uni Research

Mining “who said what, and when” in the Norwegian newspaper corpus

Andrew Salway, Knut Hofland, Øystein Reigem

Language and Language Technology Group

Centre for Big Data Analysis

Motivation

- Data about the reported speech of politicians in newspapers could be used to investigate:
  - How politicians' opinions align with public opinion
  - How different newspapers cover different politicians
  - Evidentiality in newspapers and new media (Krestel, Bergler and Witte 2008)
  - What politicians say about each other
  - …???
- Data could also be interesting for citizens and journalists

Task Definition

- **Quote extraction**: identify the text fragments that are reported speech
- **Quote attribution**: identify the speaker

- So, for each quote we should have:
  - The text of the quote, and the topic(?)
  - Who the speaker is (and metadata about the person, e.g. political party)
  - Metadata of the newspaper story, inc. date, newspaper, journalist(?)

Example input

DET SER UT TILat det samme gjelder polakker selv om lønnsforskjellene mellom Polen og Norge er store. Finansministerens bekymring er at det kan bli for få polakker i Norge. **"Jeg frykter at når den økonomiske situasjonen nå bedrer seg i Polen og Sverige, vil mange ta seg jobb i hjemlandet i stedet"**, sier **Halvorsen**. Tilgangen på importert arbeidskraft gjør norsk økonomi mer fleksibel og sikrer at høy økonomisk vekst ikke presser lønninger og rente i været. Indirekte gir hun argumenter for at det nye tjenestedirektivet i EU, som vil gjøre det lettere å tilby tjenester og arbeidskraft over landegrensene, vil gi Norge mange fordeler. Kristin Halvorsen fortjener ros når hun så åpent forteller om den nye virkeligheten hun møter som finansminister.

Aftenposten, PUBLISERT: 21.FEB.2006

http://www.aftenposten.no/meninger/leder/Hjelpen-fra-Polen-417238b.html

Example output

Politcian: Kristin Halvorsen

Party: Sosialistisk Venstreparti

Newspaper: Aftenposten

Date: 21-02-2006

Quote: Jeg frykter at når den økonomiske situasjonen nå bedrer seg i Polen og Sverige, vil mange ta seg jobb i hjemlandet i stedet

Topic: employment; immigration

Challenges

- Various ways to indicate direct quotes, and other uses of quotation marks.
- Indirect quotes: no previous work for Norwegian, but for English direct quotes are relatively rare; most reported speech is indirect (O'Keefe et al. 2012). Taking the syntactic object of speech verbs gives high precision for indirect quotes, but low recall (Pareti et al. 2013).
- A complete solution will require named entity recognition (changes over time!), and pronoun resolution.
- Classifying the topic of short text fragments could be difficult.

A simple first approach

Extract and attribute direct quotes for a small set of politician names in Aviskorpus (3.4m Norwegian newspaper articles):

- Apply the baseline method in O'Keefe et al. (2012), modified for a fixed set of names
- Evaluate and assess what more is needed

Method: quote extraction

Take all text between each pair of quotation marks («», “”), including across sentences and paragraphs:

- Only look for quotes if the story contains a politician’s whole name
- Each part of a split quote gets treated as a separate quote
- Add a test to keep track of start and end quotes, and reset if necessary

Method: quote attribution

1. Search backwards in the text from the end of the sentence that the quote appears in for a reported speech verb; the verb could be in a previous sentence.

2. If a verb is found take the name** mentioned nearest it; distance is measured by number of words in either direction.

3. If no verb is found then take the name nearest the quote.

**A name has to be one of a fixed set of politicians, otherwise we ignore the quote. So, if another name-like thing is found first we ignore the quote. Name-like means capitalised words except sentence initial words and some exclusion words.

