Automatic identification of unknown names with specific roles

Samia Touileb
Language Technology Group
Department of Informatics
University of Oslo
samiat@ifi.uio.no

Truls Pedersen
Department of Information
Science and Media Studies
University of Bergen
truls.pedersen@uib.no

Helle Sjøvaag
Department of Media
and Social Sciences
University of Stavanger
helle.sjovaag@uis.no

Abstract

Automatically identifying persons in a particular role within a large corpus can be a difficult task, especially if you don't know who you are actually looking for. Resources compiling names of persons can be available, but no exhaustive lists exist. However, such lists usually contain known names that are "visible" in the national public sphere, and tend to ignore the marginal and international ones. In this article we propose a method for automatically generating suggestions of names found in a corpus of Norwegian news articles, and which "naturally" belong to a given initial list of members, and that were not known (compiled in a list) beforehand. The approach is based, in part, on the assumption that surface level syntactic features reveal parts of the underlying semantic content and can help uncover the structure of the language.

1 Introduction

One important factor in media coverage of news is the use of diverse sources. Both prominent and marginal voices should be able to express their opinions and views, and be taken into consideration by media outlets. While prominent voices are easily identified and might already be compiled in existing lists (e.g. top level politicians or judges), the marginal ones remain unknown for most readers, and are often not included in precompiled lists. Marginal voices are not necessarily unknown by the public, but they are voices that are not that prominent in a given media coverage. In order to be able to actually quantify to what extent marginal voices are reflected in the media, it is important to be able to identify them.

Automated approaches can be used to identify the voices present in a large corpus of texts. Advances in Natural Language Processing (NLP) approaches allow the automatic identification of the named entities present in a corpus. While these advances have yielded good results for English, they are still at a basic stage for Norwegian. To the best of our knowledge, no approaches can identify the type of an entity (if it is a person name, an organization, ...) for Norwegian texts. However, there are some efforts made towards it, and we are able to identify entities from texts without being able to identify their types (Johansen, 2015). For Norwegian, it is still necessary to add a second step of analysis, usually a manual one, to be able to identify the type of each entity. Even when or if this approach succeeds, there is still a need for additional tools that classify a particular individual within a class of entities.

Linguistic theories show that words having the same context tend to share the same meaning (Harris, 1954), where context here is identified as the words appearing before and after a given word. We believe that this applies to named entities as well. We therefore argue that given a list of known person names, we can identify new unknown persons having similar roles, by analyzing the context in which their names co-occur.

In this work we focus on prominent Norwegian politicians that are mentioned in news, predefined in existing lists, and try to identify the marginal ones, which are not in these lists. The marginal politicians do not need to be unknown, they just have to not be part of a predefined list of politicians. The marginal politicians can be mayors, political representatives, or foreign politicians. Since very few or no resources

Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 150–158
Santa Fe, New Mexico, USA, August 25, 2018.

exist for Norwegian, we rely on an unsupervised approach to automatically identify these unknown politicians.

In what follows, we start in Section 2 by giving an overview of the literature on which we build our work. Then, in Section 3 we give an overview of the corpus used during this investigation, along with the seed list of politicians. In Section 4 we define the methodological steps we followed, and present in Section 5 the results of this study. A discussion of our main contributions, findings, and future avenues are summarized in Section 6.

2 Literature review

Named entity recognition (NER) approaches allow us to automatically identify entities in a text, as well as which type of entities these are, ranging from person names, to organizations and places. NER systems automatically identify entities and their types in texts. These methods have been extensively developed for English, but no off-the shelf approaches currently exist for Norwegian. Some work has been made in this direction, so called Named Entity Chunking (NEC) (Johansen, 2015), that only focuses on the identification of the entities, without specifying their types. It is therefore still necessary, for Norwegian texts, to use pre-compiled lists of entities to be able to identify which types a NEC has located in a text. Challenges emerge, however, if we want to identify specific types of persons in texts, without actually knowing these persons' names.

