has been suggested that these lexica do not reflect domain-specific characteristics and for this reason several methods for adapting sentiment lexica to different domains have been proposed (Choi and Cardie [16], Bol-

[1] www.wjh.harvard.edu/ inquirer

legala et al. [17]).

Moreover, research on algorithmic machinery has advanced, providing several frameworks for supervised algorithms. It was recently shown by Wang and Manning [18] that even simple machine learning algorithms like Naive Bayes lacking feature engineering techniques and using only simple unigram and bigram features can outperform other more sophisticated algorithms like CRF-based techniques. On the negative side, all these approaches require extensive training thus leading to the need of annotated corpora. Unfortunately, when moving to a different domain (e.g from movie to product reviews) the supervised algorithms require annotation on the new domain, which is really expensive to obtain. For this reason, domain adaptation techniques have been proposed (Titov [19], Blitzer et al. [20]) in order to leverage the annotated information and transfer it to other domains.

2.1.2 Joint learning of aspect and sentiment

Recently, Bayesian modeling of aspect and sentiment have gained interest and several methods have been implemented that either do this in complete unsupervised framework or by leveraging some kind of supervision. In our model, we restrict ourselves to using weak supervision in the form of global ratings of the reviews, that are provided by the user. We use no other information for guiding the aspect discovery. In some platforms[2], users have the possibility to provide ratings for a set of predefined aspects. This extra information can serve as additional supervision for Bayesian models; for example, Titov and McDonald [21] used the aspect ratings provided in some reviews, whereas Branavan et al. [22] leveraged pros/cons lists[3] which are provided as a complement to the review. However, since aspect ratings is not a standardized option in online reviews, it is the case that sometimes users do not provide any other information apart from the text of the review and the global rating. When moving to other domains like blog posts or news articles, it becomes clear that even supervision coming from the global ratings becomes problematic but is still easier to obtain than other, fine-grained types of information.

Mei et al. [23] extended LDA by creating a joint model of sentiment and aspect (TSM) for product reviews, which unlike traditional LDA, it is able to distinguish between aspects and opinion words. It samples a word either from the background component model or from topical themes, where the latter are further categorized into three sub-categories, i.e. neutral, positive and negative sentiment models. We have

[2]For example, `http://www.tripadvisor.com` or `http://www.amazon.com`.

[3]As for example in platform in `http://www.epinions.com/`.

to note here that although neutral words are generated from topic-specific language models, negative and positive language models are shared across all topics, and so there is no possibility for extracting topic-specific opinion words. Finally, for sentiment detection, their model requires post-processing to calculate the sentiment coverage of a document or a sentence.

On the contrary, Zhao et al. [24] created MaxEnt-LDA Hybrid which, unlike TSM, induces topic-specific opinion words. More specifically, their model generates a word than can either be a) a commonly used word (e.g. "know"), b) a word referring to a specific aspect (e.g "staff"), c) a word expressing an opinion specific to some aspect(e.g "friendly") or d) a general opinion word (e.g. "great"). However, unlike our work, they do not use the notion of polarity for clustering negative and positive opinion words, and thus cannot directly perform sentiment classification.

Finally, Jo and Oh [2] created a unification model of aspect and sentiment (ASUM), by extending a *sentence LDA*. In SLDA, unlike traditional LDA, the words of one sentence are constrained to be generated from the same topic-specific language model. In ASUM, the generative story proceeds as following; the author of the review decides on the distribution of sentiments of the review (e.g 70% positive and 30% negative). Then, he decides the distribution of the aspects for each sentiment, e.g. 50% about the staff, 25% about the rooms, and 25% about the price for the positive sentiment. Finally, for each sentence, he picks a sentiment to express and an aspect for which he has expressed this opinion. Although ASUM is the work mostly similar to ours, for distinguishing between negative and positive words they use sentiment lexicons to introduce informative prior of words. In our work, we model the sentiment of the review as an observed variable, and we use this information to help the model draw the distinction between positive, negative and neutral clusters of words.

2.2 Leveraging Content Structure for Sentiment Analysis

In the past years there has been an increase of interest in the research community for leveraging content structure in several natural language processing tasks. One of the arguably most popular views of content structure is discourse information, which has been injected in various tasks like paraphrase extraction tasks(Regneri and Wang [25], Pichotta and Mooney [26] in script learning and Meyer and Popescu-Belis [27] in statistical machine translation). Sentiment analysis field is not an exception to this trend and it has been suggested (Webber et al. [28]) that discourse information can actually improve the performance of several sentiment analysis tasks.

We review and classify the work done on this area into three fields; work that uses discourse information through sentiment polarity shifters, work that uses discourse relations as defined in the Rhetorical Structure Theory (RST), and finally work that views discourse information as local structure modeling.

2.2.1 Basics of Rhetorical Structure Theory

RST (Mann and Thompson [9]) was initially developed for text generation and aims at creating a framework for structural description of the meaning of a given text. In this theory, the main ingredient is the *elementary discourse unit* (EDU), which is usually a clause. The EDUs and higher-level discourse segments are linked via a predefined set of rhetorical relations, creating a tree-like hierarchical structure. Furthermore, each EDU can either act as a nucleus or satellite. Intuitively, the nuclei tend to provide basic information, whereas satellites provide additional information. As an example, given the sentence

1. To create your own "Victorian" bouquet of flowers,
2. choose varying shapes, sizes and forms, besides a variety of complementary colors.

in which an *Elaboration* relation holds, span 1 is considered as the nucleus since it holds the core information (i.e. the creation of the bouquet) whereas span 2 is the satellite since it elaborates by conveying additional information. Our discourse relations differ from RST relations because we only define pair-wise relations between EDUs. In our work, since this is done in a generative, unsupervised framework and discourse relations are mainly signaled by discourse cues, it is computationally very expensive and not straight-forward how to extend the markov relation between subsequent EDUs to the hierarchical representation of RST. Another critical difference is that our relations take into account changes in both sentiment and aspect. Since RST has no been built having in mind the sentiment domain, to the best of our knowledge, there is no direct way to encode in the relations of RST information for both aspect and sentiment.

2.2.2 Polarity Shifters

At the lexical level, Polanyi and Zaenen [29] proposed a scheme in which interactions between words determine their lexical valence (sentiment polarity). The prior valence of individual words is predefined and can either be expressed by a positive number (e.g. "excellent" has score 4) or negative (e.g. "annoyingly" has score -2). In their work

they define a deterministic framework for re-assigning the valence of words based on local context information, and specifically based on the properties of valence shifters like negatives (e.g. "not") and intensifiers (e.g. "very"). As an example, in the review

> The actors were **very** $good^{+}$. The play was **not** $boring^{-}$.

since **very** is an intensifier, it has the effect that it increases the absolute score of $good^{+}$. Similarly, **not** flips the valence of $boring^{-}$ from negative to positive. However, the presented method is limited to theoretical research and does not provide a systematic way for asserting the sentiment of a whole document or part of it.

Another work that adopts the use of polarity shifters was presented by Nakagawa et al. [30]. In their work, the polarity of individual words is compositionally combined through syntactic dependency graphs in the Conditional Random Fields (CRF) probabilistic framework, in order to determine the polarity of a span. More specifically, every dependency subtree in a sentence is associated with a sentiment polarity which is not observable in training data but is represented by a hidden variable. The polarity of a sentence can then be calculated in consideration of interactions between the hidden variables. For eliminating the explicit use of polarity shifters, Socher et al. [31] proposed a neural-network-based method for classifying opinion conveyed by sentences. Their input the sentence in the form of binary trees and the leafs, which are essentially the words of the sentence, are represented by distributed representations. These representations are meant to capture syntactic and semantic phenomena and so there is no need of explicit modeling linguistic information such as polarity shifters.

In a similar setting, Taboada et al. [32] adopt the locality introduced by Polanyi and Zaenen [29] and combine it with discourse information for asserting the score of a review. The discourse information is in the form of explicit discourse relations adhering to the Rhetorical Structure Theory, and are obtained by the statistical discourse parser SPADE (Soricut and Marcu [8]). SPADE labels relations that occur intra-sententially and marks the two spans that participate in the relation as *nucleus* and *satellite*. However, the authors do not explicitly use the information about the different relations, as one would expect. In contrast, they only use the information about spans being marked as nucleus or satellite and they only consider words in sentences marked as nuclei for aggregating the scores.

Our work differs substantially from this track of research; first, we do not work on the lexical level, since our method does not use any kind of prior knowledge concerning the polarity of individual words. Another fact is that SPADE operates on the sentence level, thus making it directly impossible to identify cross-sentential relations.

In our work, the sentence boundaries are not explicitly used, since we are operating on a finer-grained level than the sentence by going deeper into the structure of the sentence. Another point is that research that depends on the output of SPADE faces a natural error propagation; the performance of the method measured in F-1 score when trained on Wall Street Journal articles from the Penn Treebank was reported to be 49%. Finally, while our algorithm does not explicitly model RST relations, we can argue that we incorporate into our sentiment-aspect model some coarse-grain subset of the RST relations.

2.2.3 Discourse relations as RST Scheme

On a theoretical study, Asher et al. [33] has defined a shallow semantic representation for fine-grained contextual opinion analysis using discourse relations. In particular, they use the notion of feature structure on lexical semantic analysis and they use 5 discourse relations (CONTRAST, CORRECTION, SUPPORT, RESULT, CONTINUATION), which impose certain restrictions on the way these feature structures are combined to calculate the overall opinion expressed in a text on a given topic. The reported inter-annotator agreement for annotating with opinion categorization (i.e. whether an opinion expresses *Advice, Sentiment, Reporing* or *Judgement*) information and discourse information when computed in terms of Kappa[4] score is 95% for highly opinionated texts (movie reviews) and 73% for news articles. Although the annotation scheme presented seems consistent, the task of creating annotated corpora for the particular task is time-consuming and expensive; to the best of our knowledge, we are not aware of any annotated corpus of substantial size with sentiment and discourse information.

More recently, Zhou et al. [34] proposed a computational model for resolving intra-sentential polarity ambiguities without dealing with inter-sentential relations. The authors proposed a discourse scheme that is a subset of RST, consisting of 13 relations that are further grouped into five relations (CONTRAST, CONDITION, CONTINUATION, CAUSE, PURPOSE). Each of this group imposes constraints on the polarity of the two segments of sentences (i.e. the nucleus and the satellite). For example, CONTRAST relation indicates that the two segments should have opposite polarities, where the polarity of CONDITION, CAUSE and PURPOSE is defined by the *nucleus* segment. The authors do not use any labeled data for the discourse, but instead define patterns based on discourse cues that serve as signals for these groups and use these as seed to collect a large number of discourse instances. This is related to our method,

[4]Authors do not mention in the article which Kappa score is used

since we also model the discourse cues that signal transition from one segment to another. Once the discourse relations have been recognized in sentences and both nuclei and satellites have been identified, the task of polarity ambiguity elimination is then a deterministic process based on the constraints of the five relations. The way the constraints are applied to influence polarity differs from our proposed method, in which the sentiment, aspect and discourse information participate in a joint framework. Furthermore, we should note that the relations impose constraints only on the polarity, whereas in our work discourse classes influence also the aspect. Although we use weak supervision in terms of global rating, the authors use explicit polarity annotations of adjectives[5].

On the other hand, Snyder and Barzilay [35] do not explicitly model the discourse relation between sentences of segments. On the contrary, they globally model dependencies between polarity labels via an *agreement relation*. The agreement relation, which can be detected automatically, captures whether a user equally likes all the aspects of a review (i.e. if he has assigned the same polarity score to all of the aspects), or whether for some aspects of a review there are different degrees of satisfaction being expressed. This model then, which is based on contrastive RST relations, is coupled with a local aspect model to make an overall decision for sentiment classification.

2.2.4 Local structure modeling

In the work of Sauper et al. [5], the authors create a semi-supervised content model system for extracting key properties for a pre-specified set of aspects which incorporates a more abstract view of discourse. Their system implements document-level Hidden Markov Model which intents to capture transitions of the sentence-level latent variables.These latent variables encode aspect information in the context of multiple-aspect analysis. The emitted from the HMM words are used as the observed variables of Conditional Random Field model and the hidden variables are per-token annotations of the aspect following the IOB scheme. The output of this system was then used by Sauper et al. [36] for aspect-based sentiment analysis. Although the final system performs analysis on a fine-grained granularity, it cannot model full reviews, but rather snippets. Furthermore, in our system we model discourse structure on the sub-sentential level in the form of discourse classes for every segment, which in addition enables us to induce discourse structure that can be used for automatically generating coherent summaries of reviews. On the contrary, it is not clear to us how

[5]Although it is unclear whether these annotations come from some polarity lexicon or from explicit manual annotation

the structure induced in Sauper et al. [5] can be used in other tasks.

