Quote Extraction and Attribution from Norwegian Newspapers

Andrew Salway, Paul Meurer, Knut Hofland and Øystein Reigem
Language and Language Technology Group
Uni Research, Bergen, Norway
firstname.lastname@uni.no

Abstract

We present ongoing work that, for the first time, seeks to extract and attribute politicians' quotations from Norwegian Bokmål newspapers. Our method – using a statistical dependency parser, a few regular expressions and a look-up table – gives modest recall (a best of .570) but very high precision (.978) and attribution accuracy (.987) for a restricted set of speaker names. We suggest that this is already sufficient to support some kinds of important social science research, but also identify ways in which performance could be improved.

1 Introduction

Social science researchers are increasingly incorporating automatic text analysis techniques into their research methods in order to exploit large corpora, such as newspaper articles, social media posts and political speeches. To date, it has mostly been bag-of-words techniques, such as topic modelling, that have been used and, as such, the text is normally the basic unit of analysis (Grimmer and Stewart, 2013).

For further progress, there is a need to be able to recognize other units of analysis within texts, such as quotations attributed to their speakers in newspaper articles. The ways in which news media select and present the reported speech of politicians is important for the functioning of democratic societies. Given a data set comprising who is reported to have said what, and when, social science researchers could study the opinions of political leaders on key issues and how they relate to those of the electorate (representative democracy), and how different newspapers present the views of different politicians (media coverage and framing). This data would also be relevant for citizens and journalists to keep track of what a politician has said about a certain issue and how this changes over time. Our aim is to develop a method to generate such data sets from Norwegian Bokmål newspapers.

Quote extraction and attribution have been well studied for English-language newspapers (Krestel et al., 2008; O'Keefe et al., 2012; Pareti et al., 2013). That research highlighted how the ways in which quotes are presented, and hence the challenges for extraction and attribution, vary quite markedly between different varieties of English and even different newspapers. Hence it makes sense for us to look for a Norwegian-specific solution. Also, in contrast to previous work, our focus on generating data sets for social science researchers leads us to prioritize very high precision, at the cost of recall if necessary, so long as there is no systematic bias in how recall fails (Section 5 says more on this).

This paper presents ongoing work to extract and attribute politicians' quotes from Norwegian Bokmål newspaper stories. Section 2 gives a more precise task definition and describes the creation of a testing set of newspaper stories in which quotations are annotated. Section 3 describes the method for extraction and attribution that we have implemented so far. Section 4 presents an evaluation and discusses remaining challenges, and Section 5 makes some tentative conclusions.

2 Task Definition and Testing Set

Following O'Keefe et al. (2012), we define quote extraction to be the task of identifying continuous spans of direct speech – an instance of a speaker's words being presented exactly as they were spoken, and of indirect speech – an instance of a speaker's words being reported, but

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 293–297,
Gothenburg, Sweden, 23-24 May 2017.

not as direct speech. The span of a quote may go across sentence and paragraph boundaries. In cases in which direct quotes are embedded within indirect quotes, we take the whole span of the indirect quote, with embedded direct quotes, to be one quote. Quote attribution is then the task of associating text spans with speakers.

Direct speech is marked with pairs of quotations marks, and, in Bokmål newspapers at least, more commonly with a preceding long dash which leaves the end of the quote less explicitly marked. For English-language newspapers, O'Keefe et al. (2012) reported over 99% accuracy for extracting direct quotes by searching for text between quotation marks. Such a straightforward solution is not possible for Bokmål newspapers, in part because of the dashed quotes, and also because we observed the frequent use of quotation marks around non-speech text spans, especially to highlight words and terms, and also for film and book titles.

Determining which text spans should be considered as indirect reported speech is somewhat problematic. Indirect speech implies that the writer is, to some extent, filtering and/or interpreting the spoken words. However, judging how much filtering and interpreting can happen before it is no longer reported speech can be hard. On the one hand, it may be defined in syntactic terms, e.g. a change from a first person to a third person pronoun used to refer to the speaker, and a change from absolute to relative tense, but these criteria do not cover all cases. On the other hand, although it may be harder, it also may be more appropriate to define indirect speech in semantic or pragmatic terms, e.g. according to different categories of verbs, or the attitude that the writer is perceived to express towards the proposition. These criteria leave fuzzy boundaries, for example between a clear case of indirect reported speech such as *'Person X said that Proposition'* and something like *'Person X thinks that Proposition'*: in the latter, it is hard to know how much the writer has interpreted the spoken words. See Pareti et al. (2013), and references therein, for a comprehensive treatment of these issues.

To create annotated material for evaluation purposes, two of the authors were responsible for annotating quotes in a sample of Bokmål newspaper texts. They first worked independently on different parts of the material, and then inspected, discussed and jointly edited each other's annotations. Although the aim was to create a gold standard, at this stage in our work we are not yet confident about a few cases of indirect speech, for reasons mentioned above.