Our motivations are based on the identification of marginal voices present in news articles. The marginal voices are those who are not precompiled in known lists, but who still express opinions in the news. When looking for indicators of source diversity in the news, the aim is to ascertain the extent to which a range of voices gain access to the sphere of public debate. Based on principles of representation in democratic deliberation procedures (Bennett, 1990; Brown et al., 1987; McQuail, 1992), a pluralistic media landscape reflects the ability of media systems to curtail imbalances in the distribution of social, political and economic power in society (Jeffres et al., 2000). As the news media is the primary arena where actors can exchange ideas and opinions (Skogerbø and Krumsvik, 2015), access to this space should ideally be equal for all (Baker, 2006).

Research shows, however, that marginal voices are often excluded from this arena (e.g. (Kleemans et al., 2017; Ross et al., 2013)). This is partly because the agenda setting function of the media (McCombs, 2005), and the critical and competitive nature of journalistic professionalism (Cook, 1998; Sparrow, 1999), tend to exclude marginal voices - especially voices that do not easily fall within the established narrative frames of political journalism (Wolfsfeld and Sheafer, 2006). The exclusion of marginal voices is especially prevalent in political journalism (Alexseev and Bennett, 1995; Baumgartner and Chaqués Bonafont, 2015; Figenschou and Beyer, 2014; Shehata, 2010). By and large, these findings reflect the fact that most of this research is conducted on elite media, while in fact, local media tend to include a broader range of voices (Berkowitz and Beach, 1993; Allern, 1996; Ross, 2007). In cases where media regulation aims to preserve the type of media outlets that do, in fact, contribute to increase the diversity of voices in the news (cf., (Kulturdepartementet, 2017)), we therefore need measures to empirically establish what those outlets are. Because automatic classifiers, topic modelling and conventional content analysis methods tend to negate marginal classification, improvements to these methods are sorely needed.

Research in linguistics has shown that it is possible, by only analyzing the surface form of the language, to identify structures yielding important information (Harris, 1954). Especially in the sublanguages of specialized domain languages (Harris, 1988). Information structures are linguistic structures representing the distributional structure of the language. Aligning sentences can uncover which words are typically used with selected words in specialized languages, and words that share the same context tend to have the same meaning. We believe that the journalistic language used to discuss politicians is specialized and stylized, which makes it subject to the repeated use of some linguistic patterns.

One of the first attempts to automate the identification of information structures is the work of Lamb (1961). He developed an approach to identify in a corpus of texts the different grouping of words into fixed parts of sentences, and interchangeable parts. His method would for example identify from these

sentences "John bought a car", "John bought a bike", "John bought a house", "Norah bought a car", "Norah bought a bike" and "Norah bought a house", the following fixed part "bought a", and the interchangeable parts (John, Norah) and (car, bike, house). The fixed parts represent a grouping of sequences of horizontal elements/words (H-groups), and the interchangeable parts represent a set of vertical elements/words (V-groups). These information structures shed light on how words are being used similarly or differently in the same context next to given words.

Some of the work in the field of grammar inference are built on the insights of Harris (1954) and the work of Lamb (1961). Grammar induction approaches rely on the structural aspect of the language to extract information from textual data. These methods are usually used to induce complete grammars for a given language or text. However, some work has shown that methods developed in grammar induction can be used to automatically induce information structures that can identify what is being said about given issues, how it is being said, and can reflect some of the most important and distinctive content of a corpus (Salway et al., 2014; Touileb and Salway, 2014; Salway and Touileb, 2014).

