NARC - Norwegian Anaphora Resolution Corpus

Petter Mæhlum, Dag Haug, Tollef Jørgensen, Andre Kåsen, Anders Nøklestad, Egil Rønningstad, Per Erik Solberg, Erik Velldal, and Lilja Øvrelid

University of Oslo, Norwegian University of Science and Technology, National Library of Norway

October 16th, 2022, CRAC, COLING

Introduction

- First publicly available anaphora resolution corpus for Norwegian
- Final version will contain both Norwegian written standards: Bokmål and Nynorsk
- Currently contains Bokmål, but the Nynorsk part is close to being finalized
- We will present details regarding annotation, guidelines and inter-annotator agreement
- We also present some preliminary modelling results.

Related work

- A large number of English corpora, such as MUC (Grishman and Sundheim, 1996), Ontonotes (Weischedel et al., 2011) and ARRAU (Uryupina et al, 2020)
- One earlier Norwegian corpus, BREDT (Borthen et al., 2007) was created, but it is not openly available.
- However, the guidelines are available.

Data source

- Texts taken from the Norwegian Dependency Treebank (NDT)
- Roughly 300 000 tokens for both Bokmål and Nynorsk
- 85% news
- also contains government reports, parliamentary transcripts and blog data.

Bokmål and Nynorsk?

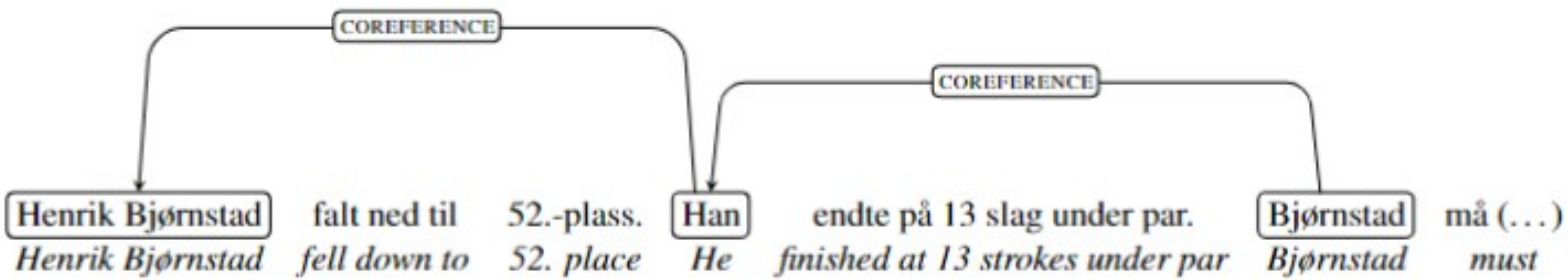
- The two oficcial written forms of the Norwegian language
- Bokmål is based on Danish, while Nynorsk is based on certain dialects.
- All Norwegians study both in school, with some exceptions
- A certain percentage of both is required in all governmental institutions
- It is important that NLP resources for Norwegian cover both.

Anaphora in NARC

- We use three different labels
  - Coreference
  - Split-antecedent
  - Bridging
- Defined from one *markable* to another. A markable is in our case a noun phrase or a determiner, such as *min* 'my'. The expression that needs to be resolved is called the anaphoric expression, and the other markable is called the antecedent.
- We separate between anaphoric and cataphoric relations.

