Inducing Information Structures for Data-driven Text Analysis

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

Samia Touileb
University of Bergen
N-5020 Bergen
Norway
samia.touileb@gmail.com

Endre Tvinnereim
Uni Research Rokkansenteret
N-5015 Bergen
Norway
endre.tvinnereim@uni.no

Abstract

We report ongoing work that is aiming to develop a data-driven approach to text analysis for computational social science. The novel feature is the use of a grammar induction algorithm to identify salient information structures from an unannotated text corpus. The structures provide richer representations of text content than keywords, by capturing patterning related to what is written about key terms. Here we show how information structures were induced from texts that record political negotiations, and how the structures were used in analyzing relations between countries and negotiation positions.

1 Introduction

There is a widespread need for automated text analysis to be integrated into research methods for computational social science (e.g. Grimmer and Stewart, 2013). In order to analyze highly diverse content, techniques tend to treat texts as bags of words, e.g. for search, to summarize content with word clouds, and to model topics. Such techniques capture the general "aboutness" of texts, but they do little to elucidate the actual statements that are made about key concepts. Conversely, structured representations of statements can be generated, up to a point, by information extraction systems but these are costly to port to new languages and domains.

Thus, we are motivated to develop a portable technique that can generate richer representations of text content than keywords. Our idea is to adapt and apply a grammar induction algorithm to identify salient information structures in the surface form of texts. It seems to us that, to the extent that there is patterning, information structures may be induced from an unannotated text corpus with little or no need for language-specific and domain-specific resources. Unlike approaches under the rubrics of unsupervised and open information extraction (e.g. Riloff, 1996; Sekine, 2006; Etzioni et al., 2008), we avoid the use of parsers, part-of-speech taggers, and pre-specified entities for which relations are sought.

The approach that we envisage fits with the ethos of exploratory "data-driven" research. Rather than approaching a corpus with a hypothesis and an a priori coding scheme, a researcher is given an overview of the content in terms of computationally tractable information structures that were induced from the corpus. Such structures map to surface forms in text and can hence be used directly in quantitative analyses for further exploration and to test hypotheses, once they have been interpreted as interesting by a researcher. Working in this way will avoid the expense and bottleneck of manual coding, and reduce the potential for biases.

In the following we motivate our use of the ADIOS algorithm for grammar induction (2.1), and introduce the Earth Negotiations Bulletin (2.2). Section 3 describes our method and discusses the information structures identified in ENB texts. Section 4 takes some preliminary steps in using these information structures to identify dyads of (dis-) agreement and to extract markers of quantifiable negotiation positions. In closing, Section 5 offers some tentative conclusions and ideas for future work.

2 Background

2.1 Grammar induction for text mining

Harris (1954; 1988) demonstrated how linguistic units and structures can be identified manually through a distributional analysis of partially aligned sentential contexts. We are struck by Harris' insight that the linguistic structures derived from a distributional analysis may reflect

Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 28–32,
Baltimore, Maryland, USA, June 26, 2014.

domain-specific information structures, especially in the "sublanguages" of specialist domains (Harris, 1988). Whilst the textual material typically analyzed by social scientists may not be as restricted in content and style as that analyzed by Harris, our work proceeds on the assumption that, at least in some domains, it is restricted enough such that there is sufficient patterning for an inductive approach.

Harris' ideas about distributional analysis have become a cornerstone for some of the work in the field of automated grammatical inference, where researchers attempt to induce grammatical structures from raw text. In this field the emphasis is on generating complete grammatical descriptions for text corpora in order to understand the processes of language learning; see D'Ulizia et al. (2011) for a review.

For example, the unsupervised ADIOS algorithm (Solan et al., 2005) recursively induces hierarchically structured patterns from sequential data, e.g. sequences of words in sentences of unannotated text, using statistical information in the sequential data. Patterns may include equivalence classes comprising items that share similar distributional properties, where items may be words or other patterns. As a toy example of a pattern, take '(the (woman|man) went to the (house|shop|pub))', with equivalence classes '(woman|man)' and '(house|shop|pub)'.

