Applying Grammar Induction to Text Mining

Andrew Salway
Uni Research Computing
Thormøhlensgt. 55
N-5008 Bergen
Norway
andrew.salway@uni.no

Samia Touileb
Information Science and Media Studies
University of Bergen
N-5020 Bergen
Norway
samia.touileb@gmail.com

Abstract

We report the first steps of a novel investigation into how a grammar induction algorithm can be modified and used to identify salient information structures in a corpus. The information structures are to be used as representations of semantic content for text mining purposes. We modify the learning regime of the ADIOS algorithm (Solan et al., 2005) so that text is presented as increasingly large snippets around key terms, and instances of selected structures are substituted with common identifiers in the input for subsequent iterations. The technique is applied to 1.4m blog posts about climate change which mention diverse topics and reflect multiple perspectives and different points of view. Observation of the resulting information structures suggests that they could be useful as representations of semantic content. Preliminary analysis shows that our modifications had a beneficial effect for inducing more useful structures.

1 Introduction

There is an obvious need for text mining techniques to deal with large volumes of very diverse material, especially since the advent of social media and user-generated content which includes dynamic discussions of wide-ranging and controversial topics.

In order to be portable across domains, text genres and languages, current techniques tend to treat texts as bags of words when analyzing semantic content, e.g. for keyword-based retrieval, summarization with word clouds, and topic modelling. Such techniques capture the general "aboutness" of texts, but they do little to elucidate the actual statements that are made about key terms in the material. More structured and deeper semantic representations can be generated by information extraction systems for relatively restricted text genres and domains, but even then they are costly to port.

We see one particular area of application in elucidating the semantic content of social media debates about controversial topics, like climate change, both for casual users, and for social scientists studying online discourses. The complex, diverse and dynamic nature of the text content in such material presents a significant challenge for elucidating semantics. On the one hand, keywords alone will not convey what is said about important concepts, nor different points of view. On the other hand, modelling the semantics for information extraction purposes does not seem feasible given the breadth and diversity of the material.

Thus, we are motivated to develop a portable technique that generates representations of semantic content that are richer than keywords, and that can be applied to broad domains. Specifically, we seek to extract important information structures from an unannotated corpus comprising texts of the same genre and relating to the same domain.

Rather than using language-specific or domain-specific resources, we assume that important information structures in such a corpus will be reflected by patterning in the surface form of texts, such that they can be identified automatically through a distributional analysis (Section 2). Our approach is to induce information structures from an unannotated corpus by modifying and applying the ADIOS grammar induction algorithm (Solan et al., 2005): the modifications serve to focus the algorithm on what is typically written about key-terms (Section 3). To date we have implemented the approach to process 1.4m English-language blog posts about climate change: proper evaluation is ongoing but we are able to show

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 712–717,
Baltimore, Maryland, USA, June 23-25 2014.

examples of the semantic representations generated, discuss how they elucidate semantic content, and suggest how they might be used for various NLP tasks (Section 4). In closing, we make tentative conclusions and describe ongoing work (Section 5).

2 Background

Harris (1954; 1988) demonstrated how linguistic units and structures can be identified (manually) through a distributional analysis of partially aligned sentential contexts. His work suggests that it should be possible to induce syntactic descriptions from samples of unannotated text.

An early attempt to apply this thinking to computational linguistics was made by Lamb (1961) who described procedures for identifying "H-groups" and "V-groups". An H-group is a horizontal grouping of items (words and groups) that tend to appear sequentially, cf. a syntagmatic linguistic unit. A V-group is a vertical grouping of items that occur in similar linguistic contexts in a corpus, cf. a paradigmatic linguistic unit. As a toy example, take the H-group `'(the (woman | man) went to the (pub | shop | park))'`, with V-groups `'(woman | man)'` and `'(pub | shop | park)'`.

In more recent times, Harris' insights have become a cornerstone for some of the work in the field of grammatical inference, where researchers attempt to induce grammatical structures from raw text, e.g. ADIOS (Solan et al., 2005). In this field the emphasis is on generating complete grammatical descriptions for text corpora in order to understand the processes of language learning, rather than text mining; see D'Ulizia et al. (2011) for a review.

The unsupervised ADIOS algorithm recursively induces hierarchically structured patterns from sequential data, e.g. sequences of words in unannotated text, using statistical information in the sequential data. Each sequence (sentence) is loaded onto a directed pseudograph with one vertex for each vocabulary item: this means that partially aligned sequences share sub-paths across the graph.

