ADIOS LDA: When Grammar Induction Meets Topic Modeling

Samia Touileb, Lubos Steskal
Department of Information Science and Media Studies
University of Bergen
{samia.touileb,lubos.steskal}@uib.no

Abstract

We explore the interplay between grammar induction and topic modeling approaches to unsupervised text processing. These two methods complement each other since one allows for the identification of local structures centered around certain key terms, while the other generates a document wide context of expressed topics. This approach allows us to access and identify semantic structures that would be otherwise hardly discovered by using only one of the two aforementioned methods. Using our approach, we are able to provide a deeper understanding of the topic structure by examining inferred information structures characteristic of given topics as well as capture differences in word usage that would be hard by using standard disambiguation methods. We perform our exploration on an extensive corpus of blog posts centered around the surveillance discussion, where we focus on the debate around the Snowden affair. We show how our approach can be used for (semi-) automated content classification and the extraction of semantic features from large textual corpora.

1 Introduction

The information released by Edward Snowden ignited huge amounts of polarized online discussions. The information leak was massively discussed, and different opinions and points of view nourished the debates. People were discussing the issue from their perspectives and framed their discussions accordingly to their own beliefs. How people frame their discussions has been extensively studied in social sciences, where framing is defined as being a selection of features (from texts) to support given ideas that describe the strengths of a text [6]. Frames direct attention to some aspects of an issue, and simultaneously they direct attention away from other existing aspects. The included and excluded aspects characterize the frames and can affect the readers' opinions. Frames can be found in the sequences of words chosen to form sentences, paragraphs and documents. Language use is thus a very important element in order to identify the frames present in a text. Language use is represented by the word choices and the word order, which in turn represent structures in the language.

This paper was presented at the NIK-2016 conference; see http://www.nik.no/.

Regularities present in a language can be exploited to induce the most frequent word sequences of a corpus [7]. Lamb's work [8] was an early attempt to automate Harris' idea, where he introduced the concept of grouping words into a sequence of horizontal elements or words (H-groups - syntagmatic relations between words) and a set of vertical elements or words (V-groups - paradigmatic relations between words), which may help to characterize the meaning. For example the sentences "The girl ate a banana", "The girl ate a strawberry", "The boy ate a banana", "The boy ate a strawberry". The H-groups here are "the" and "ate a", the V-groups are (girl, boy) and (banana, strawberry). Harris' insights have also become the foundation of some of the work in the field of grammar induction, where the focus is to induce grammatical structures from raw texts and generate complete grammatical descriptions of texts [5].

Recently, the grammar induction algorithm ADIOS (Automatic DIstilation Of Structure [12]) has been modified to be applied for text mining purposes [11]. The algorithm uncovers the most important structures around given key terms and discovers what is said about given keywords and key concepts. Recall our previous example of the girl and the boy who ate fruits. The algorithm will induce structures in the form of regular expressions, such as $(\texttt{The}(\texttt{girl}|\texttt{boy})\,\texttt{ate a}\,(\texttt{banana}|\texttt{strawberry}))$, to be read as "the girl or the boy ate a banana or strawberry".

How issues are framed can vary relative to the theme of the discussions. For example, if the theme of a discussion is nature, then oil drilling will be framed negatively. But if the theme of the discussion is economy, then oil drilling can be framed positively. The context in which an issue is discussed is therefore very important. Algorithmic topic modeling encompasses methods that uncover the various themes of topics in a corpus [2]. These algorithms examine the words of a corpus in order to determine the underlying existing topics and how they are connected. They have been extensively used in recent years and have been applied to various purposes, including framing analysis [4, 1].

In this paper we aim to investigate how these three main ideas – framing, grammar induction, and topic modeling – can be merged together in order to have a better understanding of how various issues are discussed in different topics from different perspectives. We combine a topic modeling method – Latent Dirichlet Allocation (LDA) with a modified version of a grammar induction algorithm (ADIOS) in order to uncover how the language use about the same issue can differ from the perspectives of various topics. To the best of our knowledge, these two approaches have not been combined together before.