An unsupervised grammar induction algorithm (first presented in (Solan et al., 2005)), that discovers hierarchical structures in sequential data, has been modified by Salway and Touileb (2014) into a text mining approach – henceforth referred to as SIMMS (Structure Induction for Mining Meaningful Snippets). The latter automatically, and in an unsupervised way, induces information structures from unannotated corpora. This algorithm is able to identify some of the most significant patterns (horizontal sequences of words) and equivalence classes (vertical groups of words) within the context of patterns, using statistical information, and running over a predetermined set of iterations. The patterns and equivalence classes respectively resemble the H-groups and V-groups presented by Lamb (1961). The method focuses on text snippets around key terms of interest rather than processing entire sentences. Instances of the most frequent induced patterns, representing information structures, are then replaced with unique identifiers in the input to make patterning around them more explicit in subsequent iterations. The induced structures are in the form of regular expressions, where elements of patterns are separated by whitespace, and the elements of the equivalence classes are separated by "|" representing "or". Considering the previous example of John and Norah, this approach will induce structures of the form "`(John|Norah) bought a (car|bike|house)`" which can be read as John *or* Norah bought a car *or* bike *or* house. We will in what follows refer to information structures and patterns interchangeably.

Based on this method and the linguistic theories on which it is built, we believe that given a sufficiently large corpus and a seed list containing a set of voices (both often described in specialized language and sufficiently homogeneous), we may isolate a set of new names (not contained in the seed list) which naturally extend the seed list. In our case, given a set of politicians' names, we may automatically find a disjoint set of names which will contain a relatively high portion of politicians. We therefore propose a method for expanding a list of known names to include previously undiscovered names occurring in similar contexts. In the following, we outline the steps in the implementation of an approach to enlarge a list of Norwegian politicians' names to include politicians which do not appear in the preexisting list (which we refer to as the seed list), from a corpus of news texts.

3 Data

We use two data sets: a large corpus of newspaper articles and a list of Norwegian politicians. The corpus of Norwegian news articles was scraped hourly from 125 online newspapers, between October and December 2015 and 2016. These newspapers reflect diversity in ownership (state (one outlet, the Norwegian Broadcasting Corporation), corporate, foundation, or independent), and distribution (local (92 outlets), metropolitan (25 outlets) or national (8 outlets)). This resulted in more than 600,000 articles. We show in Table 1 the newspapers, and their distribution category.

The list of politicians contained 368 Norwegian politicians in parliament and government, as well as mayors and their deputies. The names of politicians from the parliament and government were collected from the open data of the Norwegian Parliament[1]. The names of mayors and their deputies were manually

[1] https://data.stortinget.no/no/dokumentasjon-og-hjelp/dagens-representanter/

National: Aftenposten, Dag og Tid, Dagbladet, Dagen, Dagens Næringsliv, Dagsavisen, Dinside.no, Fiskeribladet Fiskaren, Klassekampen, Morgenbladet, Nationen, NRK, TV2, Vårt Land
Metropolitan: Adresseavisa, Agderposten, Avisa Nordland, Bergensavisen, Bergens Tidende, Budstikka, Drammens Tidende, Fædrelandsvennen, Glåmdalen, Gudbrandsdølen Dagn, Haugesuns Avisa, Harstad Tidende, iTromsø, Moss Avis, Nordlys, Oppland Arbeiderblad, Rogalands Avis, Romerikes Blad, Romsdals Budstikke, Sandefjords Blad, Sarpsborg Arbeiderblad, Stavanger Aftenblad, Sunnmørsposten, Telemarkavisa, Tønsbergs Blad, Varden, Østlands-Posten
Local: Altaposten, Akershus Amststidende, Arbeidets rett, Arendals Tidende, Askøyværingen, Aura Avis, Aust-Agder Blad, Avisa Nordhordland, Bladet Vesterålen, Brønnøysunds Avis, Bygdanytt, Bygdebladet Randaberg, Bømlonytt, Demokraten, Driva, Eidsvol Ullensaker Blad, Dølen, Eikerbladet, Enebakk Avis, Fanaposten, Firda, Firdaposten, Fjordenes Tidende, Fjordingen, Framtid i Nord, Fremover, Gjesdalbuen, Gjengangeren, Groruddalen, Halder Arbeiderblad, Hadeland, Helgelendingen, Hallingdølen, Hardanger Folkeblad, Helgelands Blad, Hitra-Frøya, Hålogalands Avis, iFinnmark, Indre Akershus Blad, Innherred Folkeblad, Jarlsberg Avis, Jærbladet, Kanalen, Kragerø Blad, Kvinneheringen, Kyst og Fjord, Laagendaksposten, Lierposten, Lillesandsposten, Lindesnes Avis, Lofotposten, Lygdals Avis, Lokalavisa Nordsalten, Møre-Nytt, Namdalsaisa, Norddalen, Nordre, Nye Troms, Porsgrunns Dagblad, Rakkestad Avis, Rana Blad, Ringerikes Blad, Ringesaker Blad, Røyken og Hurun Avis, Saltenposten, Sande Avis, Sandnesposten, Smaalenes Avis, Sogn Avis, Solabladet, Stjørdalens Blad, Sunnhordland, Sunnmøringen, Svalbardposten, Svelvikposten, Telen, Tidens Krav, Troms Folkeblad, Trønderbladet, Tvedestransposten, Tysnes, Tysvær Bygdeblad, Valdres, Vennesla Tidende, Vestby Avis, Vestnesavisa, Vestnytt, Vikebladet Vestposten, Østlandets Blad, Øyene, Åndalsnes Avis, Ås Avis, Asane Tidende