2.2 The Earth Negotiations Bulletin

Within political science, text corpora provide a valuable resource for the analysis of political struggle and structures. For international climate negotiations, the Earth Negotiation Bulletin (ENB) constitutes an online record of the positions and proposals of different countries, their agreements and disagreements, and changes over time. As such it can provide insights into, e.g., how institutional structures and bargaining strategies affect policy outcomes. Since 1995, every day of formal climate negotiations under the UN Framework Convention on Climate Change (UN FCCC) and the Kyoto Protocol has been summarized in a separate 2-4 page issue of the ENB[1]. The ENB seeks to cover the major topics of discussion and which negotiators (referred to by country name) said what. The publication is used by scholars to address research questions such as whether countries with more extreme positions have more or less success (Weiler, 2012) and whether democracies and autocracies (Bailer, 2012) or developed and developing countries (Castro et al., 2014) behave differently in negotiations. From our perspective, the ENB's restricted content and style makes it appropriate to test our inductive approach.

3 Inducing Information Structures

We are investigating how the ADIOS algorithm (Solan et al., 2005) can be adapted and applied for mining the content of unannotated corpora; cf. Salway and Touileb (2014). Our objective of identifying salient information structures, rather than generating a complete grammatical description, leads us to modify the learning regime of ADIOS. Firstly, we modify the way in which text is presented to ADIOS by presenting sentences containing terms of interest (for the ENB texts these were country names), rather than processing all sentences: we expect more relevant patterning in these sentences, and think the patterning will be more explicit if not diluted by the rest of the corpus. Secondly, as described in more detail below, we focus the algorithm on frequent structures through an iterative process of selection and substitution.

3.1 Method

Our data set comprised all texts from the ENB volume 12, numbers 1-594, which cover the period 1995-2013. Preprocessing involved removing boilerplate text, sentence segmentation, and making all text lowercase. Then, all sentences mentioning one or more countries were selected. Every mention of a country, or a list of countries, was replaced with the token 'COUNTRY': this serves to make patterning around mentions of countries more explicit. A list of country names was the only domain- and language-specific resource required for the processing described below.

The resulting file of 32,288 sentences was processed by an implementation of the ADIOS algorithm, in which we modified the original learning regime to bias it towards frequent structures. After one execution of ADIOS we selected the five most frequent patterns (and any patterns contained within them) and replaced all instances of them in the input file with a unique identifier for each pattern: as with the 'COUNTRY' token, we believe that this serves to make relevant patterning more explicit. We executed ADIOS and selected and substituted frequent patterns nine more times.

[1] http://www.iisd.ca/linkages/vol12/

3.2 Results

In this way 53 patterns were identified, some of which are shown in Table 1 (patterns 1-7). Here patterns and equivalence classes are bracketed and nested. The sequential items in a pattern are separated by whitespace and the alternative items in an equivalence class are separated by '|'. 'COUNTRY' stands for a mention of a country, or a list of countries. In some cases we have manually merged and simplified patterns for clarity, but the structuring that they describe was all induced automatically.

Pattern 1 captures a simple relation between countries that appears frequently in sentences like 'China supported by Saudi Arabia said…'. It could thus be used as a simple template for extracting data about how countries align with one another (see section 4.1). Patterns 2-4 represent a multitude of ways in which a country's stated positions on issues can be reported. These patterns do not describe the issues, but could be used as cues to locate text fragments that do, e.g. by taking the text that follows 'COUNTRY said|noted|recommended|(etc.)…' (see section 4.2). Patterns 5 and 6 appear to have captured a wide variety of verbs and noun phrases respectively. Presumably these verbs relate to things that countries say that they will do, or that they think should be done. The noun phrases appear to raise topics for discussion; consider how pattern 6 appears as part of 7. There were other patterns that did not contain any equivalence classes: these often captured domain terminology, e.g. '(developing countries)', '(commitment period)'.