Another line of research, which is somehow closer to our definition of discourse information, defines discourse relations that are specific for the sentiment analysis task. In particular, Somasundaran et al. [37] propose the use of *opinion frames*, which consist of two opinions that are related by virtue of having unified or opposed targets. An opinion frame consists of three pieces of information; the opinion of the two targets and information of whether the two targets actually refer to the same entity (**same**) or not (**alternative**). The opinion has a polarity (**positive** or **negative**) and can either express a sentiment or an argument. In total, the proposed annotation scheme consists of $4 * 4 * 2 = 32$ opinion frames. As an example, in the sentence:

> The **remote** has a *nice shape*, but **it** *shouldn't have* so small buttons.

there is a *same* relation between the two entities in bold and the opinion expressed towards the first entity is *positive sentiment*, whereas the opinion expressed towards the second entity is *arguing against*. This work differs from ours in two main points. First, although we also define sentiment-aware discourse relation between targets, we do not include the opinion type *arguing*. Furthermore, for the initial experiments, we only define two different types of relations concerning sentiment (*same* and *alternative*) which results in total of four discourse classes. Second, since their modeling includes manual annotation, it becomes possible to define (annotate) relations between entities that are located in arbitrary positions in a text.

Somasundaran et al. [38] propose a framework for computational implementation of *opinion frames* for the sentiment classification task. In more detail, for predicting the polarity of mentions in a given instance, the authors propose both a supervised and an unsupervised method for using discourse information, both of which relying on a local classifier trained to predict the polarity of individual mentions and using simple unigram features as well as prior polarity lexica. The supervised method consists of first initializing the polarity by using the local classifier; then, in an iterative process, a relational classifier is used, which also takes into account features that are extracted from the opinion frames annotation. The unsupervised approach is implemented as a global optimization problem by using Integer Linear Programming (ILP). With ILP, the discourse relations are encoded as constraints on the polarity interpretation. In our generative modeling, the constraints are expressed by sampling the sentiment and the aspect of segments from a product-of-experts consisting of the document-specific aspect and sentiment distribution and from the discourse-specific aspecta and sentiment distribution.

2.3 Summary

In this Section we reviewed several methods dealing with aspect-based sentiment analysis. Initial methods approached the task by decomposing it in the subtasks of aspect discovery and polarity classification. Recently, Bayesian models have been proposed that deal with the problem in a joint framework. However, some of the proposed methods exhibit limitations on their expressivity power (e.g methods either model aspect-specific polar words without drawing the distinction between positive and negative, whereas others can express the notion of polarity but cannot deal with aspect-specific polar words).

Leveraging content structure in sentiment analysis tasks has started gaining interest with the results of discourse in sentiment analysis being way promising. This work that uses discourse to bootstrap sentiment has either tried to model inter- or intra-sentential relations based on a subset of RST, or has introduced relations specifically designed for the sentiment domain. Concerning discourse usage, both supervised and unsupervised methods that have been proposed aim at enforcing with different ways the constraints that are imposed by the relations.

However, to the best of our knowledge, the work presented in this thesis is the first one to combine information about sentiment, aspect and discourse in a sub-sentential level, in a unified, almost unsupervised framework. This model is further presented in Chapter 4 and before this, we provide some technical background in Chapter 3.

CHAPTER 3

Technical Background

3.1 Bayesian Statistics

In the following few paragraphs, we will motivate the use of Bayesian Inference for estimating the parameters of our generative model. According to our notation, X represents a series of observations (e.g the outcomes of flipping a coin), y represents the outcome of the next event in a chain of series and θ are the parameters (e.g. the probability P(heads) of producing *heards* when flipping a coin). Our goal then is to estimate $\boldsymbol{P(y|X)}$.

In Maximum Likelihood Estimation (MLE), $P(y|X)$ can be approximated through $P(y|\theta^{MLE})$ where θ^{MLE} is this value that maximizes the likelihood $P(X|y)$. It is clear that in MLE, we tailor our decision uniquely on what we have observed; if all the outcomes X of a coin are heads, then with probability 1 we are going to predict y to be heads. Maximum A Posteriori (MAP) treats this shortcoming by incorporating the prior beliefs for the parameters $P(\theta)$. Thus, it is easy to show that the MLE can be considered as MAP with uniform priors. So, MAP and MLE both consider θ as quantities whose values are fixed but unknown and both of them aim at getting the best estimate for them in order to arrive to the desired $P(y|X)$.

On the other hand, Bayes estimation treats the parameters θ as random variables. The goal is not anymore to obtain the best estimate of θ, thus throwing away information, but rather to account for all possible θ and calculate an expected value. Formally, this translates to integrating over all possible θ for obtaining $P(y|X)$ such that

$$P(y|X) = \int P(y|\theta)P(\theta|X)d\theta \tag{3.1.1}$$

where from Bayes rule we have:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta} \tag{3.1.2}$$

We can see from Equation 3.1.1 that the posterior $P(y|X)$ is defined by Bayesian Estimation using a true equality and not an approximation like MLE or MAP.

For solving Equation 3.1.1 and 3.1.2, we need to compute the integral to which analytical solutions might be impossible to obtain, since we have to sum over all possible combinations of solutions. Thus, there is a need of an approximate estimation of the posterior and this is accomplished by using a sampling inference algorithm (i.e. Gibbs Sampling (Geman and Geman [39])) from the family of algorithms known as *Markov Chain Monte Carlo* (MCMC).

3.2 Markov Chain Monte Carlo

Intuitively, MCMC implements the idea that given a distribution, it is simpler to sample from the conditional than to marginalize by integrating over a joint distribution. Thus, the goal of MCMC is to enable the sampling from a distribution that asymptotically follows the target distribution, without having to compute the latter. "Markov-chain" refers to the fact that we try to construct/approximate the target distribution by drawing many samples from it. In other words, the dimensions θ of the distribution are sampled alternately one at a time, conditioned **only** on the values of all other dimensions. Unlike other methods of approximating multi-dimensional integrals (e.g. Variational inference), MCMM are based on random walks ("Monte Carlo methods") or else random sampling.

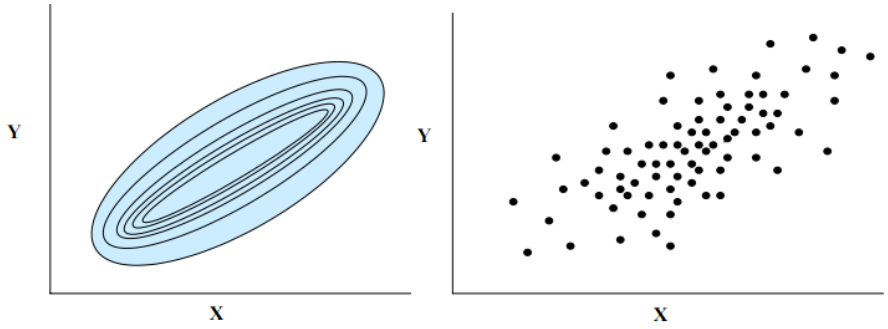

Figure 3.1: On the left, the real joint distribution of the variables X and Y. On the right, the approximation as obtained by sampling several times

As an example, imagine having two random variables X and Y. Their true joint distribution can be presented by a contour map as in Figure 3.1 on the left. In order to approximate this distribution, suppose that we know the 2 conditional distributions, e.g. given an X, we know the distribution of Y and the other way around. Then, we only need to randomly initialize X^0, and go into a loop where we first pick a value for Y^0 conditioned on X^0, then we pick a value X^1 given Y^0, and so on. Figure 3.1 on

the right presents the scattered plot of the resulting points, and we can see that this actually resembles the "real" distribution appearing on the left. Every point on the plot has been created by running the previously described iterative process for many times, and taking a sample from it.

In our inference we will use the "collapsed' variant of Gibbs sampling (Griffiths and Steyvers [40]). This means that we will integrate out some latent variables that are very hard to compute. For making the integrations tractable, conjugate priors on the posterior probabilities are preferred i.e. using prior distributions that can take the same form as the posterior.

3.3 Dirichlet Distribution as prior

In our Bayesian model, we encode our knowledge about distribution of latent variables in the model in the form of conjugate priors to these distributions. The conjugate prior to a categorical distribution is a Dirichlet distribution meaning that the resulting posterior will belong to the same family as the prior. Intuitively, in such a case, starting from what we know about some parameter prior to observing any data point and following a Dirichlet distribution, we then can update our knowledge based on the data points that follow a categorical and end up with a new distribution of the same form as the old one. This means that we can successively update our knowledge of a parameter by incorporating new observations one at a time, without running into mathematical difficulties.

More formally, a Dirichlet distribution is defined as a distribution over the K-dimensional probability simplex, which is simply a set of vectors

$$\Delta_K = \{(\pi_1, ..., \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\} \tag{3.3.1}$$

such that each entry is positive. In other words, we can consider every vector as a discrete distribution over K outcomes where π_k being the probability of outcome k and the density of the vector is

$$P(\pi_1, ..., \pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k^K \pi_k^{\alpha_k - 1} \tag{3.3.2}$$

where α_k are the parameters of the distribution or, in other words, the prior observation counts for outcomes governed by π_k. Intuitively, we can see that α_k simulates a smoothing process. For $\alpha_k < 1$, most of the probabilities returned will be close to 0, and the vast majority of the mass will be concentrated in a few of the probabilities. On the other hand, setting $\alpha_k > 1$ will result to dense, evenly distributed distributions.

3.4 Dirichlet Process: Non-parametric Bayesian prior

In our setting, since we want to induce discourse cues and we have no way of determining beforehand the different types which might be needed to account for the data, we use a non-parametric Bayesian prior, specifically the Chinese Restaurant process, which is a different perspective of the widely known Dirichlet Process (DP). The term non-parametric describes the family of methods in which the models have the ability adapt its complexity to the data. Form an intuitive example, consider the problem of clustering data. The traditional mixture modeling approaches require the number of clusters to be defined before analyzing the data. On the contrary, the Bayesian non-parametric approach estimates the number of clusters needed to model the observed data and furthermore allows for future data to exhibit previously unseen clusters.

Dirichlet Process DP (Ferguson [41]) belongs to the family of non-parametric methods, which have the property that they can model any arbitrary probability distribution as the size of the data goes to infinity. More formally, DP is a distribution over probability measures, and we can define the latter as a function from subsets of a space Θ to $[0, 1]$.

Then, we denote $G \sim DP(\eta, G_o)$ if G is a DP-distributed random variable measure having the property that for any finite set of partitions $A_1 \cup .... \cup A_N = \Theta$ as the vector $(G(A_1), ..., G(A_N))$ is Dirichlet-distributed.

The DP has two parameters: η which is is the *concentration parameter* and can been seen as the inverse-variance of the DP, whereas G_o is the base distribution and it can been seen as the mean of the DP. It is thus obvious that G should have the same support as G_o.

Chinese restaurant process paradigm Focusing on the draws from a Dirichlet Process, we can show that these draws take discrete numbers and that they define a partition or clustering of a set of objects (Teh [42]). This representation of a DP is the CRP, and we can intuitively understand this process as equivalent to assigning incoming customers to tables.

More precisely, let's assume there is a restaurant with infinitely many tables. The first customer that enters the restaurant picks the first table, and each of the following customers N has two possibilities; they either pick an already occupied table k with probability $\frac{N_k}{a+N-1}$, where N_k denotes the number of customers seating at the table k, or alternatively a new table $K + 1$ with probability $\frac{a}{a+N-1}$. Thus, after N customers sit

down, the seating plan gives a partition of N items.

The following process depicts the clustering property that CRP exhibits if we draw the analogy between tables and clusters, as well as between customers and integers. Furthermore, it is important to mention that even though we treat the incoming costumers in a sequential order, this sequence is *exchangeable*, meaning that the probability of seating arrangement does not depend on the ordering. Thus, we can treat every customer as the final one, which enables our inference to be tractable.