Table 1: Overview of the newspapers in the corpus.

gathered from various online sources. The list comprises 298 names representing county mayors, county deputy mayors, sate secretaries, municipal councils, city councils, local chairmans, county councils, city council representatives, and local councils.

4 Approach

We start with the assumption that texts featuring politicians have certain common context-dependent characteristics across the corpus. This should therefore allow for the identification of language use patterns around the politicians' names and allow us to identify politicians that were not included in the seed list.

Given a corpus (C), we produce a list of all names occurring in it and isolate those that appear on the seed list of politicians. We select all sentences from C in which some politician is mentioned and apply SIMMS to these sentences. This yields a set of patterns, from which we may isolate a sub-corpus (D) of sentences (not full texts) which match these patterns. Names will occur in a number of sentences matching a number of patterns. We take the sum of these frequencies to denote the names' score. Names which score well, are presented to a human expert for validation.

In more details, the process starts by running a Norwegian NEC analysis (Johansen, 2015) on the news corpus. The list of identified entities was manually analyzed, with some algorithmic assistance (looking up names), to only select entities referring to persons. Then, using the list of Norwegian politicians, all sentences from all news articles containing the politicians' names from the seed list were identified and extracted. These names were then substituted with the placeholder-string "POL" in the sentences. This was done to create an abstract concept of politician and create more patterning in the input texts. It yields a set of sentences where known names have been removed, but permits our approach to distinguish between entities occurring in our target list (i.e. having roles similar to politicians) from other names. In the same sentences, we substituted each appearance of a person name (non-politician) entity, as identified by NEC and not present in the politician seed list, with the placeholder-codes "PER" for persons, and "ENT" for the remaining entities. The placeholders "POL", "PER", and "ENT" do not otherwise occur in the text.

Once the sentences containing known politician names are extracted, SIMMS (Salway et al., 2014) is applied in order to automatically induce salient patterns of language use around the coded politicians' names (i.e. the string "POL"). We start by creating snippets around POL as explained in (Salway et al., 2014). These snippets are of various sizes and contain between 0-12 words on either side of the string. These are used in the various iteration phases of the algorithm. The snippets are of increasingly large sizes which allows the algorithm, in each of its running iterations, to build more patterning around the already identified patterns. These patterns are information structures of word sequences, resembling information extraction templates (Salway et al., 2014).