In Section 2 we define the methodological steps followed during this investigation. The results of this study are presented in Section 3. A discussion of our main contributions and findings are summarized in Section 4.

2 Methodology

We explore how language use around selected key terms differs within the topics of a debate. The main goal of applying topic modeling to our data is to provide a bird's eye view of our corpus and use it to extract additional semantic, topical or even framing features from individual documents. We see topic modeling as providing additional contextual information based on the entire corpus.

LDA is in stark contrast with the narrow window used in the ADIOS algorithm. We combine these two perspectives to uncover patterns that might otherwise remain hidden from each of these methods when used separately. With the use of LDA, we identify the most likely topic for each mention of a word in a document; which can help differentiate

between the various shades of a word's usage and meaning. We add this inferred topic information to each word (by adding the suffix _⟨topic_id⟩_ to each word). This way, words being used in different global contexts (topics) will be treated as different words. This allows us to go beyond basic word connotation disambiguation (e.g., river *bank* vs *bank* robbery). A word can have the same basic meaning, yet be present in multiple topics loading the word with different ways of interpretation.

Being able to incorporate this distinction, such word usage nuances will allow us for a more sensitive and insightful extraction of information structures. The analysis of the retrieved information structures can then offer a better understanding of typical sentence structures within topics and help uncover the actual connotation of a word invoked by a given topic. Conversely, it is also possible that a word is present in multiple topics without an observed shift in the word's meaning. The extracted information structures may help us identify such instances by (systematically) grouping these topically different forms of a word into V-groups.

Corpus

We harvested an English language surveillance debate corpus of roughly 100,000 blog posts. The corpus spans from March 2005 to June 2014. We repeatedly queried three major search engines Google, Bing and Yahoo! for twenty-one terms based on domain expertise. We repeated this process over the course of several weeks to allow for more variation in the results. The search results were restricted to three major blogging platforms – WordPress, Blogspot and Typepad. To get a clean text corpus, we preprocessed it using the boilerplate removal tool JusText [10].

Topic modeling with LDA

Latent Dirichlet Allocation [3] is an unsupervised probabilistic topic modeling method, operating under the bag of words representation of documents, which assumes that each document is sampled from a mixture of k topics, where a topic is a multinomial distribution over all words of the vocabulary. These distributions are not known a priori and are inferred during the learning process.

Given the topic distributions characterized by a $k \times |V|$ matrix β (where V denotes the vocabulary) and the distribution of topics over a document characterized by the vector θ, the probability of a document vector $\mathbf{w}$ is

$$p(\mathbf{w}|\beta;\theta) = \prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta)p(z_n|\theta),$$

where z_n is the topic associated with the n-th word, $p(z_n|\theta)$ is a multinomial parametrized by θ, and $p(w_n|z_n;\beta)$ is a multinomial over the words.

The objective of the method is, given the observed documents, to identify the topic distributions over each document and the topic assignment of each word instance in each document.

The parameters θ and β are assumed to be drawn from Dirichlet distributions with hyper-parameters α and η. We used the implementation of LDA provided by the MALLET toolkit [9] to infer the underlying topics.

Selecting the right number of topics and evaluating topic quality is a notorious problem that has a number of strategies to cope with, ranging from human driven qualitative analysis to fully automated quantitative methods. We were mostly focused on

ID	Top 10 keywords	Title
4	law legal data eu european mr rights uk public case	EU data protection
5	al intelligence pakistan agencies government india police security terrorist qaeda	Terrorism
6	nsa surveillance intelligence government data information program security agency national	NSA data collection
11	police law federal blog criminal attorney court crime judge california	Law and crime
12	data backup database retention file rman files policy server recovery	Data management
14	obama president bush administration house congress bill program senate white	US politics
15	intelligence war israel iran military nuclear cia china united states	Intelligence agencies
16	snowden nsa edward government security surveillance documents russia spying greenwald	Snowden
18	gt camera video surveillance system cameras phone technology device devices	Surveillance technology
19	internet data google privacy information facebook security users companies online	Internet and social media

Table 1: Selected LDA inferred topics.

having well interpretable topics without any strict requirements of the level of granularity. Due to practical reasons and the already topic specific corpus, we also preferred fewer topics to more.