Method: resources

- List of politician names, with a canonical form, unambiguous forms (versions of the whole name) and alternative forms (e.g. surname only), and with their party
- List of reported speech verbs (base and inflected forms)

Results

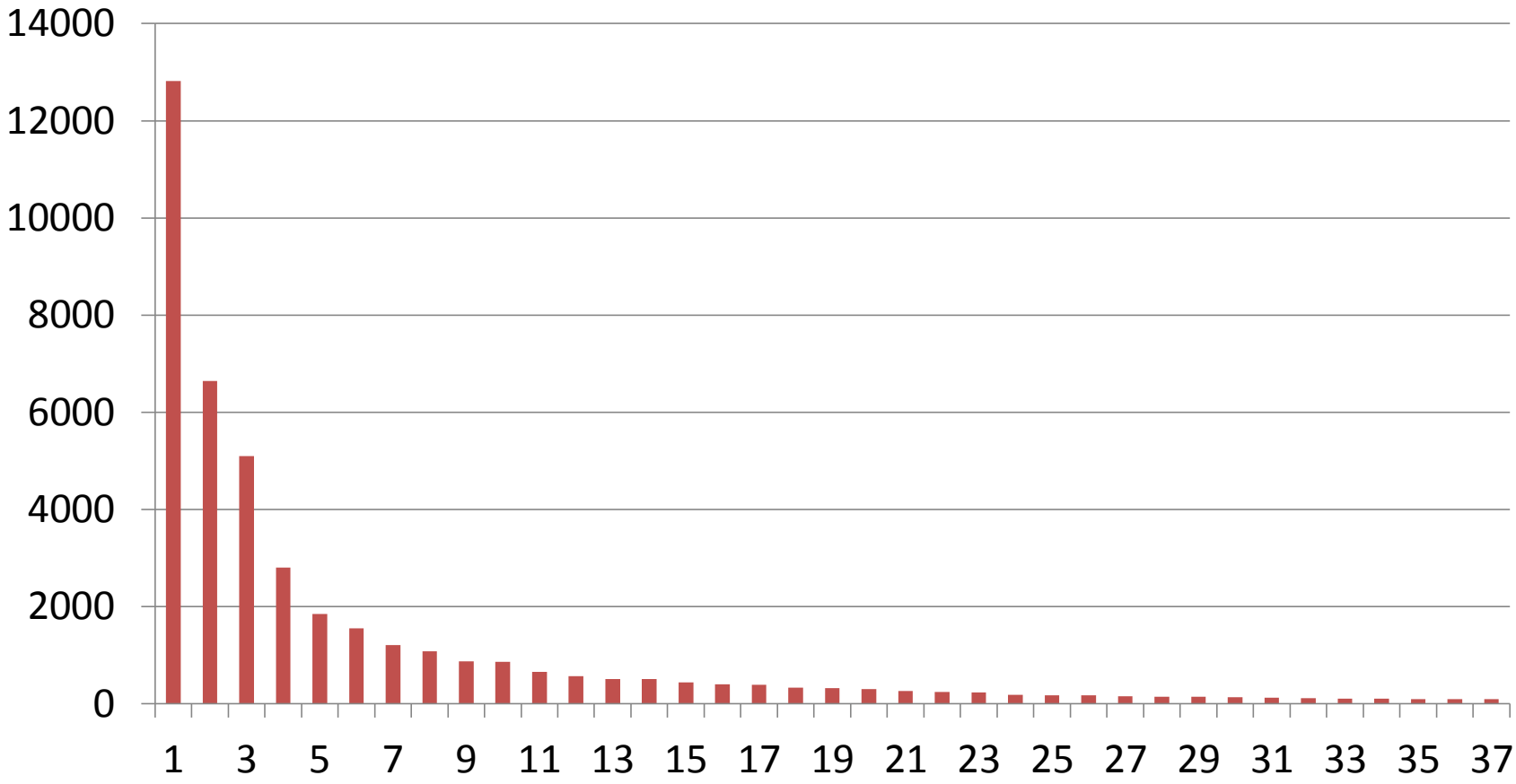
- From 3.4m newspaper articles: 45,173 text fragments identified as direct quotes in about 29,000 different articles
  - Only a small set of about 50 politician names, but probably most of the common ones
  - Recall, indirect quotes are probably much more common

Preliminary evaluation: 100 quotes

Quote is reported speech	Attributed correctly	Correct verb found	**Frequency**
Yes	Yes	Yes	**46**
Yes	Yes	No	**3**
Yes	No	Yes	**4**
Yes	No	No	**20**
No	-	-	**27**