ID.Pattern	Freq
P_0. ((justisminister|arbeidsminister|kommunalminister|helseminister| utenriksminister|finansminister|stortingsrepresentant|ENT|næringsminister |partikollega|fiskeriminister|styreleder|ENT-ordfører|ordfører |ordførerkandidat|statsminister|statsråd|samferdselsminister| statssekretær) POL) *((justice minister|labor minister|minister of local government|health minister|minister of foreign affairs|finance minister|member of parliament|ENT|industry minister|party colleague|minister of fisheries|Chairman|ENT-mayor|mayor|mayor candidate|prime minister|council of state|minister of transport|state secretary) POL)*	13912
P_14. (klima- og (miljødepartementet|miljøminister|miljøministeren| miljøvernminister|miljøvernministeren)) *(climate and (environment ministry|environment minister|the environment minister|minister of the environment|the minister of the environment))*	440
P_44. (ENT (kommunalpolitiske|justispolitiske|likestillingspolitiske| helsepolitiske|innvandringspolitiske|mediepolitisk|mediepolitiske| finanspolitiske|landbrukspolitiske|fiskeripolitiske) talsperson) *((municipality's political|justice's political|equality's political|health's political|immigration's political|media political|media's political|finance's political|agriculture's political|fisheries' political) spokesperson)*	151
P_61. ((påtroppende|kommende) byrådsleder) *((incoming|forthcoming) governing mayor)*	99
P_102. (ENT (utdanningspolitiske|helsepolitiske|finanspolitiske| fiskeripolitiske) talskvinne) *(ENT (education's political|health's political|fiscals' political|fisheries' political) spokeswoman)*	41
P_192. (sier helsepolitisk (talskvinne|talsmann)) *(says health's political (spokeswoman|spokesman))*	17

Table 2: A selection of induced structures.

A small sample of induced structures is shown in Table 2. The structures are shown in Norwegian, with English translations in italics. Structure P_0 is the most frequent of all induced structures, and seems to discuss politicians in different positions. Structure P_14 focuses on politicians related to climate and environment. Structures P_44, P_102, and P_192 discuss spokeswoman and spokesman in various political roles, while structure P_61 seem to discuss local politicians.

We automatically filtered the induced structures to keep only those that included the string "POL", or had the string in the exterior of the patterns within the matched sentences. We disregarded all structures that only included the strings "PER" or "ENT" without "POL" in their surroundings within their matched sentences. However, structures including "PER" or "ENT" and not including "POL" but having the string in the matched sentences were kept. This was done to focus our analysis on structures and sentences containing names of politicians indicated by the presence of the string "POL", and in order to identify which of the strings within "PER" are actually politicians. This process resulted in 108 patterns.

After we had discovered the patterns which seemed to identify politicians, we subsequently isolated the sentences which manifested these patterns into a sentence sub-corpus, D. We re-ran the NEC (Johansen, 2015) on this corpus to identify candidate names, which were thereafter compared to the seed list of politicians. The politicians that were present in the seed list were removed keeping only the list of new names that we believed were likely to be unknown politicians (i.e. not in the seed list). This list of unknown names was afterwards presented to human judges for manual analysis. The results of this were rather promising, as we show in the next section.

5 Results

We compute for each of the newly identified names a score representing its frequencies in the induced structures. The score is computed as the sum of all frequencies of the name in the various structures in which it appears. In Figure 1 we show the distribution of names and their frequencies. The figure on the right shows an excerpt of the most interesting range of the full range shown in the figure on the left. Most of the newly identified names have frequencies between 2 and 51, with some recurring spikes between frequency 51 and frequency 150. The frequency distribution flattens out after frequency 150 and there is only one name per frequency between frequency values 150 and 521.