We originally ran the algorithm with 20 and 60 topics and then manually evaluated topic quality by (i) inspecting the 20 most likely words of each topic; and (ii) for each topic reading through the 20 most topic specific articles.

We concluded that the 20 topic model offers mostly coherent topics of reasonable quality. One of the topics contained mostly function words and words that did not otherwise fit into the other topics. The model's hyper-parameters were inferred as the result of Mallet's hyper-parameter optimization setting.

We decided to focus our attention on three topics that seem to interact with each other and which cover an important part of the discussed story: topic *6 – NSA*; topic *16 – Snowden*; and topic *19 – Internet*. We also chose these topics since they are of general interest to media scholars. Our approach can be of course used to analyze any set of topics. A subset of the inferred topics together with the top 10 keywords for each topic is presented in Table 1. Note that we used all 20 topics for word annotation.

Inducing information structures using a modification of ADIOS

ADIOS [12] is an unsupervised algorithm that discovers hierarchical structures in sequential data. It identifies the most significant patterns (horizontal sequences similar to H-groups) and equivalence classes (vertical groups similar to V-groups) within the context of patterns, using statistical information. Each sentence of the corpus is loaded onto a directed pseudograph (loops and multiple edges are permitted) with one vertex for each vocabulary item, and where each sentence of the corpus is a path over the graph (partially aligned sentences share sub-paths across the graph). In each iteration, the most significant pattern is identified based on evaluating the ratio of flow from one node to the other relative to the in-flow of the first node. That favors frequent sequences occurring in

various contexts. During the next step, the algorithm identifies the existing equivalence classes within the context of the previously identified pattern. The algorithm identifies positions in the pattern that could be filled by different items and forms an equivalence class with those items. At the end of each iteration, the new pattern and equivalence class(es) become vocabulary items in the graph and can be embedded into non-yet-identified patterns and equivalence classes. This process forms hierarchical structures.

Salway and Touileb [11] modified ADIOS for text mining purposes in order to extract salient information structures from unannotated corpora. They modified the learning regime and how the input is presented. The input is presented as increasingly large snippets around key terms of interest. Selecting snippets around a predefined key term will lead the algorithm to only induce the structures present around it. The algorithm starts running on snippets with a small window of words around the key term and the window increases after a predetermined amount of iterations [11]. When the structures are induced, each structure is substituted with a unique ID in the text, and the algorithm proceeds running on the next snippet size. This is done in order to "force" the algorithm to find more patterns around the key terms and the previously induced structures [11]. The induced information structures capture both lexical and grammatical patterning. The structure's form has also been modified [11]: they are presented as regular expressions, where the elements of the structures are bracketed and the elements of the existing V-groups are separated by "|" representing "or".

We use the modified version of ADIOS [11] to uncover what is said about different key terms from different existing perspectives of the debate. We have selected three key terms from our blog surveillance corpus: *NSA*, *PRISM* and *Snowden*. These three keywords are the top ranking keywords relative to the intersection of the three topics *6*, *16* and *19*. We have selected the sentences containing one of the key terms and created separate snippet files for each of them, representing a total of 32.904 blog posts. The snippets were from different sizes starting with 0-3 words on the left or the right of the key term, and gradually increased to 4-6, then 7-9, and, at last, 10-12 words on either sides. This resulted in 367 structures for the key term *PRISM*, 149 structures for the key term *NSA*, and 363 structures for the key term *Snowden*.