More formally, if we assume that there are K occupied tables in the current sitting arrangement, we can estimate the probability of a particular sequence of labeled table assignments $\boldsymbol{z}$ for N costumers by making use of the chain rule. The probability of the first customer sitting at a new table k will be 1, for the second customer on the same table $\frac{a}{(1+1)-1+a}$and so on, and this will be the case for all customers in the K costumers.

$$
\begin{aligned}
P(\boldsymbol{z}|\alpha) &= 1\prod_{i=2}^{N} P(z_i|z_{i-1},\alpha) && (3.4.1)\\
&= (\prod_{i=2}^{N}\frac{1}{i-1+\alpha})\alpha^K\prod_{k=1}^{K}(n_k-1)! && (3.4.2)\\
&= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\alpha^K\prod_{k=1}^{K}\Gamma(n_k) && (3.4.3)
\end{aligned}
$$

where z_{i-1} denotes the seating arrangement of the previous $i-1$ costumers, n_k denotes the number of customers seating at the table k. To derive the result, we have made use of the properties of the Gamma function $\Gamma(x) = x-1!$ and $\Gamma(x) = \frac{\Gamma(x+m)}{(x+m-1)(x+m-2)...(x+1)x}$.

We can now extend this paradigm in order to place the labels $\boldsymbol{l} = \{l_1, l_2, ..., l_K\}$ on the tables. The label l_k of the table represents a lexical item and is generated by the base distribution G_0 when the first costumer picks an empty table. In the case of Goldwater et al. [43], the base distribution is a distribution over phonemes. According to Goldwater et al. [43], each customer represents a work token so that the number of the customers sitting at a table with a particular label l_k encodes the frequency of this lexical item. This model can be seen as a two-stage CRP, where G_0 generate labels and CRP process generates frequencies. Under this model, we can define the probability of an entire sequence of lexical items $\boldsymbol{w}$.

$$P(\boldsymbol{w}|G_0,\alpha) = \sum_{\boldsymbol{z},\boldsymbol{l}} P(\boldsymbol{w},\boldsymbol{z},\boldsymbol{l}|G_0,\alpha)d\boldsymbol{z},\boldsymbol{l} \tag{3.4.4}$$

$$= \sum_{\boldsymbol{z},\boldsymbol{l}} P(\boldsymbol{z}|\alpha)G_0(\boldsymbol{l}) \tag{3.4.5}$$

$$= \sum_{\boldsymbol{z},\boldsymbol{l}} \underbrace{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\alpha^{K(\boldsymbol{z})}\prod_{k=1}^{K(\boldsymbol{z})}\Gamma(n_k^{(\boldsymbol{z})})}_{\text{P(Partition)}}\underbrace{\prod_{k=1}^{K(\boldsymbol{z})}G_o(l_k)}_{\text{P(draws)}} \tag{3.4.6}$$

where $K(\boldsymbol{z})$ is the number of tables occupied and $n_k^{(\boldsymbol{z})}$ is the number of customers sitting at the table k under the sitting arrangement $\boldsymbol{z}$. Although passing from 3.4.5 to 3.4.6 might initially seem unnatural, we have to note that every customer is assigned the lexical item of the table he is sitting such that $w_i = l_{z_i}$. Finally, as shown in Goldwater et al. [43], the *TwoStage-CRP*$(CRP(\alpha), G_0)$ is equivalent to a $DP(\alpha, G_0)$ and throughout our inference part, we will use the DP notation.

CHAPTER 4

A Bayesian Formulation of the joint induction of sentiment , aspect and discourse information

4.1 Problem Formulation

The task for joint sentiment, aspect and discourse class induction can be formalized with the following way, in which as input we take a corpus $\{d_1..., d_n\}$ of reviews where each review is associated with the global rating $\mathbf{r} \in [1, 5]$ and a specification of the number of topics K. Furthermore, each review consists of sentences, each of which is then linearly segmented by a pipeline architecture into smaller,non-overlapping segments, the Elementary Discourse Units (EDUs) segments, a task that we will describe and motivate in more detail in Section 5.2. As a result of this segmentation, each review d is represented as an ordered sequence of L_d segments $(s_{d,1}, s_{d,2}, ..s_{d,L_d})$.

As output, we predict for every segment s of every document d a *topic assignment* $z_{d,s} \in [1, K]$, a *sentiment assignment* $y_{d,s} \in [1, M]$, a *discourse class assignment* $c_{d,s} \in [1, C]$ as well as a boundary $b_{d,s} \in [0, len(s)]$[1] indicating the size of the discourse cue. For example, given the segment

Example 2. *but the rooms were comfortable.*

if $b_{d,s} = 0$ then this segment's discourse cue is the empty string, whereas if $b_{d,s} = 1$, the discourse cue consists of the first word of the segment, that is *but*. At this point we have to mention that since a boundary in a segment maps to a discourse cue for this segment, we can replace the problem of boundary identification to *discourse cue*

[1] where $len(s)$ indicates the size of the segment s

Table 4.1: Notation for the plain sentiment-aspect-discourse

Symbol	Description
C, D, M, K	number of discourse classes, documents, sentiments, topics
Hidden Variables to induce	
y^{gl}	overall sentiment of document d
$z_{d,s}$	aspect of segment (d, s)
$y_{d,s}$	sentiment of segment (d, s)
$c_{d,s}$	discourse class of segment (d, s)
$w_{d,s}^{cues}$	discourse cue of segment (d, s)
Prior distributions	
ϕ	distribution over discourse classes
$\phi_{k,m}^{subj}$	language model of subjective words of topic k in sentiment m
ϕ_{c}^{di}	language model of discourse cues of discourse class c
ϕ	distribution over discourse classes
θ_d	distribution of document d over topics
$\psi_{y^{gl},k}$	distribution of topic k over sentiments when global sentiment is y^{gl}
Fixed distributions	
$\psi_{c,m}^{di}$	fixed distribution of sentiments of discourse class c when sentiment of previous sentence is m
$\theta_{c,k}^{di}$	fixed distribution over topics for discourse class c when topic of previous sentence is k
Hyperpriors	
γ	vector of non-symmetric priors for per-topic sentiment distribution to favor sentiment same as overall sentiment m, $\gamma \in [0, \infty]^2$
α	symmetric prior for per-document topic distribution, $\alpha \in R$
λ_k	vector of non-symmetric priors for word distribution in topic k, $\lambda_k \in [0, \infty]^W$
δ	vector of non-symmetric priors for distribution of discourse classes in collection, $\delta \in [0, \infty]^C$
η	concentration parameter of the DP

Discourse Class	Description
NoClass	Not signaled class. Favors keeping the **same sentiment and aspect**
altSame	Signaled class. Favors transition to segment with **different sentiment** but **same aspect**
sameAlt	Signaled class. Favors transition to segment with **same sentiment** but **different aspect**
altAlt	Signaled class. Favors transition to segment with **different sentiment and aspect**

Table 4.2: Discourse classes modeled by our method.

identification.

4.1.1 Definition of Discourse class

One of the advantages of our model is that we adapt the notion of "discourse" for the specific sentiment task. We decide to not work with the RST relations since they are too fine-grained and do not necessary model relevant to the sentiment domain phenomena. For deciding what the intra-sentential discourse constraints should be, we inspected reviews from TripAdvisor and Amazon. Finally, we came up 4 classes that explain the local document structure of opinionated texts. These classes together with a description are given in Table 4.2.

A very important building block of our algorithm is the incorporation of the information we have about the discourse class and the way it affects the choice of sentiment and aspect for a segment. More precisely, we encode this prior knowledge that we have in **discourse-specific** sentiment ψ^{di} and aspect θ^{di} distributions. These distributions can be thought as encoding the *transition probabilities* of the Markov process and reflect the expected sentiment and aspect of a segment **given** the discourse class of the segment, as well as the sentiment and aspect of the previous segment.

In our model, we predefine $K \times C$ distributions on aspect and $M \times C$ distributions on sentiment. To give an example of such a distribution $\theta^{di}_{c,k}$, consider the discourse class *altSame* and the number of aspects being $K = 5$. As we described in Table 4.2, *altSame* favors the transition to the **same aspect** but **different sentiment**. If the aspect of the previous segment is #4, then we would like to place the majority of the probability mass in the same aspect. Thus, we define a distribution where $\theta^{di}_{altSame,\#4,l=\#4} = 0.8$. Therefore, the rest of the probability mass is equally distributed on the rest of the $K-1$ aspects, i.e. $\theta^{di}_{altSame,\#4,l} = \frac{1-0.9}{5-1} = 0.05$, where $l \neq \#4$. Finally, the distribution $\theta^{di}_{altSame,\#4}$ is $(0.05, 0.05, 0.05, 0.05, 0.80)$

In a similar way and by following the definition of *altSame* we work for constructing the discourse-specific sentiment distribution $\psi^{di}_{altSame,m}$ when $M = 3$, **given** the sentiment of the previous segment is $m = 3$.

At this point it worths mentioning that the exact **numerical form** of the discourse-specific distributions depend on the number of sentiments M and aspects K and is left for us to experiments with different configurations. However, what matters more is the intuition that is encoded in the **general form** of the probability distribution, e.g. in the case of *altSame*, the majority of the probability mass of the aspect distribution should be gathered in the same aspect as the previous segment.

4.2 Model overview

We propose a generative Bayesian model that explains how a corpus of D documents can be produced from a set of four latent variables, i.e. the sentiment y, the aspect z, the discourse cue c and the discourse phrase boundary b which is not explicitly modeled but is rather encoded in the induction of discourse cues. These latent variables are defined on a sub-sentential level, providing a modeling in fine-grained granularity.

Global Distributions At a global level, we first draw distributions over words $\phi^{subj}_{z,y}$ for every aspect and sentiment z and y respectively and ϕ^{di}_{c} for every discourse class c. Intuitively, the unigram $\phi^{subj}_{z,y}$ is meant to encode that every aspect is associated with a language model that expresses what words are used to express positive, negative and neutral opinion. For example, the language model of the aspect *service* indicates that the word *friendly* is used to express a positive opinion, whereas the word *rude* to express a negative opinion. Furthermore, these distributions are drawn from asymmetric Dirichlet priors. On the other hand, unlike the ϕ^{subj}, the distributions ϕ^{di} are drawn from a non-parametric prior and thus are not restricted to unigrams but to phrases of general size. These distributions capture the different discourse cues that are used to connect subsequent segments.

Next, we draw two distributions over sentiments. The $\psi_{m,z}$ is drawn from an asymmetric Dirichlet prior for every global sentiment m and aspect z and it encodes the information of general opinion about specific aspects. On the other hand, the $\psi^{di}_{c,m}$ are hand-coded and denote the sentiment transition probabilities from sentiment m under the discourse class c. Finally, distributions $\theta^{di}_{c,z}$ are also hand-coded distributions over aspects and they express the aspect transition probabilities under the discourse class c. For information concerning these two distributions, we refer the reader to 4.1.1.

Document level For each document d with L_d segments, we observe a global sentiment y^{gl}. This information is taken from the product reviews directly and is our only source of supervision. Then, we draw distributions θ_d over K aspects. This aspect distribution expresses the information of what aspects are discussed to review d and with what frequency.

Segment level For every segment s of the document d, a random variable $c_{d,s}$ is drawn. Intuitively, this variable indicates if the current segment connects somehow with the previous segment, and if so what kind of relation these segments exhibit (e.g they might both talk about the same aspect).