Top 20 of the most frequent names and their frequencies, excluding the ones present in the seed list,

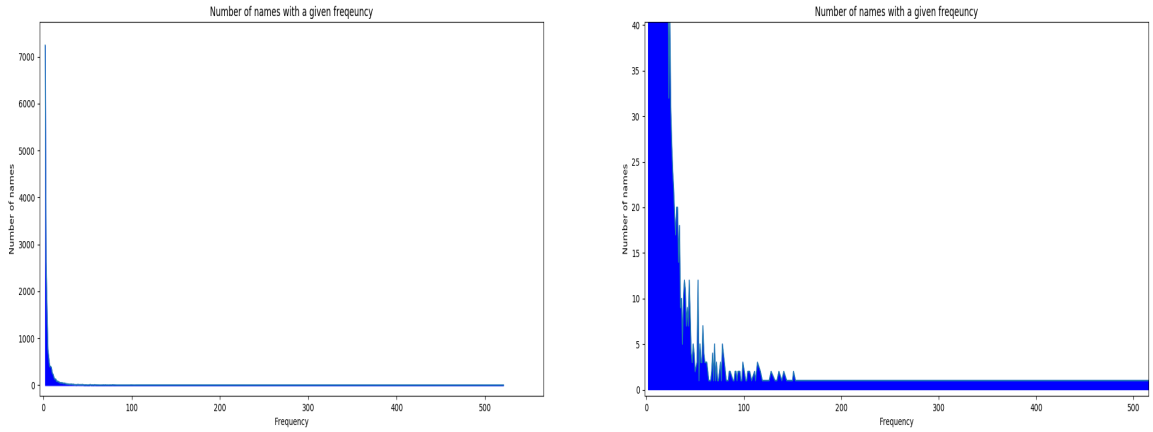

Figure 1: Number of names per frequency.

Name	Freq.	Cat.	Name	Freq.	Cat.
1. David Cameron	521	1	11. Ketil Solvik-olsen	280	1
2. Hege Storhaug	468	0	12. Monica Mæland	239	1
3. Henrik Kristoffersen	423	0	13. Kristoffer Ajer	227	0
4. Svein Aaser	410	0	14. Thorhild Widvey	218	1
5. Mads Stokkelien	375	0	15. Charles Michel	217	1
6. Benjamin Netanyahu	331	1	16. Manuel Valls	192	1
7. Ahmet Davutoglu	327	1	17. Knut Storberget	191	1
8. Angela Merkel	324	1	18. Bent Høie	180	1
9. Stefan Löfven	306	1	19. Alexander Kristoff	177	0
10. John Kerry	283	1	20. Kristoffer Barmen	172	0

Table 3: Top 20 of the most frequent names as identified by our approach, and their manually assigned categories (1: politician, 0: not politician).

are presented in Table 3. We also show which names are politicians using boolean values: 1 for politician and 0 for not politician. This decision was based on the manual analysis of the human judges.

We investigated two top-scoring intervals: a set of the best scoring names, and a subsequent set of best scoring names. The set of the best scoring names are all names that have a frequency equal or greater than 90% of the frequency of the most frequent name. The second top-scoring interval represents names with frequencies between 50 and 32. They yielded respectively 156/264 (60%) and 97/231 (42%) politicians including known politicians from the seed list. We argue that this already corroborates our assumptions about specialized language, that persons sharing the same role are discussed and talked about similarly.

Disregarding known politicians, the first set of best scoring names resulted in a sample of 167 unknown names, with frequencies spanning from 521 at maximum to minimum 51. We manually analyzed this sample of names and identified which ones were actually politicians. From these 167 names, we were not able to classify 6 names either because they were incomplete, or ambiguous. From the 161 names remaining, 33% were manually classified as politicians (53 names out of 161). The remaining 67% mostly represented athletes. The second best scoring interval of names resulted in a set of 170 unknown names, from which 8 were unclassified. From the remaining 162 we were able to classify 28 names as politicians, representing 17.28%. As in the previous interval, most of the remaining names referred to athletes. That is, had we presented the unrecognized top-scoring names, a human expert would verify every third name as a politician which we could correctly add to our seed list of politicians. Table 4 summarizes our findings on unknown politician names.