3 Results

In this section, we report our main observations about combining induced structures and topic information. In section 3.1, we show how the induced structures can be clustered into different categories in order to facilitate the analysis. In section 3.2, we describe the connotation identification and disambiguation within some of the induced structures.

Categorisation of the induced structures

Our manual analysis of the induced structures around the three predefined key terms *NSA*, *PRISM* and *Snowden*, enabled us to define and classify the structures into five different types: grammatical; storytelling; characteristic of a topic; hook structures and named entities, while about 10% of the induced structures could not be categorized.

Grammatical structures

These structures describe the grammatical linguistic behavior in regards to a certain term. Grammatical structures included essentially sequences of grammatical units or of inflected verbs. For example, the structures *S1* and *S2* in Figure 1 are two grammatical

structures. Structure *S1* shows two different verbs ("say" and "believe") that have been used in the same context between the articles "to" and "that". Structure *S2* uses two different adverbs with a personal pronoun and a verb.

S1 (to(say|believe)that)
S2 (we(already|now)know)
S3 (snowden_16_(met_16_|departed_16_|is|appeared_14_|comes|said|worked_16_|
worked_17_| registered_13_))
S4 ((snowden_16_|snowden_6_)(will|may|would| might))

Figure 1: Induced grammatical structures.

Structure *S3* shows different verbs related to Snowden in a given context in topic *16*. The verbs are from various topics and have been embedded into one V-group since they have all been used with Snowden in topic *16* in the same context. The structure is a grammatical structure because it contains an equivalence class of verbs only. Structure *S4* is formed of an equivalence class containing different inflections of the verbs "will" and "may".

Storytelling structures

The storytelling structures are the most commonly induced structures. These structures are the key phrases that enable the understanding of what is being discussed. These structures have a narrating form and capture the most salient information present in the subcorpora built around the chosen key term. Such structures do not need to be complete phrases. Partial phrases are in some cases enough to grasp the main point conveyed by the structures. The added value of topic information is to shed light on different perspectives presented in the various topics which are actually sharing the same storytelling structures.

Figure 2 shows a selection of storytelling structures. Structure *S5* shows the different ways in which "edward_16_" is talked about in topics *6* and *16*. It demonstrates that topic *16* refers to Snowden as a leaker and a whistleblower, while topic *6* refers to him only as whistleblower. It is important to bear in mind that this case happens solely in the context of "prism_16_" or "prism_6_", and the context of "edward_16_". *S6* talks about how and why NSA collects data. Even if the word data is not present explicitly in the structure, it is very clear what the structure is about. The V-group (ability_6_|goal_6_|power_14_) shows the different existing angles from which the issue is discussed, from the NSA's ability to do it to its actual goal and its real power to do so.

Structure *S7* mostly represents topics *16* and *6* and represents Snowden's opinions. It does not represent his opinions per se, but it shows the precursor of his opinion. It would be very easy to find the articles where Snowden says and gives his opinion and perspective of the story. The structure can be used for example as a query to extract the missing information from the texts, in order to extract and visualise the entire information.

Structure *S8* shows the three charges that were held against Edward Snowden. The added topic information enabled us to see how topics *9* and *16* discuss the same issue from different perspectives. While topic *16* relies on presenting all the charges, topic *9* is mostly concerned with the criminal charges. It is again important to stress that this is uniquely true in the context of the word "charges_16_".