Patterns 1-6 all have a relatively shallow structure. In order to induce further structure we made new input files, based on what we saw in the initial patterns. We chose the most frequent 'speech acts' from patterns 2-4, and for each one made a separate file containing only sentences that contained 'COUNTRY SPEECH_ACT', e.g. one file that contained all the sentences matching 'COUNTRY expressed'. Each speech act file was processed with 10 iterations of selection and substitution (cf. section 3.1). The resulting patterns, including 8-10 in Table 1, do indeed have richer structures and show in a more nuanced way how countries' positions are reported in the ENB texts.

These results are encouraging for the idea of inducing information structures from an unannotated text corpus. The examples shown in Table 1 would not surprise anybody who was familiar with the ENB material. However, they provide a useful summary view of what is typically written about countries. Further, since they relate directly to surface forms in the text, they may be valuable for guiding further quantitative analyses, e.g. by pinpointing where significant expressions of country positions, arguments and affinities are to be found.

1. (COUNTRY ((supported|opposed) by) COUNTRY)
2. (COUNTRY (said|noted|recommended|explained|responded|stressed| questioned|addressed|reiterated| reported|urged|amended|invited…)); *the equivalence class contains 51 words*
3. (COUNTRY ((clarified|urged|reported) that)
4. (COUNTRY ((presented|demanded|outlined|favored (the|a))
5. (to (apply|safeguard|undertake|link|deliver…)); *the equivalence class contains 63 words*
6. (the (merit|cost effectiveness| merits|importance|idea…) of); *the equivalence class contains 84 words*
7. ((COUNTRY (noted|said|questioned|…)) (the (merit|cost effectiveness| merits|importance|idea…) of))
8. (COUNTRY expressed ((disappointment|concern) that)|((support|appreciation) for)|((readiness|willingness) to)|(satisfaction (with the) (outcome|reconstitution|functioning|work) (of the)))
9. (COUNTRY called (((for |on) (parties|(developed countries)) to)|((for a) (cautious|three phased|common|phased|bottom up|budget|global) approach to)|(for an (overview|elaboration|analysis|evaluation|examination) of)))
10. (COUNTRY highlighted ((the (need|basis) for)|(the (benefits|possibility|establishment) of)|(the (consideration|impact|impacts) of)|(the (use|involvement) of)|((the need to) (err|focus) on)|(the (role|importance) (of the))))

Table 1: Ten of the patterns automatically induced from Earth Negotiations Bulletin texts.

4 Using selected information structures

Here we describe our first steps in using some of the induced structures to infer coalitions (4.1) and to scale negotiation positions (4.2).

4.1 Dyads of support and opposition

The pattern '(COUNTRY ((supported|opposed) by) COUNTRY)', cf. Table 1, was used as a regular expression to extract instances where relations between countries were recorded with respect to stated positions. This gave 1145 instances of support, and 592 of opposition, often involving multiple countries; recall that 'COUNTRY' may stand for a list of countries. A count was made for each pair of countries in support and opposition, with a distinction made between 'C1 supported by C2' and 'C2 supported by C1'. Figure 1 is a scatterplot made from these counts. It shows, for example, that the US very often finds its statements supported by Canada. Further, whilst the US tends to support the EU relatively often, the EU supports the US only about as often as it opposes the US.