Quote attribution requires resolution of pronouns and other nominal references. In our work to date, with a focus on social science applications, we have simplified this situation by defining a closed set of speakers comprising 99 Norwegian politicians, subjectively selected for their prominence from lists of governments over the past 20 years. This made it feasible for our method to include a manually compiled look-up table with gender information and alternative forms of full names (which may vary over time, e.g. after marriage). We have not yet attempted to resolve nominal references such as *'the trade minister'*, which are time-sensitive, but these are annotated in the test set and hence have a negative impact on recall in evaluation (this challenge is discussed in Section 4).

The sample of Bokmål texts was taken from a Norwegian monitor corpus of newspaper texts – Aviskorpus (Andersen and Hofland, 2012). Having retrieved all texts which contained a politician name, a speech verb and a nearby quotation mark, we selected every 220th text for the sample. For this, a list of 64 speech verbs was compiled from a data-driven analysis of the co-texts of politicians' names in newspaper articles, and extended with synonyms. It should be noted that it appeared that a very few verbs account for the vast majority of reported speech, i.e. *'si' ('to say')*, *'mene' ('to think')* and *'kalle' ('to call')*.

After manually removing articles in which the mentioned politicians were not speakers this gave 162 texts from 10 newspapers for 2001-2016; most appeared to be regular news stories of 600 words or less. A total of 1031 quotes were annotated comprising 690 instances of direct speech and 341 of indirect speech. This ratio is different from the 50:50 estimated for English-language newspapers (Pareti et al. 2013). We also note that the majority of direct quotes – 630 out of 690 – appear with a preceding dash, rather than in a pair of quotation marks.

3 Method

The main idea is to take the subject and the complement of speech verbs as speaker and reported speech respectively. Two extensions to improve recall were conceived, in part from our initial corpus-based investigations into how quotations are expressed, and in part from analysis of the annotated data (Section 2), which was later used for evaluation. Thus, there may be

potential for a kind of "overfitting", but the extensions comprise a few simple heuristics which we believe are generally applicable.

For each text that contains the full version of a politician's name, each sentence is analyzed by a statistical dependency parser for Norwegian Bokmål that was developed, in other ongoing work, using the Stanford Neural Network parser framework (Chen and Manning, 2014). It was trained with a dependency treebank that was derived from the INESS NorGramBank gold standard treebank, i.e. the subset of manually disambiguated and checked parses (Dyvik et al., 2016). The word embeddings were trained on 1.58 billion tokens from news, parliamentary and popular science texts.

If an inflected form of a speech verb is detected, then its grammatical subject and complement are extracted. Only speech verbs that subcategorize for a sentence complement were considered; a list was formed through a search for verbs allowing a sentence complement in the NorGram grammar lexicon. In practice, this meant that the vast majority of detected instances were for the verbs *'si' ('to say'*) and *'mene'* (*'to think'*). The subject is taken to be the speaker and the complement is taken to be the reported speech. We do not explicitly distinguish direct and indirect speech, but note that complements with a complementizer are mostly indirect speech.

In the simplest case the subject is one of the politicians' full names. In the case that it is a surname, or another variant of their name, then the full name is looked up, and if the full name is mentioned somewhere in the current story then this is taken to be the speaker. If the subject is a third person singular pronoun (*'han', 'hun'*) then the pronoun is resolved to the most recent politician's name if it has the correct gender.

The prevalence of dashed quotes requires an extension to the core method. Sometimes these are contained within a single sentence that starts with a long dash, followed by direct speech, followed by a speech verb and speaker. However, there are also cases where the direct speech continues over several sentences: only the first sentence starts with a dash, and the verb and speaker come only in the final sentence. Thus, in those cases where the complement comes before the speech verb, we use a simple regular expression to check whether the current sentence, or any preceding sentence in the paragraph, starts with a dash. If so, then the text span from the dash to the end of the complement is taken to be the quote.

This extension to deal with dashed quotes was refined to deal with dialogs such as when a journalist is interviewing a politician. Here dashed quotes typically follow each other with alternate quotes coming from the politician but the politician may only be mentioned with a speech verb near the final quote. So, once a dashed quote is found we look backwards for a sequence of dashed quotes in which alternating quotes end with a question mark, and then attribute every other quote to the politician who is attributed to the final quote.

A second extension was implemented to deal with some of the cases in which the parser failed to find either the subject or the complement of a speech verb. Each sentence is tested for a simple pattern comprising a comma followed by a speech verb and a personal pronoun or a politician's name within three tokens to the right. If this pattern matches, then the text span from the start of the sentence to the comma is taken to be a quote.

4 Evaluation

The quote extraction performance of the core method, and of the two extensions, was evaluated on the basis of the testing set (Section 2) with measures of recall and precision. Here recall is the proportion of the quotes in the testing set that were extracted by the method; either wholly or at least partially. Precision is the proportion of the extracted quotes that were actually quotes according to the testing set; either wholly, or at least partially.

The results for quote extraction are presented in Table 1, with a best recall of .570 and a best precision of .978. The need for the extension to deal with dashed quotes that go over multiple sentences is highlighted by the increase from .246 to .409 for recall of whole quotes, whilst there is little change in the value for recall of at least partially extracted quotes (.503 to .509). Adding one pattern to capture quotes that were missed by the parser gives a useful increase in recall (.409 to .469, for whole quotes), at the expense of some precision (.974 to .951).