The next step is to draw the aspect $z_{d,s}$. Unlike the latent variable for the discourse, we want to generate the aspect not only according to some global structure, which in this particular case is the document-level structure expressed in the θ_d, but we want to enforce some soft constraints that stem from from the discourse structure encoded in the value of $c_{d,s}$. This hybrid nature of the process can be achieved by a technique known as *product-of-experts*(PoE) (Hinton [44]). Intuitively, it denotes a technique where we draw a variable from a product of several distributions, each of which act as a different source of information. This technique can also been seen as a log-linear-model as also discussed by Smith et al. [45]. By using the multiplication operation and not the additive, we succeed in making the negative effect of a probability more strong and influential. In the context of multilingual POS-tag induction, Snyder et al. [46] incorporated with the PoE in the generative model the intuition that if a tag is inappropriate for the monolingual setup, the same should apply for the multilingual setup as well. In our case, we want to generate the $z_{d,s}$ by relying on the beliefs of the discourse class expressed by the $\theta^{di}c, z_{d,s-1}$, and the topic information of the document expressed by the document-specific distribution θ_d. Once the aspect $z_{d,s}$ is drawn, we condition on this and the value of the sentiment $y_{d,s}$ is drawn in a similar process as the described above from the product of the two distributions $\psi_{y^{gl},z_{d,s}}$ and $\psi^{di}_{c,y_{d,s-1}}$

Generation of Discourse cues DP have been successfully used in the past for structure induction, like morphological induction where the goal is to induce word boundaries (Goldwater et al. [47]), as well as grammar induction (Cohn et al. [48]). In our model, we are using the *TwoStage-CRP*, which as Goldwater et al. [43] for discourse cue induction, which is a task analogous to word segmentation (Goldwater et al. [47]). In the latter, given a dataset of continuous phonemes, the task is to induce word boundaries. In this setting, the base distribution G_0 is expressed over phonemes. In our

setting, we can view the cue induction as general phrase segmentation; we are inducing the boundaries of phrases and this is achieved by having a base distribution G_0 over words. Thus, we define a process where the labels of the tables represent different discourse connectives sampled from the base distribution which is a bi-gram language model, and customers are simply the number of segments starting with this phrase. This non-parametric component allow us to induce phrases of various size. Finally, since we want to encode the intuition that the discourse classes are signaled by different cues, every class Thus, the generation of discourse cues $w_{d,s}^{cue}$ is done conditioned on the discourse class $c_{d,s}$.

Bag-of-words generation Once the sentiment and the aspect of the segment have been selected, we then condition on these in order to generate the words of the segment from the language model $\phi_{z_{d,s},y_{d,s}}^{subj}$. From this description it is clear that unlike the generation of the discourse cues which is more structured since the process can result to phrases of general size, the generation of the rest of the content of the segment does not adhere to any word order

4.3 Formal Generative process

Having provided the overview of the model as well as some theoretical background concerning the TwoStage-CRP, we can now proceed with the formal description of the generative process, whose plate diagram cab be found in Figure 4.1. Table 4.1 presents a description of the variables that are used. Shaded variables denote hyperpriors that are predefined, arrows denote conditional dependencies between variables, doubly-circled variables present observed data and boxes around variables denote that the variables inside the box are repeated as many times as the label of the box specifies. However, we have to note that this plate notation is a simplified version, which does not encode the conditional dependencies between the sentiments and aspects of subsequent segments.

Our algorithm takes us input a set of D documents, where each document d is an ordered sequence of L_d segments and each segment is represented as an observed bag-of-words. The number of topics K, sentiments M , discourse classes C as well as the prior discourse-specific sentiment ψ^{di} and aspect θ^{di} distributions are predefined. Furthermore, in our setting, the global sentiment y^{gl} of every review is observed. We have to note that this is the only supervision needed for our algorithm. Our algorithm then induces a set of latent variables, that can explain the generation process of our corpus. The following paragraph describes the generation process.

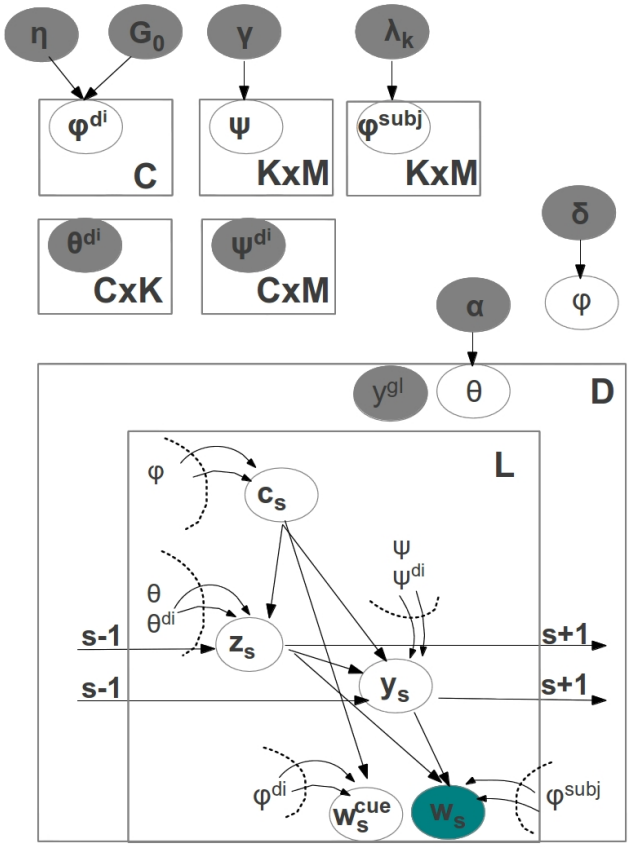

Figure 4.1: Plate notation for the generative model defined in Section 4.3. Table 4.1 describes the use of the variables that appear in plate diagram. Gray shaded variables represent parameters that are observed, blue shaded data that are observed and all the other are hidden variables. Arrow denote conditional dependencies.

1. For every aspect k and sentiment m draw unigram language models $\phi_{k,m}^{subj} \sim Dir(\lambda_k)$.
2. Draw discourse class distributions, $\phi \sim Dir(\delta)$.
3. For every classes c draw discourse cues distribution $\phi_c^{di} \sim \text{DP}(\eta, G_o)$
4. For every classes c and aspect k draw aspect transition distribution $\theta_{c,k}^{di}$
5. For every classes c and sentiment m draw sentiment transition distribution $\phi_{c,k}^{di}$
6. For every global sentiment m and topic z, draw sentiment distribution $\psi_{m,k} \sim \text{Dir}(\gamma)$.
7. For every document d with L_d segments and y_d^{gl} global sentiment:
   (a) Draw aspect distribution, $\theta_d \sim Dir(\alpha)$.
   (b) For every segment s in the review:
      i. Draw discourse class $c_{d,s} \sim \phi$
      ii. Draw discourse cue $w^{cue} \sim \phi_{c_{d,s}}^{di}$
      iii. Draw aspect $z_{d,s} \sim \theta_d * \theta_{c_{d,s}, z_{d,s-1}}^{di}$
      iv. Draw sentiment $y_{d,s} \sim \psi_{y_d^{gl}, z_{d,s}} * \psi_{c_{d,s}, y_{d,s-1}}^{di}$
      v. Draw rest of the words $\boldsymbol{w}_{d,s} \sim \phi_{z_{d,s}, y_{d,s}}^{subj}$

Encoding of prior knowledge in hyperpriors The hyperpriors play the role of prior knowledge encoded in pseudocounts. In our model, we use both symmetric and non-symmetric priors. Non-symmetric priors are the ones that have different values across different latent variables (i.e. they are represented as vectors) and symmetric are priors that have the same value and are represented as scalar values. Now, we give a description of our hyperpriors and their usefulness:

1. $\boldsymbol{\lambda_k}$ is a vector of size $|W|$ (i.e. the size of the vocabulary) and it is a non-symmetric prior, meaning that it assigns different pseudocounts to different words according to the topic k. As an example, we would like the word $staff$ to obtain a higher prior probability when it is considered for topic #1 and at the same time word $breakfast$ to have lower prior for topic #1 but higher for topic #3.

2. δ is a vector of size C and is again a non-symmetric prior and encodes our prior belief that the class *NoClass* is used more often than the other classes.

3. γ is our last non-symmetric prior and is vector of priors with size 2. It encodes our prior belief that it is more probable for the sentiment of an aspect in a review to be in accordance to the global sentiment y^{gl} and so it penalizes assignments of sentiment values to segments that are different than the global sentiment.

4. η is the concentration parameter of the DP. The smaller the value of the concentration parameter, the more sparsely distributed is the resulting distribution, with all but a few parameters having a probability near zero.

5. α is again a symmetric prior used to control sparsity in topic assignments.

4.4 Inference

The posterior distribution of our latent variables that we need to evaluate takes the form:

$$p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}} | \mathbf{w}) = \frac{p(\mathbf{w} | \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}}) p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}})}{\int_{\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}}} p(\mathbf{w}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}})} \tag{4.4.1}$$

.

In order to reduce the effective number of parameters in the model we will deploy collapsed Gibbs sampling and we will integrate out the distributions ϕ^{di}, θ, ϕ^{subj}, ψ and ϕ. The integration of these parameters are enabled by the use of conjugate priors. Since in our setting these parameters are categorical distributions (i.e. multinomial distributions with a unique trial) over latent variables, we use Dirichlet distribution as prior (for more details see Section 3.4).

In our case, we will *jointly* sample all the hidden variables for a segment, conditioned on all other values variables of the other segments obtained by the **previous iteration**. What is left then is to define the sampling step for our sampler; i.e. the conditional probability distribution of the hidden variables for a segment (d, s), given the assignments of latent variables in all the other segments. Mathematically, this can be obtained as:

$$Pr(z_{d,s}, y_{d,s}, c_{d,s}, w^{cues}_{d,s} | \boldsymbol{w}', \boldsymbol{z}', \boldsymbol{y}', \boldsymbol{c}', \boldsymbol{w}^{\boldsymbol{cues}'}) = \frac{Pr(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}})}{Pr(\boldsymbol{w}', \boldsymbol{z}', \boldsymbol{y}', \boldsymbol{c}', \boldsymbol{w}^{\boldsymbol{cues}'})} \tag{4.4.2}$$

where the superscript $'$ in the latent variables denote the vectors of the variables excluding the latent variable of segment s in document d. As we observe, the numerator is

just the joint probability of the model, whereas the denominator is the probability of the hidden variables of all the segments apart from the segment that we are sampling for. Thus, for obtaining the full conditional in Equation 4.4.2 that will allow us to simulate the posterior in Equation 4.4.1, we need to first evaluate the joint distribution.

Simplifying Joint probability According to our model, the joint distribution for the entire document collection is

$$Pr(\mathbf{w}, \mathbf{z}, \mathbf{y}, \mathbf{c}, \boldsymbol{w}^{\boldsymbol{cues}}, \theta, \psi, \phi, \phi^{subj}, \phi^{di}, \psi^{di}, \theta^{di}; \eta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \alpha, \boldsymbol{\lambda_k}) \tag{4.4.3}$$

The semicolon indicates that the values to its right are parameters for this joint distribution. Intuitively, this means that the variables to the left of the semicolon are conditioned on the hyperpriors given to the right of the semicolon. By following the model's generative story and the independence assumptions encoded in the plate notation depicted in Figure 4.1, the joint distribution can be decomposed into a product of several factors:

$$Pr(\phi|\delta)Pr(\phi^{subj}|\boldsymbol{\lambda})Pr(\theta|\alpha)Pr(\psi|\gamma)Pr(\mathbf{c}|\phi)Pr(z|\mathbf{c}, z_{-1}, \theta, \theta^{di}) \tag{4.4.4}$$

$$Pr(y|\mathbf{c}, y_{-1}, y^{gl}, \psi, \psi^{di})Pr(w|, y, z, \phi^{subj})Pr(w^{cues}|\phi^{di}, \boldsymbol{c})Pr(\phi^{di}|\eta, G_o) \tag{4.4.5}$$

The next step for arriving to the joint probability as in the nominator of Equation 4.4.2 is to integrate from the joint distribution the latent variables corresponding to the multinomials, i.e. $\theta, \psi, \phi^{di}, \phi^{subj}$. Every integrated parameter participates only in some parts of the joint distribution, and so we can break the integrals in several factors, according to the latent variables that they influence.

$$\int Pr(\mathbf{w}, \mathbf{z}, \mathbf{y}, \mathbf{c}, \boldsymbol{w}^{\boldsymbol{cues}}, \theta, \psi, \phi, \phi^{subj}, \phi^{di}, \psi^{di}, \theta^{di}; \eta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \lambda, \alpha)d\theta, \psi, \phi, \phi^{subj}, \phi^{di} \tag{4.4.6}$$

$$\int Pr(\mathbf{z}|\mathbf{c}, z_{-1}, \theta, \theta^{di})Pr(\theta|\alpha)d\theta * \tag{4.4.7}$$

$$\int Pr(\mathbf{y}|\mathbf{c}, y_{-1}, y^{gl}, \psi, \psi^{di})Pr(\psi|\boldsymbol{\gamma})d\psi * \tag{4.4.8}$$

$$\int Pr(\mathbf{c}|\phi)Pr(\phi|\boldsymbol{\delta})d\phi \tag{4.4.9}$$

$$\int Pr(\mathbf{w}|, y, z, \phi^{subj})Pr(\phi^{subj}|\boldsymbol{\lambda})d\phi^{subj} * \tag{4.4.10}$$

$$\int Pr(w^{cues}|\phi^{di}, \boldsymbol{c})Pr(\phi^{di}|\eta, G_o)d\phi^{di} \tag{4.4.11}$$

Aspect probabilities In order to compute the probability of the aspect assignment in the corpus, we will focus on Equation 4.4.7. There, we have to integrate out the multinomial θ. For the time being, we will ignore $\boldsymbol{c}$ and θ^{di}, since they are not affected by the integration of θ.