The perception of Snowden is presented in structure *S9*. There are two main perspectives: Snowden as a whistleblower; and Snowden as a U.S. spy. The two

S5 ((prism_16_|prism_6_)(leaker_16_|whistle_16_ blower_16_|whistleblower_16_|whistleblower_6_)edward_16_)

S6 (the(ability_6_|goal_6_|power_14_)(of(thensa_6_)|thensa_16_)to collect_6_)

S7 ((snowden_16_|snowden_6_|edward_16snowden_16_|((mr_4_|mr_16_)snowden_16_)|((says|after|reports_16_|journalist_16_|interview_16_|interviewed_16_|interviewed_6_|since)(edward_16snowden_16_|(edward_6_|(snowden_16_|snowden_6_)|snowden_16_)))) (asserted_16_|sits_16_|says|stranded_16_|was| appeared_16_| claims_16_)he)

S8 ((criminal_16_|criminal_9_|felony_16_| espionage_16_|three)charges_16_)

S9 ((whitles_16_blower_16_|former(u_s|us)(spy_16_|spy_6_))edward_16_snowden_16_)

S10(u_s_(cyber_15_espionage_16_|cyber_16_espionage_16_|data_6_monitoring_6_)program_6_)

Figure 2: Storytelling structures.

perspectives are present in topic *16*, while topic *6* only refers to him as a spy. Structure *S10* shows how the *PRISM* program is talked about. Topics *15* and *16* refer to it as "*cyber espionage*", while topic *6* calls it a "*data monitoring program*". The expression "*cyber espionage*" seems to have more negative connotations than "*data monitoring*". Espionage refers to the activity of spying in order to uncover hidden secrets, while monitoring suggests the act of observing, listening and maintaining a constant surveillance over something.

Characteristic of a topic

Some of the induced structures exclusively, or to a greater extent, contained words from a unique topic. These structures are thus very representative of what is being said in that unique topic and characterise the topic's content. Figure 3 illustrates such structures. *S11* is a structure characterising mainly topic *16*, but the structure also shows that topics *16* and *6* overlap in how they discuss Edward Snowden. Many induced structures showed a considerable overlap between topics *6* and *16* when discussing our three key terms *NSA*, *PRISM* and *Snowden*.

Structure *S12* is also a characteristic structure of both topics *6* and *16*. The elements (words) of the structures are mainly from topics *6* and *16*, except for one occurrence of topic *19*. This is, once more, an example of the important overlap between topics *6* and *16* when discussing Edward Snowden.

S11((whistleblower_16_|(whistle_16_blower_16_)|whistle_16_blower_16)(((edward_16_snowden_16_|snowden_6_)has)|(edward_6_(snowden_16|snowden_6_))))

S12((snowden_16_|snowden_6_|edward_16_snowden_16_|((mr_4_|mr_16_)snowden_16_)|((says|after|reports_16_|journalist_16_|interview_16_|interviewed_16_|interviewed_6_|since)(edward_16_snowden_16_|(edward_6_|(snowden_16_|snowden_6_)|snowden_16_))))(leaks_6_|revelations_19_|revelations_6_|disclosures_6_))

Figure 3: Structures that characterise topics' content.

Hook structures

Hook structures are structures that might hint to interesting information about the corpus' content. They usually raise questions which a simple regular expression matching from the text will provide the answers to. These kinds of structures are very interesting and can be used as templates for information extraction. Figure 4 presents some hand picked examples of such structures. Structure *S13* tells what NSA will or not be able to do. What NSA will do is not explicitly present in the structure, but it is possible to retrieve the remainder of the missing information. The same case applies to structure *S14*, where it is implicitly apparent that this might hint to an interesting bit of information. Structure *S16* is also a hook structure that hint to more information about what the PRISM program actually is and is not.

Structure *S15* is particularly interesting. It shows two perspectives (implicitly about Snowden) from topic *16* and one from topic *9*. The two points of view from topic *16* are more of a moral judgement perspective, while topic *9* indicates a more pragmatic judgement (being a criminal and not following the law). The fact that the person being described is not present in the structure makes it a hook structure.

S13((the nsa_6_|the nsa_16_) will (still|soon|not) be able to)
S14(to (eavesdrop_6_|spy_16_) on)
S15((the|(of the|a|or)) (hero_16_|criminal_9_|traitor_16_))
S16(((the ((prism_16_|prism_6_)|(program_16_|program_19_|program_6_)) is)|
((prism_16_|prism_19_|prism_6_) is)) not a)

Figure 4: Structures categorised as hook structures.