The performance for quote attribution was measured as attribution accuracy, i.e. the proportion of the extracted actual quotes that were attributed to the correct speaker. For 'parse + dashed quotes', 519 out of 526 quotes were correctly attributed, giving an accuracy of 0.989.

It is not surprising that attribution accuracy is very high because a quote will only be extracted if one of the politician's names is found as its subject, or as the resolution of a pronoun.

Method	Recall whole *(partial)*	Precision whole *(partial)*
Parse only	254/1031 = .**246** *519/1031 = .**503***	518/531 = .**976** *519/531 = .**977***
Parse + dashed quotes	422/1031 = .**409** *525/1031 = .**509***	523/537 = .**974** *525/537 = .**978***
P. + d.q. + simple pattern	484/1031 = .**469** *588/1031 = .**570***	583/613 = .**951** *588/613 = .**959***

Table 1: Evaluation of quote extraction.

Consideration of the quotes that were not recalled suggests several ways in which the method could be extended. It seems the most common problem is the occurrence of nominal references such as *'handelsministeren'* (*'the trade minister'*) in subject position. It would be expensive to create a knowledge-base of which politician held which role at which time. Our next step will be to evaluate a simple look around method, similar to how we resolve pronouns and to the baseline method described by O'Keefe et al. (2012). Beyond that, we may look to connect nominal references and politicians with evidence from the text, e.g. introductions like *'the trade minister, NAME …'*.

Recall could also be improved by capturing more ways in which quotes are signaled. Firstly, we may extend the method to look for more kinds of constructions such as: (i) extrapositions in which a quote is split around the verb phrase, e.g. *'[Q part 1] sa Politician at [Q part 2]'* (*'[Q part 1] said Politician [Q part 2]'*); (ii) particle verbs, e.g., *'legge til' ('to add'*); and, (iii) other constructions that can take sentential complements, e.g. *'ifølge'* (*'according to'*).

Of course, however many constructions are added, the performance of the parser will still be a limit on extraction and attribution results. For example, we think that problems with the parsing of co-ordination could be significant for explaining the difference between partial recall (.509) and whole recall (.409). The parser has a labeled attachment score of about 89% which, although it is quite close to the state-of-art, means that 1 in 10 attachments will be incorrect. More could be done to catch the cases where parsing fails, i.e. by using more regular expressions to match known patterns, although a fall in precision would be expected. If recall was to be prioritized over precision, or quotes for an unrestricted set of speakers were needed, then the use of machine learning with a large set of diverse features should also be considered. However, even then the best recall achieved in the literature (for English) is 0.54 (P=0.66) for whole quotes and 0.74 (P=0.79) for at least partial quotes (Pareti et al., 2013).

5 Conclusions

This paper initiated work on quote extraction and attribution from Norwegian Bokmål newspapers, with the aim of creating data sets to support social science research in areas such as democratic representation, media coverage and media framing. The creation of a testing set of annotated quotations may be seen as a step towards a gold standard to be used as the basis for future work, but tricky issues remain around the definition and annotation of indirect quotes.

Our method for extraction and attribution addresses some of the characteristics of Bokmål quotations (more direct speech and the common use of dash quotes). We have identified ways to improve on our method, particularly for recall, but it can be argued that the levels of recall, precision and attribution accuracy that were achieved already may be sufficient for some social science research.

Having high levels of precision and attribution accuracy means that researchers can trust that almost all of the extracted text spans are quotes and that they are attributed to the correct politician. It seems likely that very high precision would be a prerequisite for using the data in social science research, unless it was to be checked manually; note, we estimate many 10,000's quotes for 99 politicians in Aviskorpus. Conversely, it seems to us that a modest recall value (c. 0.5) would be acceptable if the set of quotes is considered to be a good sample, i.e. if there is no bias towards particular newspapers or politicians when the method fails to extract quotes. Whilst we cannot see any way in which the method is biased towards certain politicians (so long as the data in the look-up table is accurate), it is possible that the idiosyncratic style of some newspapers could have an impact, and this must be investigated.

Acknowledgments

We thank the anonymous reviewers for their insightful and constructive input.

References

Gisle Andersen and Knut Hofland. 2012. Building a large corpus based on newspapers from the web. In: Gisle Andersen (ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*: 1-30. John Benjamins.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. *Procs. 2014 Conference on Empirical Methods in Natural Language Processing*: 740-750.

Helge Dyvik, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes. 2016. NorGramBank: A 'Deep' Treebank for Norwegian. *Procs. 10th International Conference on Language Resources and Evaluation, LREC 2016*: 3555-3562.

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.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *Procs. 6th International Language Resources and Evaluation Conference, LREC 2008*.

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A Sequence Labelling Approach to Quote Attribution. *Procs. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: 790-799.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. *Procs. 2013 Conference on Empirical Methods in Natural Language Processing*: 989-999.