The general form of the probability mass of a categorical distribution has the following form:

$$Pr(\mathbf{x}; n, \pi) \quad = \quad \pi_1^{x_1} .... \pi_K^{x_K} \tag{4.4.12}$$

where K is the number of possible outcomes, n is the number of trials, x_i is the support of outcome i and π_i is the probability of outcome i. If we multiply the probability mass with the distribution of the Dirichlet prior, whichtakes a form similar to Equation 3.3.2, and then we **integrate** over the multinomial parameters π, we obtain a *Dirichlet compound multinomial* (DCM) distribution.

For the specific problem of topic assignment, we have:

$$\begin{aligned} Pr(\mathbf{z}|\alpha) &= \int Pr(\boldsymbol{z}|\theta) Pr(\theta|\alpha) d\theta && (4.4.13) \\ &= \prod_d^D Pr(\boldsymbol{z}|\theta_d) Pr(\theta|\alpha_d) d\theta_d && (4.4.14) \\ &= \prod_d^D DCM(N^{topics}; \alpha) && (4.4.15) \end{aligned}$$

where D denotes the number of documents and N^{topics} denotes the data structure that keeps track of the topic assignments of the **segments** in every document d. For a Dirichlet prior with parameters $\alpha = (\alpha_1, ..., \alpha_K)$, the DCM, having integrated out the prior for one document θ_d, assigns the following probability to a series of topic observations $z = (z_1, ..., z_N)$ in the document d with length N:

$$DCM(\mathbf{z}; \boldsymbol{\alpha}) \quad = \quad \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(N_{d,k}^{topics} + \alpha_k)}{\Gamma(\sum_{k=1}^K N_{d,k}^{topics} + \alpha_k)} \tag{4.4.16}$$

Finally, having plugged in categorical discourse-specific aspect distribution for all documents D and using a symmetric prior over the asepcts, the topic assignments can be obtained as:

$$\begin{aligned} Pr(\mathbf{z}|\alpha, \theta^{di}, \boldsymbol{c}, \boldsymbol{z^{-1}}) &= \prod_{d=1}^D \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{k=1}^K \Gamma(N_{d,k}^{topics} + \alpha)}{\Gamma(\sum_{k=1}^K N_{d,k}^{topics} + K\alpha)} * && (4.4.17) \\ & \prod_{c=1}^C \prod_{k=1}^K \prod_{k'=1}^K \theta_{c,k,k'}^{di \; N_{k,k'}^{diTop}} && (4.4.18) \end{aligned}$$

where N^{diTop} is a data structure that keeps track of the transitions of aspects. Intuitively, the above formula specifies that for obtaining the probability of some aspect assignment to all segments in our document, we will have to account for the number of times each topic has been used in our corpus. Furthermore, the last factor of the formula encodes the fact that the discourse class assignments also influence the result since they define a different discourse-specific aspect distribution.

Sentiment probabilities Moving on to Equation 4.4.8, with a similar work flow, we forget for a while about the discourse-specific sentiment distribution ψ^{di} and the discourse class assignments $\boldsymbol{c}$, and we integrate out the prior distribution $\psi_{o,k}$ for global sentiment o and topic k, thus obtaining the following DCM:

$$DCM(\mathbf{y};\gamma) = \frac{\Gamma(\sum_{m=1}^{M}\gamma_m)}{\prod_{m=1}^{M}\Gamma(\gamma_m)}\frac{\prod_{m=1}^{M}\Gamma(N_{o,k,m}^{sents}+\gamma_m)}{\Gamma(\sum_{m=1}^{M}N_{o,k,.}^{sents}+\gamma_m)} \tag{4.4.19}$$

Therefore, the probability of the sentiment assignments for all possible topics and global sentiments is:

$$Pr(\mathbf{y}|\gamma,\psi^{di},\boldsymbol{c},\boldsymbol{y^{-1}}) = \prod_{o=1}^{M}\prod_{k=1}^{K}\frac{\Gamma(\sum_{m=1}^{M}\gamma_m)}{\prod_{m=1}^{M}\Gamma(\gamma_m)}\frac{\prod_{m=1}^{M}\Gamma(N_{o,k,m}^{sents}+\gamma_m)}{\Gamma(\sum_{m=1}^{M}N_{o,k,.}^{sents}+\gamma_m)} \tag{4.4.20}$$

$$\prod_{c=1}^{C}\prod_{m=1}^{M}\prod_{m'=1}^{M}{\psi_{c,m,m'}^{di}}^{N_{m,m'}^{diSents}} \tag{4.4.21}$$

where M denotes the number of possible sentiments, N^{sents} denotes the data structure that keeps track of the sentiment assignments of **segments** for all possible combinations of global sentiments and topics and $N^{diSents}$ denotes the data structure that keeps track of the transitions of sentiments.

Again here, the intuition is that how probable a sentiment assignment is depends on how popular the specific assignment is in the dataset as well as on how well the discourse classes are distributed in the segments, since discourse classes also influence the sentiment of a segment through the prior belief that is encoded on distribution ψ^{di}

Discourse Class probabilities Similarly, if we integrate out the prior distributions ϕ, Equation 4.4.9 is simplified to:

$$Pr(\mathbf{c}|\boldsymbol{\delta}) = \frac{\Gamma(\sum_{c=1}^{C}\delta_c)}{\prod_{c=1}^{C}\Gamma(\delta_c)}\frac{\prod_{c=1}^{C}\Gamma(N_{c,.}^{cues}+\delta_c)}{\Gamma(\sum_{c=1}^{C}N_{c,.}^{cues}+\delta_c)} \tag{4.4.22}$$

where C denotes the number of discourse classes and N^{cues} is the discourse class specific distribution over discourse cues.

Language Model of Words Again, following the same patterns, we obtain the DCMs after integrating out all prior distributions $\phi_{k,m}^{subj}$ from Equation 4.4.10. Thus, we can obtain the probability of the words in our corpus:

$$Pr(\mathbf{w}|\boldsymbol{y},\boldsymbol{z},\lambda) = \prod_{k=1}^{K}\prod_{m=1}^{M}\frac{\Gamma(\sum_{w=1}^{W}\lambda_{k,w})}{\prod_{w=1}^{W}\Gamma(\lambda_{k,w})}\frac{\prod_{w=1}^{W}\Gamma(N_{k,m,w}^{subj}+\lambda_{k,w})}{\Gamma(N_{k,m,.}^{subj}+\sum_{w=1}^{W}\lambda_{k,w})} \tag{4.4.23}$$

where W is the size of the vocabulary and N^{subj} is the data structure that keeps the count of words being under a specific sentiment and aspect.

Language Model of Discourse Cues For the discourse cues, integrating out the distribution ϕ^{di} results to Equation 3.4.6 but without summing over all the possible partitions:

$$Pr(\boldsymbol{w}^{\boldsymbol{cues}}|\boldsymbol{c}, G_0, \eta) = \prod_{c=1}^{C} \frac{\Gamma(\eta)}{\Gamma(N_{c,\cdot}^{cues} + \eta)} \eta^{L_c} \prod_{i=1}^{L_c} \Gamma(N_{c,l_i}^{cues}) \prod_{i=1}^{L_c} G_0(l_i) \quad (4.4.24)$$

where L_c is the number of unique phrases used in discourse class c, N^{cues} is the discourse-class-specific distribution over discourse cues and thus, N_{c,l_i}^{cues} denotes the number of times cue l_i has been used seen in segments generated by the discourse class c. Finally, $N_{c,\cdot}^{cues}$ is the number of segments generated by the class c.

Joint sampling of aspect, sentiment, discourse class and discourse cue Having updated all the factors of the joint distribution depending on which integrated variable they depend, we can put everything together and calculate the updated joint distribution $Pr(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{w}^{\boldsymbol{cues}}; \alpha, \boldsymbol{\delta}, \boldsymbol{\lambda}, \gamma, \eta, G)$. From this joint, we can then derive the full conditional distribution for a segment index (d, s), i.e., the update equation from which the Gibbs sampler draws the hidden variables associated with segment that segment, i.e. z_s^d, y_s^d, c_s^d, $w_{d,s}^{cue}$. For doing so, we have to compute Equation 4.4.2. Through careful computations and cancellations of terms, we can derive the target conditional distribution which will serve as the update step of our sampler:

$$Pr(z_{d,s} = z, y_{d,s} = y, c_{d,s} = c, w_{d,s}^{cues} = w^{cues}|...) \propto \quad (4.4.25)$$

$$ASPECT * SENTIMENT * CUES * DISCOURCE * WORDS \quad (4.4.26)$$

where:

- $ASPECT = \frac{N_{d,z}^{topics} + \alpha}{N_{d,\cdot}^{topics} + K\alpha} * \theta_{c,z_{s-1},z}^{di}$ and is derived from Equation 4.4.18 and 4.4.18
- $SENTIMENT = \frac{N_{y_d^{gl},z,y}^{sents} + \gamma_{y==y^{gl}}^{y^{gl}}}{N_{y_d^{gl},y}^{sents} + (\gamma_0^{y^{gl}} + \gamma_0^{y^{gl}} + \gamma_1^{y^{gl}})} * \psi_{c,y_{s-1},y}^{di}$ and is derived from Equation 4.4.21 and 4.4.21
- $CUES = \frac{\eta G_0(w^{cues})}{\eta + N_{c,\cdot}^{cues}}$ **if** cue phrase has already been seen for the discourse class c, **else** $\frac{N_{c,w^{cues}}^{cues}}{\eta + N_{c,\cdot}^{cues}}$ and they are derived from Equation 4.4.24
- $DISCOURCE = N_c^{cues} + \delta_c$ and is derived from Equation 4.4.22
- $WORDS = \prod_{w=1}^{W} \frac{N_{z,y,w}^{subj,-s} + N_{z,y,w}^{subj,-w^{cues}} + \lambda_{k,w}}{N_{z,y,w}^{subj,-s} + \lambda_{k,w}} * \frac{N_{z,y}^{subj,-s} \sum_{w=1}^{W} \lambda_{k,w}}{N_{z,y}^{subj,-s} + N_{z,y}^{subj,-w^{cues}} + \sum_{w=1}^{W} \lambda_{k,w}}$ and is derived from Equation 4.4.23

We have to note that is an unnormalized probability. In order to normalize it, we have to compute all possible combinations of latent variables, sum them up and divide each $Pr(z_{d,s} = z, y_{d,s} = y, c_{d,s} = c, w^{cues}_{d,s} = w^{cues}|...)$. Then we can construct the distribution, and sample the tuple of the 4 hidden variables. Furthermore, we have to note that $-s$ in the structure N^{subj} denotes the counts without considering the counts of the segment, whereas $-w^{cues}$ denote the counts of the words in the segment s without considering the words that are part of the w^{cues} discourse cue. Finally, $G_0(w^{cues})$ indicates the probability of the discourse cue as obtained by the bi-gram language model.

4.5 Summary

In this section, we provided the problem formulation of our work. We defined a generative process which used three different latent variables to explain the data generation, i.e. the sentiment, aspect, discourse class. Furthermore, we defined the four discourse classes which are used in order to introduce a local structure on the subsentential level. The key point of our generative process in the induction of discourse connectives that are located in the beginning of each segment. The discourse cue induction is equivalent to phrase segmentation and is modeled by a TwoStage-CRP. Finally, since exact inference of our method is intractable, we derived step-by-step the Gibbs algorithm and computed the conditional distribution of our latent variables which serves as the update equation of the sampler.