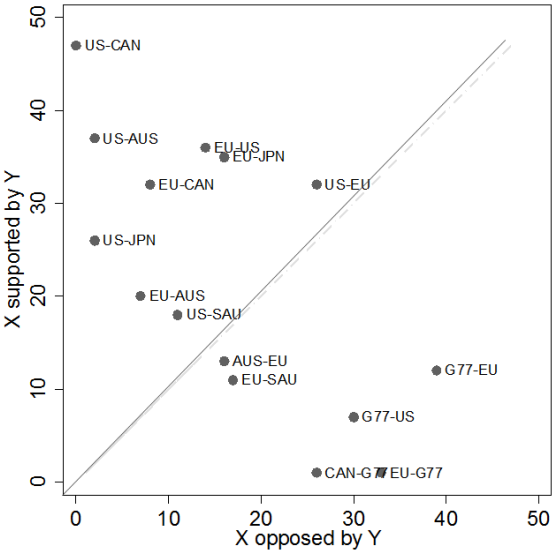

Figure 1: Dyads of support and opposition

4.2 Scaling negotiation positions

Patterns 2-4 from Table 1 were combined into a regular expression to extract instances of the statements made by countries. For each country a file was made with the text following every instance of 'COUNTRY said | noted | recommended | (etc.)', until the end of the sentence. The collection of country files was then analyzed with Wordfish (Slapin and Proksch, 2008): this tool, which implements a scaling model, positions texts (here reflecting countries) on a political/ideological dimension based on the relative frequency of discriminating words.

For the 40 countries with the most statements, the parameter indicating country position on the induced dimension ranged in ascending order from Austria (-2.38) via Belgium, Germany, the UK, Switzerland, the US, Canada, Australia, Norway, France, Russia, New Zealand to Japan (-.62) and on to Papua New Guinea (-.26), Tuvalu, Peru, Mexico, Brazil, Argentina, Malaysia, South Korea, Colombia, Saudi Arabia, Chile, Kuwait, Nigeria, Grenada, Uganda, Bangladesh, China, Egypt, the Philippines, South Africa, Indonesia, Venezuela, Iran, Bolivia, Barbados, India and Algeria (1.44).

The method thus perfectly identifies the main cleavage in international climate negotiations between developed and developing countries (cf. Castro et al., 2014). The bifurcation is robust to alternative specifications. Among the ten most discriminating lemmas used by developing countries are 'equal', 'distribut', 'resourc', 'histor', and 'equiti', suggesting an emphasis on fairness and rich countries' historical emissions.

5 Closing Remarks

The novel use of a grammar induction algorithm was successful in elucidating the content of a corpus in a complementary way to bag-of-words techniques: some of the induced structures were useful for guiding subsequent analyses as part of a data-driven approach to computational social science. Specifically, in this case, the structures facilitated text analysis at the statement level, i.e. statements about country relations and countries' positions. This meant we could plot country relations and scale country positions even though our source texts were not organized by country.

Given its inherent portability, we see the potential for applying the grammar induction approach to many other corpora, most obviously the other 32 ENB volumes, and other texts with similarly restricted content and style, e.g. parliamentary proceedings. It remains a largely open question as to what happens when the text input becomes more heterogeneous, but see Salway and Touileb (2014) regarding the processing of blog posts.

In ongoing work we are seeking to understand more about how the parameters of the ADIOS algorithm, and the modifications we make, affect the set of structures that it identifies. Also we are considering evaluation metrics to validate the induced patterns and to measure recall.

Acknowledgements

We are very grateful to Zach Solan for providing an implementation of the ADIOS algorithm, and to Knut Hofland for his help in creating our corpus of ENB texts. This research was supported by a grant from The Research Council of Norway's VERDIKT program.

References

Stefanie Bailer. 2012. Strategy in the climate change negotiations: do democracies negotiate differently? *Climate Policy* 12(5): 534-551.

Paula Castro, Lena Hörnlein, and Katharina Michaelowa. 2014. Constructed peer groups and path dependence in international organizations. *Global Environmental Change*.

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.

Oren Etzioni, Michele Banko, Stephen Soderland and Daniel S. Weld. Open Information Extraction from the Web. *Comms. of the ACM* 51(12): 68-74.

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3):267-297.

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

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

Eileen Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text. *Procs. 13th National Conference on Artificial Intelligence (AAAI-96)*:1044-1049.

Andrew Salway and Samia Touileb. 2014. Applying Grammar Induction to Text Mining. To appear in *Procs. ACL 2014*.

Satoski Sekine. 2006. On-Demand Information Extraction. *Procs. COLING/ACL 2006*: 731-738.

Jonathan Slapin and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52(3):705-722.

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

Florian Weiler. 2012. Determinants of bargaining success in the climate change negotiations. *Climate Policy* 12(5):552-574.