Structures of named entities

These structures capture salient named entities. Some of the induced structures are: (edward_16_snowden_16_), (glenn_16_greenwald_16_), (the white_14_house_14_), (der_16_spiegel_16_), (hong_16_kong_16_), (united_16_states_16_). These structures provide us with information about some of the main actors mentioned in the discussions. Most of the structures categorised as named entities belong to topic *16*.

Connotation identification and disambiguation

Having the information structures extracted allows us to address the problem of word disambiguation and connotation distinction. Notice that the grammatical structures we extracted contain equivalence classes of words that appear in similar local contexts. Each word (except for stop words) comes together with a topic flavor identified by LDA – let us call this topic-specific instance a *word form*.

If a word always co-appears in equivalence classes with all its forms, it is likely that the meaning of the word is the same in all topics. On the other hand, if two word forms never appear together in an equivalence class but have equivalence classes of their own, then it is likely that they are used in different contexts and have different connotations or meanings.

In what follows, we suggest a measure for word disambiguation that we use on our data in order to identify potentially ambiguous and unambiguous word usages, in the context around the three selected key terms *NSA*, *Snowden* and *PRISM*.

For each non–stop word w in the information structures, our algorithm has extracted all equivalence classes that contained w. It has then annotated each such class with the

different forms of this word present in the respective equivalence classes. For example, if the word w was present in two equivalence classes (`w_0|u_7|w_3`) and (`w_0|w_2|v_6`), where `u` and `v` are some other words, then it would annotate these two classes as $\{0,3\}$ and $\{0,2\}$, respectively.

This way, we have for each word a set of annotations – a profile – representing what forms of that word co-occur together in equivalence classes. This allows us to define a coherence score for each profile. Intuitively, if all annotations in a profile are mutually disjunct, the profile is least coherent. On the other hand, if a profile contains only one annotation, then it is maximally coherent. A higher number of intersections between annotations means a higher coherence score, and vice versa.

Let a word w exist in k different word forms and let P_w be its profile. We associate P_w with a characterization vector $\chi(P_w) = (p_1, p_2, \cdots, p_k)$ where p_i represents the ratio of annotations that share this word form:

$$p_i = \begin{cases} 1 & \text{if } |P_w| = 1 \\ \frac{\left|\{A|(i \in A) \wedge (A \in P_w)\}\right| - 1}{|P_w| - 1} & \text{otherwise.} \end{cases}$$

For example, if the form i is only used by a single annotation, then $p_i = 0$, but if it is used by all annotations, p_i will be 1.

Comparing two k-dimensional characterization vectors $\chi = (p_1, p_2, \cdots, p_k)$ and $\chi' = (p'_1, p'_2, \cdots, p'_k)$ we say that χ is more coherent than χ' if $p_j \geq p'_j$ for all j. This is just a formalization of our intuition regarding the number of annotation intersections.

We have just defined a partial ordering on characterization vectors and thus on profiles. However, since this is only an (partial) ordering, there are many scoring functions that will be consistent with this ordering. We chose to use the average of the p_i values of the characterization vector χ.