- 27% of extracted text fragments are not quotes: (i) boilerplate/html in corpus; (ii) other uses of quotation marks, e.g. titles
- 67% of actual quotes are attributed correctly
- Also need to look at recall

Number of quotes of length n



Top 10 politicians

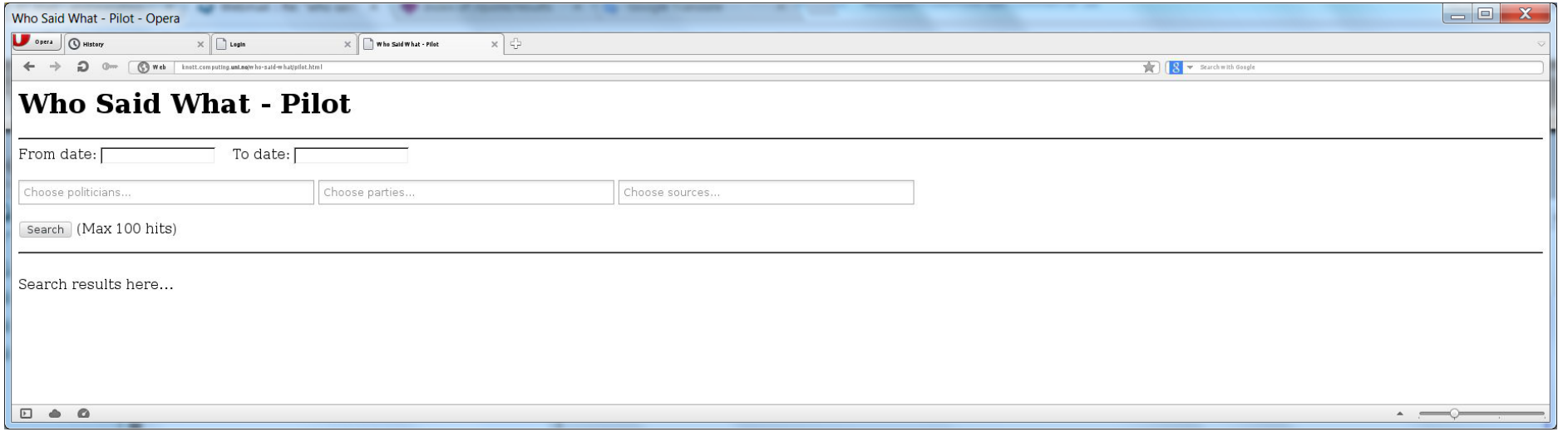
	Number of direct quotes
Jens Stoltenberg	4594
Erna Solberg	2780
Siv Jensen	2589
Carl I . Hagen	1896
Trond Giske	1853
Kristin Halvorsen	1661
Kjell Magne Bondevik	1263
Per Sandberg	1134
Lars Sponheim	1028
Jonas Gahr Stør	1008

Top 10 reported speech verbs

Inflected form	**Frequency**
sier	12725
sa	5675
mener	3531
skriver	2823
kaller	1362
skrev	753
tror	614
uttalte	570
si	567
lover	552

Base form	***Frequency***
si	*19455*
mene	*4105*
skrive	*3773*
kalle	*2192*
uttale	*944*
tro	*892*
love	*872*
snakke	*775*
fortelle	*629*
hevde	*610*

Search interface: under development



References

Krestel R., S. Bergler and R. Witte (2008). **Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles**. Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008.

O'Keefe et al. (2012). **A Sequence Labelling Approach to Quote Attribution**. Procs. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 790–799.

Pareti et al. (2013). **Automatically Detecting and Attributing Indirect Quotations**. Procs. 2013 Conference on Empirical Methods in Natural Language Processing, pp. 989–999.

Discussion points

- Is this data useful for social science research? Any past work using such data? How should the data be made accessible for analysis?
- How good does the data need to be in order to be useful? Is exact quote extraction important?
- How to improve precision and recall with more linguistic analysis and language technology?
  - O'Keefe et al. (2012) show some improvement using a large set of features for quote attribution
  - Pareti et al. (2013) show some progress with extraction of indirect speech using a mix of linguistic features
  - Need for more refined linguistic analysis, e.g. to distinguish assertions and beliefs (cf. Pareti et al. 2013)