A total of 67% of the newly identified politicians, in the first set of best scoring names, represented Norwegian politician names (local and national), and 33% were international politicians (Canada, UK,

Results	Set of best scoring names	Subsequent set of best scoring names
#unknown names	167	170
#incomplete/ambiguous names	6	8
#politicians	53 (33%)	28 (17.28%)
#Norwegian politicians	36 (67%)	23 (82.15%)
#international politicians	17 (33%)	5 (17.85%)

Table 4: Number of politicians (Norwegian and international) identified by our method, in both investigated top-scoring intervals.

Germany, USA, Sweden, France, Spain, Russia, Palestine, Poland, Croatia, Italy, Israel, Turkey, Denmark, Belgium). In the remaining names of the first set of best scoring names, 59% referred to football players, football coaches, general managers of football clubs, or football commentators. About 9% referred to athletes in other sports like cycling, and handball. A total of 12% were person names known in winter sports like skiing, ice hockey players, and alpinists. The remaining 20% were political activists, journalists, police, jurists, lawyers, celebrities, union leaders, musicians, and company leaders.

In the second set of best scoring names, a total of 82.15% of the newly identified politician names were Norwegian politicians, and 17.85% were international ones (from Canada, Germany, and UK). The person names that were not politicians were mostly related to sports. These were football players/coaches, other sports like handball, ski or ice hockey athletes and represent 50.31% of the 161 identified names. The remaining 16.69% were political activists, journalists, celebrities, union leaders, musicians, and company leaders.

6 Conclusion and future work

In this paper, we have addressed the problem of automatically identifying person names sharing similar political roles from a corpus of news articles. We have used a text mining approach able to automatically induce patterns of language use around known politician names from a predefined seed list. This approach is able to group together words appearing in similar contexts, and hence person names sharing similar roles.

Automatically identifying and extracting names is a difficult task, especially in Norwegian, as no off-the shelf named entity recognition (NER) approaches exists. In this work, we have used a named entity chunker (NEC) (Johansen, 2015) that is able to identify entities, but not their types. In order to be able to differentiate person names from other types of entities, we in part relied on computer assisted manual analysis. We have shown that our approach is able to identify new names that naturally extend the seed list, as these new names have similar roles.

There are many parameters that may influence this approach. If our seed list is actually complete, there would be no names to add and our approach would only present names not belonging to it. If our seed list is heterogeneous, or contains names belonging to a group which is generally not discussed by a specialized language, we may expect the human expert to be presented with many names which do not naturally extend our list.

Although we have only tested our approach on a list of politicians' names, we feel there is good reason to believe that many other natural sets of names satisfy the conditions our approach relies on. It suffices that the entities named in the seed list share some attributes or characteristics which makes it likely that they will be discussed in similar ways in some specialized language.

Most of the names identified were not politicians, but a relatively high proportion were. We removed all known politicians before compiling the results presented in this article. This justifies the fact that 33% is regarded as good results for a first step. We believe that this might be due to the nature of the corpus. The corpus was a collection of news articles covering all newsworthy issues from sport to politics. Based on the induced structures, it seems that the way politicians are talked about, the words used to describe them or discuss issues around them, resembles the way athletes and sports related events are talked about.

We plan to improve our approach by first running a topic modeling approach on the corpus, to filter out

the news articles not covering political issues, and then re-run the presented approach. This will give, we believe, a more focused corpus as input, and we might be able to solely identify new names of politicians. We also think that it would be interesting to use word embeddings, and investigate if politicians' names are nearer in the vector space model as opposed to non-politicians.

References

Mikhail A. Alexseev and W. Lance Bennett. 1995. For whom the gates open: News reporting and government source patterns in the united states, great britain, and russia. *Political communication*, 12(4):395–412.

Sigurd Allern. 1996. *Kildenes makt: Ytringsfrihetens politiske økonomi*. Pax.