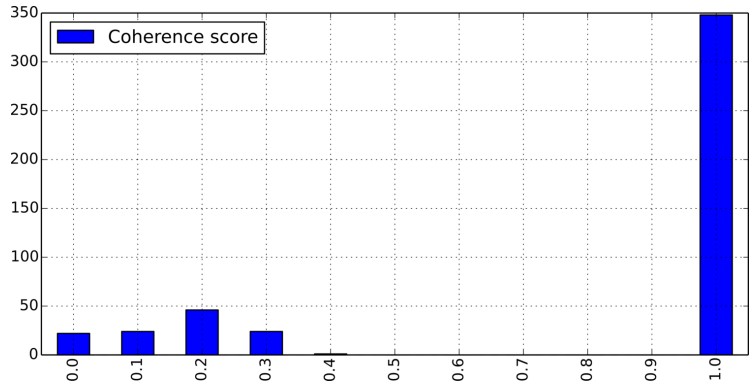

Figure 5: Word coherence histogram

We apply this coherence scoring function to identify the most ambiguous and disambiguous words. Figure 5 shows a histogram of coherence score distribution of topical words present in the extracted information structures. We can see that roughly 25% of the words have a score smaller than 0.4 and roughly 75% of the words have a score of 1.0.

A high coherence score does not necessarily mean that the word is not used in an ambiguous way; it only means that our method was not sensitive to picking up this ambiguity with respect to the word's usage in the context of the chosen keywords. Table 2 lists the 15 most *ambiguous* terms.

term	score	equivalence classes
public	0.000	$\{6\},\{16\}$
reports	0.000	$\{6\},\{16\}$
work	0.000	$\{6,8,12,19\},\{17\}$
believed	0.000	$\{5,14,17\},\{16\}$
president	0.000	$\{2,6,14\},\{16\}$
monitor	0.000	$\{18\},\{6\}$
employee	0.000	$\{16\},\{6\}$
appeared	0.000	$\{14\},\{16\}$
worked	0.000	$\{16,17\},\{6\}$
executives	0.000	$\{0,19\},\{6\}$
previously	0.000	$\{8,9,14,16,19\},\{6\}$
made	0.000	$\{4,6,15\},\{17\}$
track	0.100	$\{1\},\{6\},\{6,16,17,19\}$
claims	0.100	$\{6,9,17,19\},\{6\},\{16\}$
part	0.111	$\{17\},\{0,2,4,6,8,9,14,17,19\}$

Table 2: Twenty most ambiguous terms

An example of different connotations can be seen for the word *president* in Figure 6.

S18(((((prism_16_|prism_19_19|prism_6_) was) (launched_15_|launched_19_|
launched_6_) ((under|from) the) ashes_6_of) ((president_14_|presidents_2_|
president_6_) george_6_w)) ((bush_14_'s|bushes_14_'s| bush_6_'s) secret_6_))
S19(((russian_16_president_16_)(vladimir_16_ (putin_15_|putin_16_)))

Figure 6: The word *president* is used to refer to different persons and carry a different connotation.

Validation of the connotation disambiguation method

We argued that a low coherence score should be indicative of a word's ambiguity, while a score of 1.0 should indicate the opposite. In order to verify this hypothesis, we manually inspected all words with scores lower than 0.25 together with their respective information structures and assessed if they are in line with our expectations.

We classified the connotation of each word as *ambiguous*, *disambiguous*, or *unknown* based on its role in the associated information structures. If different annotations of the same word had actually different meanings or connotations, the word was labeled as ambiguous. If the word was actually used in the very same meaning, it was labeled as disambiguous. When it was not possible to identify if the word use is either ambiguous or disambiguous, it was classified as unknown.

After manually inspecting the relevant words, we concluded that around 44% of the words were identified as ambiguous, 15% as disambiguous, and 40% as unknown. When examining only the words with a zero coherence score, the distribution was 50%, 25%, and 25%, respectively.

This result might be interpreted in numerous ways, depending on the actual ambiguity of the words in the *unknown* group. We, however, believe that these results are encouraging for a further development of our approach.

Additional errors could have been introduced by the choice of the coherence scoring function, as well as by the performance of both topic modeling and information structure

extraction algorithms.

It is also worth noticing that a big proportion (75%) of the topical words presented in the information structures was classified as disambiguous from the information structure perspective. To offer a further qualitative validation of this approach, one could manually classify the relevant instances of the topical words in the context of the entire blog post where they were inferred from. This would, however, be a task beyond the scope of this paper.