CHAPTER 5

Experiments

To the best of our knowledge, this is the only work that aims at evaluating directly the joint information of the sentiment and aspect assignment at the sub-sentential level of full reviews, since most of the existing work focuses on indirectly evaluating the produced models by classifying the overall sentiment of sentences (Titov and McDonald [21], Brody and Elhadad [13]) or even reviews (Nakagawa et al. [30], Jo and Oh [2]). Furthermore, since the contribution of this work is the induction of discourse information relevant for the specific task in a joint framework, we believe that the evaluation on such a high level would not reveal the true performance of this new feature. However, even in the work by Sauper et al. [36] where the analysis is on a finer-grained granularity, their model considers individual snippets out of document context and so their evaluation follows a different hypothesis than ours.

5.1 Data

For our experiments, we had to create our own dataset from scratch, since to the best of our knowledge, the particular task of sub-sentential sentiment analysis on product reviews has not been tackled before. For building our dataset, we crawled hotel reviews from TripAdvisor[1]. This decision was mainly empirically driven by the fact that in a qualitative examination we did, Tripadvisor reviews tended to be better structured (i.e. longer sentences, use of explicit discourse connectives etc) than reviews from other websites, e.g. Amazon.

The crawler that we built only fetched reviews written in English, disregarding any other noise/non-English reviews; for minimizing the amount of those reviews, we mainly fetched reviews from hotels located in the United States and London. On the

[1] www.tripadvisor.com

	Neutral	Positive	Negative
#Reviews	2394	7660	3505
#Segments	322935		
#Words	35982		

Table 5.1: Statistics characterizing our dataset

other hand, we did not discard non-native reviews or reviews that were not signed as helpful. However, we did filter out very short reviews (i.e. reviews with less than 6 sentences) since we do not expect very short reviews to exhibit a helpful structure. Moreover, we tried not to have too skewed data regarding the global rating; since this kind of supervision is very cheap and there are plenty of data, this assumption is not too severe for our model. Finally, we ended up with 13559 reviews, written for 20 different hotels. Table 5.1 summarizes the statistics characterizing our dataset.

5.2 Preprocessing

For the preprocessing of the reviews, we followed standard text normalization techniques concerning lowercasing and tokenization using the OpenNLP toolkit[2]. We have to note that, although it is common in topic modeling literature to remove stop words in order to create topics with less noise, this is not possible in our setting. This is due to the fact that we expect stop words to be part of discourse phrases (e.g. "and", "my only complaint").

In our work, segmenting a sentence into smaller clauses is of vital importance. As we mentioned in Section 1.2, a rather critical assumption of our model is that the first few words of a sentence indicate a certain predisposition concerning the sentiment and aspect of a sentence. However, it is a usual case that a sentence is not homogeneous with respect to the sentiment and aspect, and especially longer sentences tend to exhibit a more elaborated structure. In these cases, the flow of information is enabled by the use of discourse connectives in the intra-sentential level. In order to cope with these cases and bring these connectives to the beginning of every fragment, we apply a method for linearly segmenting them into smaller, non-overlapping segments. This process is also known as "Discourse segmentation", where the goal is to decompose discourse into elementary discourse units (EDUs). Discourse segmentation is usually seen as a preprocessing step to discourse parsing, where the goal is to identify and name relations that exist between EDUs.

[2] `http://opennlp.sourceforge.net/projects.html`

For segmenting the reviews, we experimented with two pieces of software, SPADE[3] and SLSEG[4]. For creating the input for both softwares, we applied POS-tagging with the Stanford Tokenizer[5] and syntactic parsing with Charniak's parser[6].

SPADE is based on a supervised framework and learns to introduce segment boundaries based on lexicalized grammar rules. On the other hand, SLSEG is a rule-based tool that has been developed in order to treat some of the disadvantages of the former. More specifically, it tries to treat some problems including having too small EDUs, having EDUs that are not very informative (e.g. "He said that" or "I think that"), or having EDUs with no verb at all. The reported F-score for the segmentation task in online reviews was found to be 79% (Tofiloski et al. [49]).

5.3 Manual Annotation

For evaluating our method, we compiled a gold standard based on a subset of our dataset (i.e. 65 reviews). We asked from nine annotators to help us create the gold standard.Eight of them were assigned 1/8 of the reviews each and the 9th annotator annotated all the reviews, in order to allow computation of the inter-annotator agreement. In cases of inter-annotator disagreement, the annotations of the 9th annotator were accepted as the gold standard. All annotators were presented with the whole review partitioned in EDU's from the preprocessing step and were asked to annotate every segment with the **aspect** and **sentiment** it expresses.

Table 5.2 presents the labels together with the distribution at the final version of the gold standard. The labels above the horizontal line appear as possible[7] rateable aspects in TripAdvisor. The label "rest" was inserted in order to capture cases where segments referred to aspects that are very rarely discussed and can serve as the "garbage collector" class for segments that contain personal stories or whatever is not directly relevant for a review. Furthermore, we introduced the aspect "recommendation" that tends to capture opinion expressed for the hotel as a whole (e.g. *We could highly recommend the hotel* or *Avoid staying in the hotel*).

Finally, 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

[3] `http://www.isi.edu/licensed-sw/spade/`

[4] `http://www.sfu.ca/~mtaboada/research/SLSeg.html`

[5] `http://nlp.stanford.edu/software/tagger.shtml`

[6] `ftp://ftp.cs.brown.edu/pub/nlparser/parser05Aug16.tar.gz`

[7] When submitting a review to TripAdvisor there is the possibility of rating these aspects but is not mandatory

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

Table 5.2: Statistics characterizing our annotated data

the joint task of sentiment and aspect annotation. Table 5.3 and 5.4 present the contingency tables for the aspect and sentiment labeling respectively. On a sentence-level annotation, Ganu et al. [50] report IAA between 54% and 80% for separate aspects and sentiments.

This numerical difference between our IAA and the one reported in previous work indicates that the sentiment-analysis task on a finer-grained level than the one of sentence level is very challenging and can be considered as a more demanding task than the one of multi-label or even binary sentiment classification of reviews. We suspect that one of the reasons for this relatively low IAA is the discourse-segmentation step on the preprocessing task. Our annotators commented that sometimes the segments were very short, containing no information whatsoever about the actual task of sentiment analysis. This is also partially explained by the fact that the class with the majority of disagreements is the class "rest". This serves as a motivation for perceiving discourse analysis jointly with the sentiment analysis task in contrast to our current pipeline architecture.

5.4 Experimental Setup

For our experiments, we let our sampler run for 2000 iterations and we labeled our dataset with the last assignment of latent variables on the segments. The hyperpriors were set empirically by manually inspecting the languages models produced when running our model on a subset of our dataset. The final values are:

	value	service	food	recom	rest	amen	rooms	clean	sleep	loc
value	**40**	2	1	2	7	1	2	0	0	0
service	2	**168**	1	1	54	6	10	4	0	0
food	0	2	**61**	2	10	6	0	0	0	0
recom	5	1	0	**66**	45	0	1	0	1	2
rest	3	14	3	11	**243**	2	11	4	0	15
amen	1	3	8	2	39	**100**	19	6	0	2
rooms	0	3	0	2	33	15	**236**	14	12	1
clean	0	0	0	0	3	0	7	**49**	0	0
sleep	0	0	0	3	9	0	6	0	**37**	1
loc	0	2	2	4	11	3	0	0	0	**99**

Table 5.3: Contingency table for aspect discovery. Elements on the diagonal represent agreement between the annotators

	positive	negative	neutral
positive	**414**	9	99
negative	5	**386**	108
neutral	38	45	**437**

Table 5.4: Contingency table for aspect discovery. Elements on the diagonal represent agreement between the annotators

- $\alpha = 10^{-3}$
- $\gamma^{m}_{m'} = 5 * 10^{-4}$ for $m \neq m^{'}$ and $\gamma^{m}_{m'} = 10^{-3}$ otherwise
- $\eta = 10^{-3}$
- $\delta^{c} = 10^{3}$ for $c = NoClass$ an $\delta^{c} = 10^{-4}$ for all the other classes

In order to set for every word w in the vocabulary an informative prior $\lambda_{k,w}$ for every topic k, we modified our model in order to have only one latent variable, i.e. the aspect. This model is thus agnostic to sentiment and any other information related to discourse. After running this model for 2000 iterations, we set $\lambda_{k,w} = P(w|z_k) * coef$. The probability of the word can be obtained by estimating the distribution of the language model $\phi^{subj}_{k,w} = \frac{N^{subj}_{k,w}+\lambda}{\sum_{w=1}^{W} N^{subj}_{k,.}+W\lambda}$. $coef$ is set to 10^4 for our experiments[8]. The discourse-specific and sentiment-specific distributions are set manually. Furthermore, the base distribution generates phrases from a bi-gram language model. Finally, we constraint the size of discourse cues to by up to 3 words, in order to limit the search space and enable our sampler to converge easier.

[8]The final priors range in the scale of 10^{-4}

5.5 Baseline

For creating our baseline *SentAsp*, we ran our model with the discourse module disabled. This means that in the generative process, as defined in Section 4.3, the generation of segments is modified in the following way, whereas all the other parts are left the same:

1. Draw the aspect based on the document-specific topic distribution, $z_{d,s} \sim \theta_d$
2. Draw the sentiment based on the document-specific topic distribution, $y_{d,s} \sim \psi_{z_{d,s}}$
3. Draw **all** words from the topic- and sentiment-specific language model, $\boldsymbol{w}_{d,s} \sim \phi^{subj}_{z_{d,s},y_{d,s}}$

We should note here that this model shares commonalities with the one presented by Jo and Oh [2]. Observing the latent variables, we see that there is no longer the discourse class latent variable as well as the discourse cue latent variable. Furthermore, there is no Markov assumption between subsequent segments. Thus, this discourse-agnostic model is less restrictive than ours and we expect to be easier to learn.

We have to note that the *SentAsp* was run with the same configurations as the discourse model (i.e. for the same number of iterations and with the same hyperpriors for every different number of topics K).

5.6 Evaluation metrics

Measuring the effects of *Discourse* model against *SentAsp* is not a trivial process. This is due to the fact that the two models infer abritrary topics (i.e. the latent topic #1 in our model does not necessarily correspond to the latent topic #1 in *SentAsp*), and there is no way, that we know of, to assign the same labels or to co-ordinate the inferred topics, since the sampling involves a random process. One way of evaluation would be to manually assign to gold classes the latent topics. However, this is not so straightforward, since the assignment for both systems should be equally fair and this is rather impossible to achieve, even if it was done by the same annotator.

Ideally, we would like to evaluate the models in an extrinsic way, i.e. by plugging both models in an application and more specifically an application of the sentiment summarization task, and compare the two models in terms of the effects the models have on the performance of the system. However, mainly due to time limitations, this was not feasible.

Thus, we draw ideas from other tasks which, similar to ours, that result in induction

of clusters. More precisely, our evaluation is inspired from tasks related to word sense induction (Agirre and Soroa [51]) and semantic role labeling (Lang and Lapata [52]). The organizers of the Semeval-2007 task "Word sense induction and discrimination" presented two frameworks for evaluating the induced clusters of words. Those frameworks include an unsupervised approach (5.6.1) which aims at evaluating the clustering property of the models, as well as a supervised one (5.6.2), where the clustering is converted to a classification problem and then standard IR metrics are applied.

5.6.1 Unsupervised Evaluation

Given a set of segment assignments to gold classes $\{S_i\}_1^K$ (i.e. where a class represents a pair of aspect-sentiment) and set of segment assignments to clusters $\{C_j\}_1^A$ (i.e. where a cluster represents a pair of latent topic-sentiment), we define three measures: *purity*, *entropy* and *F-score*.

Purity is a measure of the degree to which the predicted clusters meet the goal of containing only instances with the same gold class. Therefore, to compute purity, each cluster is assigned to the class which is most frequent in this cluster

$$Purity(C_i) = \frac{1}{|C_i|} \max_j |C_i \cap S_j| \tag{5.6.1}$$

and by summing up all the purity values for the A clusters, we can compute the value for the entire clustering solution.

$$Purity(\{C_i\}_1^A) = \sum_j \frac{|C_i|}{|D|} Purity(C_j) \tag{5.6.2}$$

However, one thing to comment here is that if each document is set to its own cluster, then purity becomes 1, which is the maximum value.