In each iteration, the most significant pattern is identified with a statistical criterion that favors frequent sequences that occur in a variety of contexts. Then, the algorithm looks for possible equivalence classes within the context of the pattern, i.e. it identifies positions in the pattern that could be filled by different items and forms an equivalence class with those items. At the end of the iteration, the new pattern and equivalence class become vocabulary items in the graph, so that they can become part of further patterns and equivalence classes, and hence hierarchical structures are formed. For us, the terms "pattern" and "equivalence class" equate to the previously mentioned "H-group" and "V-group": we prefer the simplicity and literalness of these terms and use them henceforth.

3 Approach

For text mining purposes we do not see the need to induce a complete grammar for the corpus that we are mining. Rather, we are struck by Harris' further observation that the linguistic structures derived from a distributional analysis may reflect information structures, especially in the "sublanguages" of specialist domains (Harris, 1988). Thus, we propose to use a grammar induction algorithm to identify the most salient information structures in a corpus and take these as representations of important semantic content.

ADIOS has been evaluated on an interesting range of text corpora, and other kinds of sequential data. However, to the best of our knowledge, it has not been shown to successfully process a corpus with the scale and diversity of material that we envisage, e.g. 1.4m blog posts relating to climate change. This, along with our objective of identifying salient information structures rather than a complete grammatical description, led us to modify the learning regime to ADIOS. In the rest of this section we explain the modifications: please see 4.2.1 for a detailed description of how they were implemented.

To address the large scale and complexity of language use in social media, we modify the way in which text is presented to ADIOS by focusing separately on text around key terms of interest, rather than processing all sentences en masse. Our thinking here is in part influenced by the theory of local grammar (Gross, 1997), i.e. the idea that language is best described with word classes that are specific to local contexts, rather than general across the language.

Firstly, for each key term, we present only text snippets that contain that term: we expect there to be more salient patterning in snippets around a single key term because of repetition in the kinds of things written about it. Secondly, blog posts contain long and complex sentences so we process the clause containing a key term, and ignore the rest of the sentence. Thirdly, since we expect the key term to form more significant

units with words in its close proximity, we present the clauses in increasingly large snippets around the key term.

A further modification targets the most frequent and meaningful structures. After each iteration in which H-groups and V-groups are induced, the most frequent H-groups are filtered to remove any containing large V-groups which are likely to be more semantically nebulous. Instances of the selected H-groups are replaced with common identifiers in the input file so that patterning around them is more explicit in subsequent iterations.

4 Implementation

Here we report our first attempt to apply grammar induction to text mining. We chose to work with a corpus of blogs relating to climate change because they provide a challenging scenario with complex semantics, in which diverse topics – causes, effects, solutions, etc. – are discussed from multiple perspectives – scientific, political, personal, etc. – and with different beliefs (section 4.1).

We describe how we modified the learning regime of the ADIOS algorithm in order to induce H-groups and V-groups from an unannotated corpus (4.2.1). At this stage in our work, our focus is on observing the kinds of information structures that can be identified in this way, and in considering their potential applications as representations of semantic content (4.2.2). We also analyzed how results were affected by our modifications, i.e. the use of incrementally bigger snippets rather than complete clauses, and the iterative selection and substitution of frequent H-groups (4.2.3).

4.1 Input data

We used a corpus of about 1.4m unannotated English-language blog posts from 3,000 blogs related to climate change (Salway et al., 2013). Based on the relative frequency of words compared with a general language corpus, and the use of n-grams, we identified a set of domain key terms, e.g. 'climate change', 'greenhouse gases', 'carbon tax', 'sea levels'. From these we selected 17, with a mix of high (10,000's), medium (1,000's) and low (100's) frequencies.

For each key term we crudely extracted every clause it occurred in by taking a clause to be a sequence of words between punctuation. Pre-processing involved conversion to lower case, joining the words of key terms to make single items, e.g. 'greenhouse_gases', and substituting 'dddd' with 'YEAR', and other digit sequences with 'NUMBER': these changes all serve to make patterning more explicit.

Then, from the clauses for each key term, snippets of varying sizes were created. A snippet file for a key term is defined by (min-max) where there must be at least min words to one side of the key term, and no more than max words either side. Sets of snippet files were created for three different increment values: i = 2 (0-2, 3-4, 5-6, 7-8, 9-10, 11-12); i = 3 (0-3, 4-6, 7-9, 10-12); and, i = 4 (0-4, 5-8, 9-12).

4.2 Modifying the ADIOS learning regime

4.2.1 Method

In Section 3 we explained the rationale for our modifications to the ADIOS learning regime. They are detailed in steps 1 and 3-5 below.

For one key term and one increment value:

(1) INITIALIZE. Set the current input file to be the first snippet file for the key term and increment value, i.e. the smallest snippets.

(2) INDUCE CANDIDATE H-GROUPS AND V-GROUPS. Run the ADIOS algorithm over the current input file with default parameter values, except E=0.9 (cf. Solan et al. 2005).