C. Edwin Baker. 2006. *Media concentration and democracy: Why ownership matters*. Cambridge University Press.

Frank R. Baumgartner and Laura Chaqués Bonafont. 2015. All news is bad news: Newspaper coverage of political parties in spain. *Political Communication*, 32(2):268–291.

W. Lance Bennett. 1990. Toward a theory of press-state relations in the united states. *Journal of communication*, 40(2):103–127.

Dan Berkowitz and Douglas W. Beach. 1993. News sources and news context: The effect of routine news, conflict and proximity. *Journalism Quarterly*, 70(1):4–12.

Jane Delano Brown, Carl R. Bybee, Stanley T. Wearden, and Dulcie Murdock Straughan. 1987. Invisible power: Newspaper news sources and the limits of diversity. *Journalism Quarterly*, 64(1):45–54.

Timothy E. Cook. 1998. *Governing with the news: The news media as a political institution*. University of Chicago Press.

Tine Ustad Figenschou and Audun Beyer. 2014. Elites, minorities and the media-primary definers in the norwegian immigration debate. *Tidsskrift for Samfunnsforskning*, 55(1):23–51.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Zellig S. Harris. 1988. *Language and information*. Columbia University Press.

Leo W. Jeffres, Connie Cutietta, Leslie Sekerka, and Jae-won Lee. 2000. Newspapers, pluralism, and diversity in an urban context. *Mass Communication & Society*, 3(2-3):157–184.

Bjarte Johansen. 2015. Named-entity chunking for norwegian text using support vector machines. In *Norsk Informatikkonferanse (NIK)*, Ålesund, Norway.

Mariska Kleemans, Gabi Schaap, and Liesbeth Hermans. 2017. Citizen sources in the news: Above and beyond the vox pop? *Journalism*, 18(4):464–481.

Oslo: Kulturdepartementet. 2017. Det norske mediemangfoldet – en styrket mediepolitikk for borgerne. In *Norwegian media diversity: A strengthened media policy for citizens*.

Sydney M. Lamb. 1961. On the mechanization of syntactic analysis. *Conference on Machine Translation of Languages and Applied Language Analysis II*, pages 674–685.

Maxwell McCombs. 2005. A look at agenda-setting: Past, present and future. *Journalism studies*, 6(4):543–557.

Denis McQuail. 1992. *Media performance: Mass communication and the public interest*. Sage.

Karen Ross, Elizabeth Evans, Lisa Harrison, Mary Shears, and Khursheed Wadia. 2013. The gender of news and news of gender: a study of sex, politics, and press coverage of the 2010 british general election. *The International Journal of Press/Politics*, 18(1):3–20.

Karen Ross. 2007. The journalist, the housewife, the citizen and the press: Women and men as sources in local news narratives. *Journalism*, 8(4):449–473.

Andrew Salway and Samia Touileb. 2014. Applying grammar induction to text mining. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 712–717, Baltimore, USA.

Andrew Salway, Samia Touileb, and Endre Tvinnereim. 2014. Inducing information structures for data-driven text analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 28–32, Baltimore, USA.

Adam Shehata. 2010. Marking journalistic independence: Official dominance and the rule of product substitution in swedish press coverage. *European Journal of Communication*, 25(2):123–137.

Eli Skogerbø and Arne H. Krumsvik. 2015. Newspapers, facebook and twitter: Intermedial agenda setting in local election campaigns. *Journalism Practice*, 9(3):350–366.

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

Bartholomew H. Sparrow. 1999. *Uncertain guardians: The news media as a political institution*. JHU Press.

Samia Touileb and Andrew Salway. 2014. Constructions: a new unit of analysis for corpus-based discourse analysis. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 634–643, Phuket, Thailand.

Gadi Wolfsfeld and Tamir Sheafer. 2006. Competing actors and the construction of political news: The contest over waves in israel. *Political Communication*, 23(3):333–354.