4 Discussion and future work

In this paper we have explored the use of a method combining a grammar induction algorithm with a topic modeling algorithm. Both methods have been combined in order to uncover how different topics dealing with the surveillance issue discuss the three key terms *NSA*, *PRISM* and *Snowden*. We have shown that the structures induced from a corpus augmented with topic information uncovered the different aspects of the language use between the various topics.

We believe that this method, the combination of ADIOS and LDA, gives an overview and a description of the corpus' content; it also captures how different opinions are expressed differently and similarly in various topics. In addition, we were able to identify that topics *6* and *16* discussed our three key terms in a similar way, even though the topics themselves are distinct. This investigation has been conducted in order to understand the effects of combining such two different methods, and to explore how they could be applied for text analysis. The aim of the analyses could be to uncover the main positions in the issue and navigate through how they are being discussed and framed.

We also think that our approach gives more insights into the content of topics, and highlights their key content. We foresee a researcher using our approach to identify the major topics in a corpus, and how issues are discussed in each of them. This can lead to further investigations into the content of for example one specific topic of interest, or into what extent topics are similar or dissimilar in their language use. Our approach can also be used as part of identifying frames for framing analysis in social science research, as well as for content analysis of large corpora.

In summary, this paper makes contributions that can be seen from three different perspectives. Firstly, the induced structures with the topic information give a large overview of what is being discussed in real-world data regarding the perception of the Snowden affair as perceived by the bloggers' communities. We do not know if all perspectives can be found, but we believe that the results of this investigation shed light on a big portion of what is being and how it is being discussed.

Secondly, this work can also be seen as a description of an approach that can be used on large scale data to get better insights into the discourse structures. These structures can be further extended into different algorithms that capitalize on the different roles identified by the induced information structures. Some of the induced structures can help identify some of the main actors in the debate, while some others can represent information extraction patterns.

Thirdly, we believe that the method we are presenting in this work is a step towards a probabilistic language model capable of capturing both local (at the language use level) and global (at the topical level) relationships between words and sentences. Combining topic information with automatically induced information structures allowed us to understand and grasp the main points and perspectives discussed in a corpus from different topics.

Finally, both ADIOS and LDA are unsupervised methods, which makes this method portable between languages and data types (blogs, tweets, newspaper articles, etc.).

In future work, we aim to further develop the disambiguation analysis, to simplify equivalence classes by contracting word forms that are used interchangeably. In addition, it would be interesting to explore the possibilities of identifying a more automated method for categorizing information structures into the six categories identified in this paper. We also aim to determine what are the differences in topics and information structures that could be induced from the corpus pre- and post-Snowden affair. It will be interesting to see which topics die or emerge after the Snowden revelations, as well as the effects on the language use around some specific key terms.

References

[1] AHMED, A., AND XING, E. P. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (2010), Association for Computational Linguistics, pp. 1140–1150.

[2] BLEI, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84.

[3] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[4] DIMAGGIO, P., NAG, M., AND BLEI, D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics 41*, 6 (2013), 570–606.

[5] D'ULIZIA, A., FERRI, F., AND GRIFONI, P. A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review 36*, 1 (2011), 1–27.

[6] ENTMAN, R. M. Framing: Toward Clarification of a Fractured Paradigm. *J Communication 43*, 4 (dec 1993), 51–58.

[7] HARRIS, Z. S. Distributional structure. *Word* (1954).

[8] LAMB, S. M. On the mechanization of syntactic analysis. *Int. Conf. Machine Translation of Languages and Applied Language Analysis* (1961), 674–684.

[9] MCCALLUM, A. K. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[10] POMIKÁLEK, J. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic, 2011.

[11] SALWAY, A., AND TOUILEB, S. Applying grammar induction to text mining. *Procs. ACL* (2014).

[12] SOLAN, Z., HORN, D., RUPPIN, E., AND EDELMAN, S. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America 102*, 33 (2005), 11629–11634.