(3) SELECTION. Filter the 5 most frequent H-groups to keep those that meet the following criterion: if the H-group contains a V-group then the V-group must contain < 6 elements. If none of the 5 most frequent H-groups remain then go to (5).

(4) SUBSTITUTION. For each selected H-group, replace all instances of it in the current input file with a common identifier. Iterate 10 times from (2).

(5) TRANSITION. Until the final snippet file is reached, set the current input file to be the next largest snippet file and substitute identifiers for the instances of all H-groups selected so far. Go to (2).

This process was executed for 17 key terms, with three increment values (i = 2, 3, 4). For further comparison, for each key term it was executed with complete clauses (ten iterations with selection and substitution) and with complete clauses (one iteration).

1. (((to (combat|fight))| (to (battle|slow|minimise|mitigate|tackle))) **climate_change**)
2. (**climate_change** (summit|adaptation|talks|meetings|convention))
3. (((greenhouse gases)|emissions|gases|(carbon emissions)|pollution) blamed ((for|to) **global_warming**))
4. ((cause|causes) (of **global_warming**))
5. ((dangers|signs|effect|consequences|perils) (of **global_warming**))
6. (to (confuse|mislead|educate) the public) *// from **global_warming** snippets*
7. ((anthropogenic|manmade|(man made)) **global_warming**)
8. ((would|should|to|must) (control|reduce|regulate|regulating|release) **greenhouse_gases**)
9. ((source|emitter|emitters|producers) of **greenhouse_gases**)
10. (the (effects|impact) ((under|of) ((a|its|the) **carbon_tax**)))
11. (a (modest|$_NUMBER a tonne|global|simple) **carbon_tax**)
12. ((will|would|to) (push|raise|elevate) (**sea_levels** (around|by)))
13. (((due to)|(caused by)) ((climate change)|(global warming))) *//from **sea_levels** snippets*
14. ((((the|global|some|sophisticated|complex) **climate_models**) (hint|show|indicate) that)

Table 1. A small selection of H-groups induced from snippets for a variety of key terms (in bold).

4.2.2 Results and potential applications

Table 1 presents a small selection of 14 H-groups that were induced from snippets with various key terms and increment values. Here, H-groups and V-groups are bracketed and nested. The elements of H-groups are separated by white space and the elements of V-groups are separated by '|'. Recall that the induction process selects frequent H-groups which, based on our assumptions, should reflect important semantic content.

This output would benefit from some post-processing, which is part of ongoing work. For example, in 1 there are two V-groups containing verbs that would be more elegantly expressed as a single V-group. There are also H-groups in which not all V-group alternatives make sense with the rest of the containing H-group due to over-generalization, e.g. 'to' in '…blamed ((for|to) global warming)' in 3. Despite these issues, some interesting and potentially useful structures are induced.

Some H-groups, we assume those resulting from the most stylized use of language in blogs, could perhaps be taken as the basis for information extraction templates, e.g. 11 where '$_NUMBER' is a slot for different amounts of tax, and 12 which captures various ways in which predictions about the amount of sea level rise can be written.

Other H-groups highlight some of the things typically written about key terms by grouping together different expressions of canonical statements, e.g. 3, 8 and 13. These could be used as a basis for summarizing the most important points of a topic, i.e. by taking 10,000's sentences and reducing them to 10's H-groups.

For broad topics it is desirable to perform finer-grained text classification and retrieval. The induction of H-groups such as 4 and 5 helps to identify different facets of a topic. In this case, the H-groups flag the causes of global warming and the effects of global warming as sub-topics, and show different ways in which they may be expressed.

The alternation in V-groups contained by H-groups may reflect different beliefs and opinions which could be used for text classification and opinion mining. In 14, the V-group 'hint|show|indicate' reflects different degrees of confidence that bloggers have in climate models. In 6, the alternatives in 'confuse|mislead|educate' reflect positive and negative views about public communication in the climate debate.

Semantically related terms, such as those captured in 1 and 5, have very different connotations and as such reflect different beliefs: consider the difference between someone writing about the 'effect of global warming' and the 'perils of global warming'. In other cases, alternation reflects different ways to say the same thing, e.g. the more or less synonymous terms that are captured in 2, 7 and 9 which would be useful for query expansion.