Entropy measures how the various classes are distributed within each cluster. The entropy for a cluster C_i is defined as:

$$Entropy(C_i) = -\frac{1}{logA} \sum_j \frac{|C_i \cap S_j|}{|C_i|} log(\frac{|C_i \cap S_j|}{|C_i|}) \tag{5.6.3}$$

and for the entire clustering solution as:

$$Entropy(\{C_i\}_1^A) = \sum_j \frac{1}{A} Entropy(C_j) \tag{5.6.4}$$

From the definition of the entropy we note that the lower the value is, the better the clustering solution is.

F-score We also evaluate in terms of F-Score, but in a manner adapted to match the clustering problem. More precisely, if we consider the segments of a cluster as the "retrieved" examples, we can define precision and recall for a pair of a cluster/class as the fraction of correctly "retrieved" examples normalized by the cluster size and the fraction of correctly "retrieved" normalized by the class size respectively. Then, the F-score for this pair is defined as:

$$F - Score(C_i, S_j) = \frac{2 * prec(C_i, S_j) * rec(C_i, S_j)}{prec(C_i, S_j) + rec(C_i, S_j)}, \tag{5.6.5}$$

where *prec* is the precision value and *rec* the recall. The F-Score of a class S_i is the maximum F value scored at any cluster C_i

$$F - Score(S_i) = \max_{C_j} F - score(S_i, C_j) \tag{5.6.6}$$

and the final F-Score for the clustering solution is calculating by summing up the individual F values for all the classes

$$F - Score(\{S_i\}_1^K) = \sum_i \frac{|S_i|}{|D|} F - score(S_i) \tag{5.6.7}$$

5.6.2 Supervised Evaluation

A more realistic setup for evaluation, which is similar to the procedure we would follow if we incorporated our model to a sentiment summarization platform, is the one where we attempt to match the induced clusters to the gold classes. This is achieved by splitting the gold standard into two subsets. We use the training portion to learn a $1-1$ mapping from the gold classes to the induced clusters. For our purposes, we restrict the mapping in such a way so that one gold class is mapped to only one induced cluster. Finally, we apply this learnt mapping to the testing portion and we evaluate with the metrics found in the IR literature.

The problem of finding such a mapping is reduced to the problem of finding a maximum weight matching in the bipartite graph $G = ((X, Y), E)$ where X represents the induced clusters $\{C_i\}_1^A$ and Y the gold classes $\{S_j\}_1^K$. The weights $w(i, j)$ of the edges $e_{i,j}$, where $i \in [1, A]$ and $j \in [1, K]$, are defined as the fraction $\frac{|C_i \cap S_j|}{|C_i|}$ and intuitively describes the precision of mapping the i cluster to the j class.

5.7 Summary

In this Section, we provided the description of the experimental setup. We manually annotated a subset of the crawled reviews yielding an IAA close to 63% when measured

in terms of Cohen's Kappa score. We created a baseline *SentAsp* which is a discourse-agnostic variant of our model. Evaluating those two algorithms is not a trivial task. We were inspired by the evaluation performed in other induction tasks like word sense induction. We thus define two frameworks. The first is an unsupervised framework which does not aim at matching the induced clusters with the classes of the gold standard, but instead uses metrics from the clustering literature. On the other hand, in the supervised evaluation we split the gold-standard to training and testing portion, we use the training to induce 1-1 mappings between clusters and gold classes and we apply these mappings to the testing portion and evaluate with the standard IR metrics.

CHAPTER 6

Results and Analysis

6.1 Quantitative analysis

6.1.1 Unsupervised evaluation

As we have already described in Section 5.6.1, we conduct an intrinsic evaluation of the *Discourse* model against the discourse-agnostic model *SentAps* in terms of *purity*,*entropy* and *F-score*. Furthermore, we include the baseline *1clSegm* which assigns every segment in a single cluster. Table 6.1 presents the results of our experiments.

Unfortunately, this evaluation is not very helpful for drawing many conclusions. In terms of F-Score, which is more "balanced" than the other two metrics since it accounts for both clusters being pure and classes not being distributed in several clusters, the results are mixed. Even though for the realistic setup, where we set the number of clusters to be equal to the number of classes, the *Discourse* model performs better than the *SentAsp* model, in all other configurations it is either the case that the two models

Topics	Model	Purity	Entropy	F-score
-	1clSegm	1	0	0.034
5	*SentAsp*	0.241	0.330	**0.239**
	Discourse	0.210	0.296	0.206
10	*SentAsp*	0.197	0.288	0.176
	Discourse	0.205	0.296	**0.194**
20	*SentAsp*	0.19	0.223	0.149
	Discourse	0.231	0.247	**0.170**
30	*SentAsp*	0.220	0.218	0.151
	Discourse	0.207	0.207	**0.152**

Table 6.1: Results in terms of purity, entropy and F-score. Values are in the range [0-1]

are very close or the *SentAsp* scores higher values of F-score. However, in all cases, we cannot test statistical significance, and this is one of the main drawbacks of this evaluation.

A second thing to note is that although we would expect to get the "best clustering" with the realistic configuration of $K = 10$, this is not the case for either models, since clusterings were found better when having the smaller number of clusters. Pedersen [53], in the context of word sense induction, empirically supported that F-score is sensitive to the number of clusters, which explains this consistent decrease in the reported F-score values of both models.

6.1.2 Supervised evaluation

For the supervised evaluation, we perform 10-fold cross validation. This means that we split our entire gold standard in 10 disjoint sets and in every fold we use the 9 sets to induce the mappings as we described in 5.6.2 and the 10th to perform the evaluation. Finally, we report results averaged across the 10 folds. The reported results are measured in terms of precision, recall and F1. Since our gold standard is very skewed, we believe that micro-averaging would not reveal the true performance of the systems and this is the reason we macro-average the results across different classes.

Table 6.2 presents the results on the entire gold standard. Apart from the two models *Discourse* and *SentAsp*, we also present the results of for two baselines, *mostFrequent* where all the instances are classified to the most frequent class, i.e. the "garbage collector" class *rest* and *randomClassDistribution* which assigns a random label according to the distribution of the labels in the training set.

Looking at the results, first of all we have to comment that since we apply a 1-1 mapping from the gold classes to the induced clusters, the number of *false positives* will decrease, since some topics will not be assigned to any class. Therefore, it is natural to observe a slight increase in the precision as the number of topics increases. Naturally, the exact opposite effect causes recall to decrease.

We observe that the *mostFrequent* baseline is relatively low. Even though the precision, recall and F1 for the most frequent class is 19.85% (i.e. the micro-averaged results), when macro-averaging the results and since the scores for all the other classes are 0.0% we end up having macro-averaged scores of 0.7%. On the other hand, our random baseline which takes into account the class distribution performs better than the *mostFrequent* baseline, but is still rather low. One of the reasons why we do not present results in terms of micro-averages, is that we expect to be very difficult for our

Topics	Model	Macro-precision	Macro-recall	Macro-F1
-	mostFrequent	0.7	0.7	0.7
	randomClassDistribution	3.9	3.8	3.8
5	*SentAsp*	10.48	12.43	9.32
	Discourse	13.55	14.90	**9.95**
10	*SentAsp*	15.03	10.20	9.17
	Discourse	16.51	13.79	**10.83** **
20	*SentAsp*	16.66	8.17	8.81
	Discourse	14.9	8.99	**9.07**
30	*SentAsp*	16.38	6.64	8.01
	Discourse	16.9	10.5	**9.68**

Table 6.2: Results in terms of macro-averaged precision, recall and F1. Values are in the range [0-100]. ** denotes statistical significance with p<0.01 as calculated with paired t-test

models to correctly classify the most frequent class. This is because even though it is defined in the annotation as the "garbage collector" class, our Bayesian model will still try to induce some structure out of this class and thus its elements most probably will not end in the same induced cluster.

Overall, our results seem encouraging, since by adding an extra set of latent variables modeling the discourse information not only we do not lose in performance, but instead we achieve in all the configurations better results than the discourse-agnostic model *SentAsp* and for the configuration K=10 our results are significantly better. When comparing all the results in terms of F1, we find that the optimal number of topics is 10. This seems to be a natural behavior since our gold annotation also contains 10 classes.

For getting better insights into our model, we conducted an experiment similar to the one presented in Somasundaran et al. [38]. There, the authors annotate their dataset with opinion frames and then conduct different analysis for the instances connected by some frame (*Connected*) and for the instances that are not connected to any other neighboring instances (*Singleton*). In our setup, we do not have gold annotations for the discourse, since we induce them in an unsupervised framework. One could argue that we could divide our dataset in two disjoint sets depending on whether an instance/segment contains a discourse cue in the beginning of it. We therefore follow this setup and we use the annotated explicit and implicit connectives that exist in the Penn Discourse Treebank (Prasad et al. [54]), thus creating a list of 240 possible connectives. In total, we marked 549 instances (i.e. 35% of our dataset) as being "connected". Table 6.3 presents the 9 most frequent connectives that account for the 85% of the "connected" instances.

Connective	Frequency
and	36%
but	21%
so	8%
if	4%
when	3%
because	2%
although	2%
as	2%
after	2%

Table 6.3: List of the connectives that appear in our dataset with frequency larger than 1%. The rest 80% of the discourse cues account for the 15% percent of the instances.

Unfortunately, this method does not guarantee that what will end up being in the "connected" will actually contain genuine discourse connectives; that is because there exists sense ambiguity in the discourse vs. non-discourse usage (Pitler and Nenkova [55]) of connectives like *and*, *when* etc. On the other hand, since we do not draw an explicit correspondence between the discourse classes modeled by the RST theory and the ones induced by our model, we expect that in the "singleton" class we will find instances starting with a cue that is relevant for the sentiment domain but not marked as a cue in the Penn Discourse Treebank.

Even though we acknowledge the problem, we still perform the experiment in order to provide a better insight into the performance of the model. Table 6.5 presents the macro-averaged F1 values for the setup with the higher results for both models (for K=5 and 10). For this evaluation we report results on 7-cross validation, since the size of the sub-datasets are smaller.

Globally, we observe that the cues marked as "connected" create problems to both systems leading to relatively low results as compared to the "singleton" cases. For this reason, we manually inspected the segments of the "connected" dataset but after removing those classified as *rest*, since these are already a difficult case we have already mentioned before. Table 6.4 summarizes the different source of difficulties that we identified. In the examples 1-4 it is extremely difficult to identify the aspect since there is no explicit mention to it, but in the segment there is a pointer (marked with bold) to some mention of the aspect in (some) previous segment. On the other hand, in the examples 5-7 there is ambiguity on the choice of aspect since for example in segment 7 the tea facilities can refer to the breakfast the hotel (label *food*) or to the facilities in the room (label *rooms*). Finally, examples 8-10 are too short and not informative at all, and

	Content	Aspect	Sentiment
1	*but* certainly off **it** greatness	value	negative
2	*although* just for one night **this** was fantastic	service	pos
3	*and* while small **they** are very nice	rooms	pos
4	*but* **it** is not free for all guests	amenities	neg
5	*and* the water was brown	clean	neg
6	*then* this is the hotel for you	location	pos
7	*and* no tea making facilities	rooms	neg
8	*when* i checked out	service	pos
9	*and* if you do not	service	neg
10	*when* we got home	clean	neu

Table 6.4: Segments from the "Connected" dataset for which the manual assignment of sentiment and aspect is ambiguous without the local context, grouped according to different type of ambiguities

Topics	Model	Singleton	Connected
5	*SentAsp*	9.36	7.96
	Discourse	10.06	8.08
10	*SentAsp*	9.93	5.42
	Discourse	9.79	11.45

Table 6.5: Results for the subset of the instances that contain a discourse cue (Connected) and the subset that do not contain (Singleton)

this clearly indicates that the behavior of the segmentation tool sometimes does not match the given application and thus it would make sense to consider it jointly with the whole sentiment analysis task.