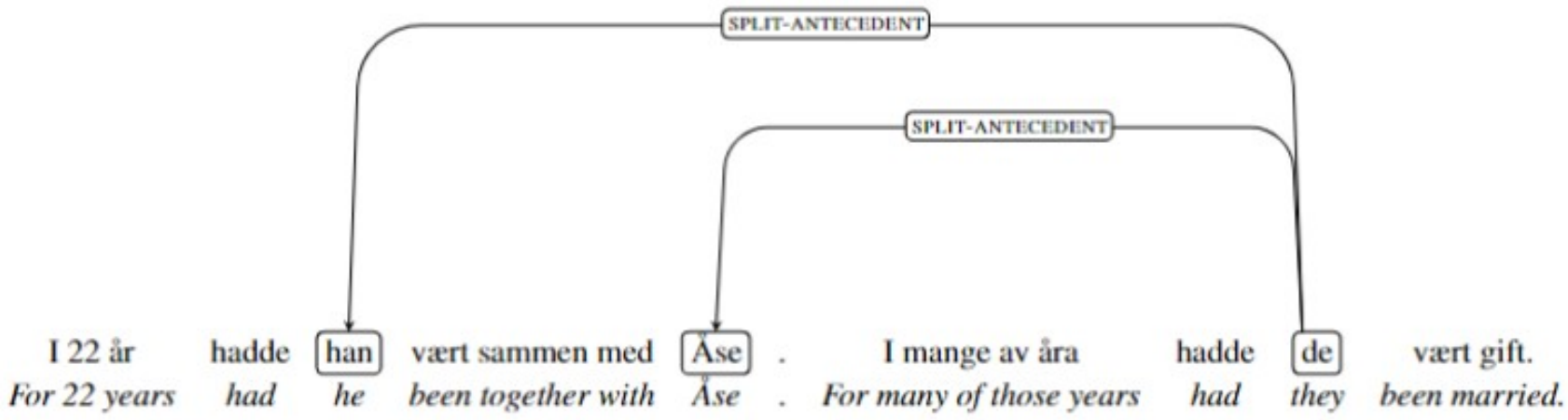
Coreference

- We recognize two broad categories of coreference:
  - Anaphors
  - Repeated coreferring entities
- Anaphors need to be resolved to an antecedent to be interpreted.
- Repeated coreferring entities are markables such as proper names and first and second person pronouns, which are not inherently anaphoric, but can corefer with earlier markables.



Split-antecedent

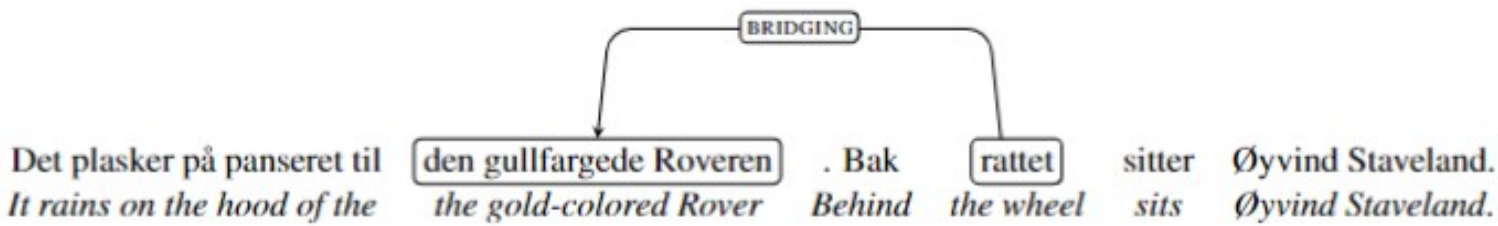
- The split-antecedent label is used when two or more non-coordinated antecedents are referred to by a single anaphoric expression.



- Coordinated expressions are treated as single markables.
  - **Kari og Ola** gikk bortover veien. **De** var sultne.
  - ***Kari and Ola*** *walked along the road.* ***They*** *were hungry.*

Bridging

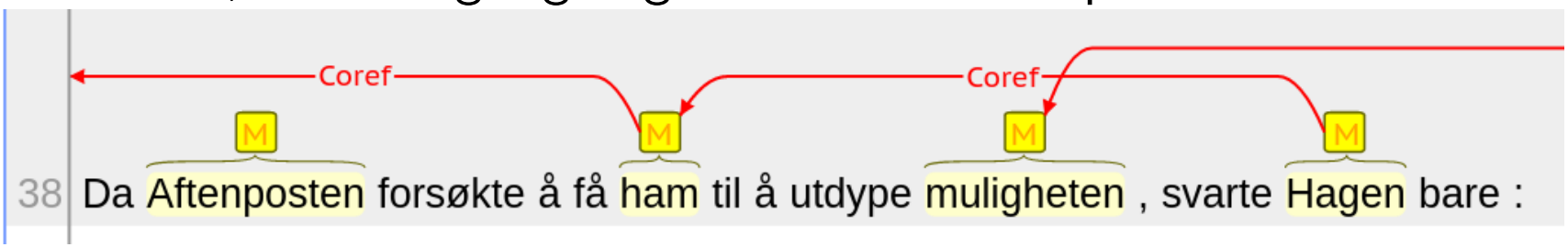
- Bridging indicates an anaphoric relation between two markables that are not coreferent, but associated in such a way that the correct identification of the anaphoric referent requires that the hearer establishes the relation to the antecedent



We do not identify sub-types of bridging.