Key Term	Clauses	Number of different H-groups and *total instances*									
		i=2		i=3		i=4		clauses-10		clauses-1	
climate change	48241	198	*47000*	105	*52745*	86	*57799*	8	*31611*	698	*123531*
global warming	27582	191	*25998*	155	*30001*	104	*31850*	40	*32315*	397	*57388*
greenhouse gases	20345	174	*30148*	136	*34009*	94	*33846*	28	*25213*	552	*65167*
carbon tax	7751	106	*6727*	84	*8341*	80	*9859*	36	*11393*	128	*14988*
sea levels	6448	138	*8322*	121	*10246*	118	*11020*	55	*12090*	240	*16752*
climate models	6276	98	*5041*	91	*6020*	74	*6399*	26	*6061*	142	*11058*
emissions trading scheme	2989	86	*2243*	65	*3802*	68	*3140*	50	*7680*	96	*8118*

Table 2. Numbers of different H-groups and total instances generated from different input data.

4.2.3 The effects of our modifications

The numbers of H-groups generated by different executions of the induction process for each key term are shown in Table 2, i.e. three executions using snippets with different values of i, and two executions using clauses for comparison (cf. 4.2.1). The 10 omitted key terms (less than 1,000 clauses each) generated less than 25 H-groups for each value of i.

The high frequencies for clauses-1 are because no selection of H-groups took place, i.e. we simply take the normal ADIOS output. Based on our own inspections, some potentially useful H-groups were found in this output but, compared with other outputs, it was more common to see H-groups with large and semantically nebulous V-groups. This observation supports the iterative selection and substitution of H-groups with a limit on the size of V-groups. We also looked at the average number of V-groups in H-groups for each execution, as a way to compare the amount of structure in H-groups. This number was consistently lowest in results for clauses-1 which further supports our modifications.

A few potentially useful H-groups were observed in results for clauses-10, for which selection and substitution were applied. However the low numbers of different H-groups compared with all values of i suggests that it is better to use snippets as input rather than clauses.

The way in which the ratio of different H-groups and total instances varies for values of i suggests that starting with larger snippets (i=4) results in fewer H-groups but that these will capture more instances, i.e. they are more general. Whilst the H-groups for clauses-10 have many instances these tend not to capture useful patterning, i.e. they tended to describe combinations of key terms and function words.

5 Closing Remarks

At this stage in the research any conclusions must be tentative. However, it seems to us that the use of grammar induction to elucidate semantic content for text mining purposes shows promise. The H-groups shown in Table 1 provide richer semantic descriptions of the domain than keywords do, and we noted potential applications for high-level summarization of a whole corpus, the creation of information extraction templates and finer-grained text classification and retrieval. Importantly, the technique for generating H-groups would not require adaptation for use on a different corpus. The analysis in 4.2.3 suggests that the modifications that we made to the ADIOS learning regime had a beneficial effect.

Without a thorough evaluation we cannot make strong claims. In particular, we have little sense of the technique's recall, i.e. we do not know what information structures it missed. That said, it might be argued that since we expect the technique to be consistent in identifying patterning in the surface form of texts then its success will depend on the extent to which key terms are written about in consistent ways. This will of course vary between text genres and domains. Work has started on another corpus with more restricted language use and richer structuring was induced (Salway et al. 2014).

In other ongoing work we are looking more into the effects of the various parameters of ADIOS, and the necessity for our modifications. We are also seeking a deeper understanding of how the statistical information exploited by ADIOS relates to that which is captured by n-gram language models to describe sequences of words (cf. H-groups), and by established techniques to form semantic classes based on shared linguistic contexts (cf. V-groups).

Acknowledgments

We are very grateful to Zach Solan for providing an implementation of the ADIOS algorithm, and to Knut Hofland and Lubos Steskal for their roles in creating the NTAP blog corpus. This research was supported by a grant from The Research Council of Norway's VERDIKT program.

References

Arianna D'Ulizia, Fernando Ferri and Patrizia Grifoni. 2011. A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review* 36(1):1-27.

Maurice Gross. 1997. The Construction of Local Grammars. In: E. Roche and Y. Schabes (eds.), *Finite-State Language Processing*. The MIT Press, Cambridge MA: 329-354.

Zellig Harris. 1954. Distributional Structure. *Word* 10(2/3):146-162.

Zellig Harris. 1988. *Language and Information*. Columbia University Press, New York.

Sydney Lamb. 1961. On the Mechanization of Syntactic Analysis. *Int. Conf. Machine Translation of Languages and Applied Language Analysis.*

Andrew Salway, Knut Hofland and Samia Touileb. 2013. Applying Corpus Techniques to Climate Change Blogs. *Procs. Corpus Linguistics 2013*, Lancaster University.

Andrew Salway, Samia Touileb and Endre Tvinnereim. 2014. Inducing Information Structures for Data-driven Text Analysis. To appear in: *Procs. ACL Workshop on Language Technologies and Computational Social Science*.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. *Procs. of the National Academy of Sciences* 102(33):11629-11634.