The *Discourse* model always results to better results than the *SentAsp* in the "connected", indicating that this modeling of discourse correlates with the "traditional" view it. Surprisingly, in the case of K=10, which is the setup where we use as many topics as classes, the F1 value even doubles. On the other hand, the "singleton" cases do not seem to be so sensitive to the discourse modeling leading to almost the same performance in both models. Therefore, from this results we can conclude that the *Discourse* model indeed induces some latent discourse structure.

cluster	Top words
clean#neg	room bathroom not dirty old walls floor carpet small bed clean rooms looked peeling wall stains
location#pos	hotel restaurants great street around good places right corner food nearby breakfast area close
location#neg	hotel subway get bus station walk easy right not take can airport street very around city train
service#pos	staff very stay helpful hotel friendly made great help desk always room feel service concierge front
rooms#neu	room tv large bed very bathroom small nice area flat space desk screen shower bedroom size closet
sleep#neg	not room have hotel very had rooms clean small could would much stay really out noise great only new bathroom
recom#pos	to stay again hotel would recommend here definitely will back go return staying next york time highly
food#neu	breakfast good room food very service restaurant coffee hotel great free bar nice buffet drinks morning lobby
amenities#pos	hotel very rooms lobby room nice great location new beautiful small good clean history grand little areas decor
rest#neu	stayed hotel nights here night just stay weekend two week husband returned spent days last wife trip recently

Table 6.6: Top words extracted from some clusters after manually filtering out stopwords from the language model produced when setting the number of topics to $K = 10$. The mappings were obtained from one fold of the cross-validation evaluation

6.2 Qualitative analysis

6.2.1 Language Models

To evaluate the quality of the clusters produced by the *Discourse* we looked at the language model obtained with $K = 10$ topics, which is the configuration that results in the best performance. As a general remark, the topics were very difficult to inspect because of the noise inserted by the stopwords. As we mentioned in 5.2, stopwords are an important building block of discourse connectives, and therefore we are not filtering them out. Furthermore, we observed that neutral words in every topic are very hard to explain, and it seems that they are a mixture of positive, negative as well as no-polarity bearing words. However, this behavior was more or less expected if we consider the contingency table presented in 5.4 and the fact that our annotators tended to disagree more on the assignment of neutral sentiment.bf

Table 6.6 presents the top words in some of the clusters the *Discourse* built. The

labels were assigned to the clusters by applying the mapping process we previously described in Section 5.6.2. As a general remark, most of the selected clusters seem to clearly "explain" the assigned classes (e.g *clean#neg*, *service#pos*, *food#neu*). However, there were also cases where we could not really identify in the cluster many words that were topically related to the assigned class. For example, in the cluster *sleep#neg* there is no strong evidence to explain neither the aspect nor the sentiment. Another example is the case of *location#neg*, where although we can spot words related to the *location* aspect, there are no words that express negative polarity, apart from the word "not". We believe that this pitfall of our model would be treated to some extent if we were able to include in the language models n-gram phrases and not just unigrams, as we do now, since we expect phrases to be more discriminant of sentiment than pure unigrams (e.g. we would induce *not easy* instead of *not* and *easy*). This future direction Finally, it is interesting to observe that the cluster mapped to the *rest#neu* is not simply a "garbage collector" cluster like the class itself. This seems to confirm our intuition that the model tries to learn a structure for the *rest#neu* and it groups together segments that refer to the abstract topic of "conditions for staying the hotel".

6.2.2 Induced discourse cues

To investigate the quality of the discourse structure that our model induces, we examined the extracted cue phrases for every discourse class. Table 6.7 presents a selection of cues that best explain the discourse class they have been associated to. 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 class *sameAlt* we can mostly observe phrases that tend to introduce a new aspect since an explicit mention to 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 like in the example:

Example 3. *It is not surprising that it is just steps from both the Four Seasons and Taj hotels.{**location#pos**} In addition , the staff is very friendly and courteous.{**staff#pos**}*

Class *altSame* is maybe the most interesting, since it seems to include cues that contain some anaphoric expressions, which might refer to previous mentions of an aspect in the discourse (i.e. previous segment). Another interesting fact is that we found the expressions *unfortunately, fortunately, the only thing* in the same class, since all indicate a change in sentiment. Finally, *altAlt* can be viewed as a mixture of the other two classes. Furthermore, in this class we can find expressions that are usually used at

Discourse class	Cues
altAlt	the rooms was, the hotel is, 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
altSame	but, and, it was, and it was, and they, although, and it, but it, but it was, however, which was, which is, which, this is, this was, they were, the only thing, even though, unfortunately, needless to say, fortunately
sameAlt	the location is , the room was, the hotel has, the hotel, the hotel is, and the room, and the bed, breakfast was, our room was, the staff were, in addition, good luck

Table 6.7: Induced cues that explain the discourse class

the end of some review since at this point we certainly change aspect and some times even sentiment from the previous segment. Some examples of these cases are *overall, we will definitely* and even the misspelled version of the latter *i would definately*.

However, there are some cases which cannot be explained. The phrases *my only complaint*, *the only problem*, *the only drawback*, *the bad things*,*now the negatives* are all extracted as cues by our model but associated with the discourse class that favors keeping the same sentiment (*sameAlt*). This is counter-intuitive though, since we would expect these phrases to indicate an alternation of sentiment and it may reflect peculiarities of the dataset or a pitfall of our model in some very specific cases.

6.3 Summary

In this section we presented both a quantitative and a qualitative analysis of our Bayesian model. In the unsupervised evaluation the results were difficult to interpret, since it has been shown that F1-score is sensitive to the number of clusters. On the other hand, the supervised evaluation confirmed our intuition that adding discourse information in a sentiment-aspect generative model yields better results. For the scenario where we have the same number of clusters as the number of classes, our system achieves the best results and it even significantly outperforms the discourse-agnostic model. Furthermore, we observed that the "connected" segments (i.e. segments that start with a discourse cue) contains cases which are very difficult to classify even manually when looking these segments in isolation to the context. Thus, our system even in these difficult cases performs better than the discourse-agnostic model.

The qualitative analysis of the language models did not yield so clear results. Examining these models a difficult task, since there are many stopwords in them, thus inserting noise. The fact that these models are based on unigrams results to limited

expressivity, thus the modeling of n-grams would lead to more discriminative phrases for each aspect and sentiment. Finally, investigating the induced cues gave us a better insight of the behavior of our model and the results verify that the induced structure provide meaningful modeling of the discourse structure even under this unsupervised approach.

CHAPTER 7

Conclusion

In this work, we presented a Bayesian model with weak supervision which jointly induces the aspect and sentiment information from opinionated texts by taking into account the discourse structure. For obtaining a fine-grained analysis, we work on the sub-sentential level by linearly segmenting our sentences with a discourse segmentation tool. The modeling of the local structure is realized with a set of two latent variables; the discourse class and the size of the discourse cue. The former aims at encoding the different constraints that subsequent segments of texts should adhere to. The latter denotes the cue phrase of every segment that signals the relation that two subsequent segments exhibit. For the parameter estimation of our Bayesian we use Gibbs sampling, which is enabled with the use of conjugate priors.

We conducted extensive quantitative and qualitative analysis. The former indicated that our model yielded significantly better performance than the discourse-agnostic model, demonstrating that inducing discourse information appropriate for sentiment is a promising task. When conducting separate analysis on the dataset, we observed that the segments that are connected via some discourse connective with other segments are more sensitive to the discourse model. This interesting finding proposes that the induced structure is related with the "traditional" view of discourse analysis. On the other hand, the qualitative analysis of the induced structure empirically verifies our results, since the majority of the induced cues are meaningful discourse connectives for the discourse class they have been associated with, even if our model does not use any kind of supervision for the discourse part.

7.1 Future work

There are a number of future directions we could follow. Concerning the representation of discourse, since we empirically verified that our model induces some meaningful structure, we would like to increase the expressivity of our model. In the current setup, the assignment of aspect and sentiment of a given segment is constrained by its discourse class and the aspect and sentiment of the previous segment. However, when analyzing reviews we found some discourse connectives (e.g. *in addition to*, *although*, *in comparison to*) which indicate that the assignment of sentiment and aspect on a given segment is going to be influenced by the **next** segment as in the Example 4, and not the previous one.

Example 4. ***In addition to*** *our spacious room, the shower was fantastic .*

Thus, we would like to encode all these linguistic intuitions in the definition of new discourse classes.

In general, our current representation models linear dependencies, which results in a flat and thus restrictive structure. However, in many theories, discourse exhibits a tree-like structure, which could be approximated in our by considering discourse to exhibit a hierarchical structure and implemented in a Hierarchical Bayesian model. Furthermore, this hierarchical structure will allow us to model discourse in different granularities and not only on the local structure.

An interesting point is that in the current model we are treating only explicitly marked discourse phenomena, since the concept of discourse is realized through the discourse connectives. If we would like to try and model implicit relations, we could let our model to condition not only on subsequent segments (i.e. **previous** or **next**), but also on further removed segments, thus allowing for longer distance dependencies.

Another direction concerns the pipeline of tasks that the preprocessing step includes, which results to a natural error propagation. From the manual annotation that we conducted and the analysis in Section 6.1.2, it became clear that the discourse segmentation is not the optimal one, creating some times very short segments, whereas other times it fails to distinguish segments in a sentence. We believe that this component should be tuned for the specific task of aspect-based sentiment analysis. Generative models are by their nature very flexible, thus allowing for easier extensions. Therefore, we are looking into ways for incorporating the **unsupervised** discourse segmentation (Dowman et al. [56], Eisenstein and Barzilay [57]) into our current Bayesian framework, which would allow us to test our model and the induced cues in languages other than English.

Bibliography

[1] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *"Proceedings of the ACL"*, 2004.

[2] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

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

[4] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

[5] Christina Sauper, Aria Haghighi, and Regina Barzilay. Incorporating content structure into text analysis applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[6] Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, 2009.

[7] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 2011.

[8] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003.

[9] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 1988.

[10] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[11] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[12] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, 2008.

[13] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[14] Theresa Ann Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh, 2008.

[15] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LRECâĂŹ06*, 2006.

[16] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domainspecific sentiment classification. In *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[17] Danushka Bollegala, David Weir, and John Carroll. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

[18] Sida Wang and Christopher Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 2012.

[19] Ivan Titov. Domain adaptation by constraining inter-domain variability of latent feature representation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[20] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, 2007.

[21] Ivan Titov and Ryan T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, 2008.

[22] S. R. K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. Learning document-level semantic properties from free-text annotations. *J. Artif. Int. Res.*, 2009.

[23] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

[24] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[25] Michaela Regneri and Rui Wang. Using discourse information for paraphrase extraction. In *Proceedings of EMNLP-CoNNL 2012*, 2012.

[26] Karl Pichotta and Raymond Mooney. Using discourse relations to improve script learning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 2012.

[27] Thomas Meyer and Andrei Popescu-Belis. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL-2012 Workshop on Hybrid Approaches to Machine Translation (HyTra)*, 2012.

[28] B. Webber, M. Egg, and V. Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 2011.

[29] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*, 2004.

[30] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[31] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP'11*, 2011.

[32] M. Taboada, K. Voll, and J. Brooke. Extracting sentiment as a function of discourse structure and topicality, 2008.

[33] Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters*, 2008.

[34] Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[35] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007.

[36] Christina Sauper, Aria Haghighi, and Regina Barzilay. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

[37] Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008.

[38] Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, 2009.

[39] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[40] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

[41] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1973.

[42] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. 2010.

[43] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Producing power-law distributions and damping word frequencies with two-stage language models. *J. Mach. Learn. Res.*, 2011.

[44] Geoffrey E. Hinton. Products of experts. 1999.

[45] A. Smith, T. Cohn, and M. Osborne. Logarithmic opinion pools for conditional random fields. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[46] Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. Unsupervised multilingual learning for pos tagging, 2009.

[47] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. In *In 46th Annual Meeting of the ACL*, 2009.

[48] Trevor Cohn, Phil Blunsom, and Sharon Goldwater. Inducing tree substitution grammars. *Journal of Machine Learning Research*, 2010.

[49] Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009.

[50] Gayatree Ganu, Noemie Elhadad, and AmÃl'lie Marian. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, 2009.

[51] E. Agirre and A. Soroa. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007.

[52] Joel Lang and Mirella Lapata. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

[53] T. Pedersen. Umnd2: Senseclusters applied to the sense induction task of senseval-4. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007.

[54] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[55] Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009.

[56] M. Dowman, V. Savova, T.L. Griffiths, KP Kording, J.B. Tenenbaum, and M. Purver. A probabilistic model of meetings that combines words and discourse features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2008.

[57] Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.