Pre-annotation

- All texts were pre-annotated
- The syntactic information from NDT was used to identify possible markables, including noun phrases and some determiners including possessive pronouns and quantifiers such as *alle*.
- Not perfect. Some cases, such as the reflexive pronoun *seg* remained unmarked.
- Goal is to minimize errors based on syntactic delimitation
- We observe that on average only 2 markables were changed per document, indicating high agreement with the pre-annotations.



Annotation process

- Mainly 6 annotators worked on the Bokmål part of the corpus, and 4 on the Nynorsk part, with some overlap.
- Weekly meetings with annotators to update guidelines and resolve issues.
- Annotation performed using BRAT(Stenetorp et al., 2012)

Review and Curation

- All documents were re-annotated in one of two ways:
  - Curation for files used for inter-annotator agreement
  - Review for files annotated by a single annotator

Inter-annotator Agreement

- 59 documents
- Divided into 5 groups of 10 and one group of 9
- All annotators annotated at least one group, but some annotated more

	Overall F_1	**Anaphor κ**	**Cataphor κ**	**Coref. κ**	**Bridging κ**	**Split Ant. κ**
Scores	0.83	0.82	0.80	0.84	0.44	0.66

Corpus statistics 1

Type	Value
Documents	326
Sentences	15125
Tokens	231363
Total markables	6979
Used markables	26005
Singletons	43788
Single word markables	34
Discontinuous markables	499
COREFERENCE relations	19420
BRIDGING relations	990
SPLIT-ANTECEDENT relations	292
COREFERENCE clusters	5350
BRIDGING clusters	962

Corpus statistics 2

Type	Value
Anaphor relations	20425
Cataphor relations	277
Sentences per document	46.4
Tokens per document	709.7
Markables per document	214
Avg. COREFERENCE cluster length	4.7
Avg. BRIDGING cluster length	2.0
Avg. COREFERENCE distance	70.4
Avg. BRIDGING distance	32.1
Avg. SPLIT-ANTECEDENT distance	53.9

Corpus statistics 2

- Pronominal coreference accounts for about 38 % of the references in the corpus.
- The 12 most common anaphoric expressions are all pronouns, with the first noun, *Norge* 'Norway' as number 13.
- *jeg* 'I' is the most common pronoun, followed by han, de and hun.
- 71% of all anaphoric expressions occur only once in the corpus.

Experiments

- We apply a word-level coreference resolution framework (Dobrovolskii, 2021).
- Predict candidate antecedents for each token, before reconstructing the full spans.
- 80-10-10 split for train-dev-test
- Evaluating using five different metrics (MUC, B^3, $CEAF_e$,LEA,CoNLL Mean F_1)
- Evaluate two different language models: the monolingual NorBERT2 and the multilingual XLM-R.

Results

Model	MUC			B^3			$CEAF_e$			LEA			CoNLL
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	Mean F_1
NorBERT2	**90.40**	79.35	84.52	**63.15**	**62.71**	**62.93**	**55.52**	33.54	41.82	**61.94**	**61.50**	**61.72**	**63.09**
XLM-R	84.97	**84.51**	**84.74**	61.09	49.09	54.44	51.17	**51.17**	**51.17**	58.87	47.11	52.34	**63.45**

- NorBERT2 and XLM-RoBERTa performed better during initial testing
- High MUC scores indicate that the model was able to properly group mention clusters
- Lower B^3 and $CEAF_e$ scores indicate the presence of inaccurately assigned mention clusters.
- The LEA score also represents lack of entity assignment within discovered clusters.
- Scores are comparable to existing work on coreference resolution
- However, there are likely issues with entity resolution and assignment

Future Work

- Finalize the Nynorsk part of the corpus
- Re-align the coreference data with other annotation layers on the treebank
  - Part-of-Speech-tags
  - Named entity recognition
  - Dependency syntax
- Experiment further on the full dataset

`https://github.com/ltgoslo/NARC`

Acknowledgements

We want to express our gratitude to the many annotators involved with annotating the datasets: Fredrik Aas Andreassen, André Nilsson Dannevig, Marie Emerentze Fleisje, Jennifer Juveth, Annika Willoch Olstad, Anne Oortwijn, Stian Ramstad, Lilja Charlotte Storset, Veronica Dahlby Tveitan and Alexandra Wittemann. We are grateful for the initial funding from Teksthub, and to Språkbanken for the main funding of the project.

Thank you

Thank you for your attention.

경청해 주셔서 감사합니다