– IN5550 –

Neural Methods in Natural Language Processing

Home Exam: Task Overview and Kick-Off

Andrey Kutuzov, Egil Rønningstad, David Samuel,

Erik Velldal & Sondre Wold

University of Oslo

April 18th, 2023

UNIVERSITY OF OSLO

NO EN

← Studies ← Courses

- **Final Exam**
- Obligatory assign-ments
- slides
- videos

IN5550 - Spring 2023

Final exam

Between Thursday April 25 and Thursday May 16, there will be a home exam in the form of a small research project. To pass the exam, students need to submit and have accepted a scientific paper to the 5th IN5550 Teaching Workshop on Neural Natural Language Processing (WNNLP 2023).

Background

Final examination in this class takes the form of a 'home exam', i.e. a project that students can work on over a period of three weeks. Like for the exam-qualifying obligatory assignments, group work (in teams of up to three students) is encouraged (but **PhD-students** enrolled for **IN9550** must complete the home exam individually). Team composition needs to be declared at the start of the exam period and, cannot be changed after April 26, 2023. Also, each team needs to decide beforehand which of the available tracks they want to research; these 'tracks' of the exam will be introduced in the lecture on April 18 (examples from previous years have included named entity recognition, negation resolution, and sentiment analysis). Further background on the tracks and supporting data and code are available through the course GitHub repo. Please announce your team composition and choice of track no later than April 26 by emailing the course contact address `in5550-help@ifi.uio.no`.

The main exam period will be Tuesday, April 25, to Tuesday, May 16, 2023. Once a team

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

Fifth IN5550 Teaching Workshop on Neural NLP (WNNLP 2023)

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

Fifth IN5550 Teaching Workshop on Neural NLP (WNNLP 2023)

Standard Process

(1) Experimentation

(2) Analysis

(3) Paper Submission

(4) Reviewing

(5) Camera-Ready Manuscript

(6) Presentation

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

Fifth IN5550 Teaching Workshop on Neural NLP (WNNLP 2023)

Standard Process

(0) Problem Statement

(1) Experimentation

(2) Analysis

(3) Paper Submission

(4) Reviewing

(5) Camera-Ready Manuscript

(6) Presentation

WNNLP 2023: Call for Papers and Important Dates

General Constraints

- Three specialized tracks: Targeted Sentiment Analysis, Definition Generation, Machine Translation.
- Long papers: up to 8 pages (minimaly 5), excluding references, in ACL Rolling Review style.
- Submitted papers must be anonymous: peer reviewing is double-blind.
- Replicability: Submission backed by code repository (area chairs only).

WNNLP 2023: Call for Papers and Important Dates

General Constraints

- Three specialized tracks: Targeted Sentiment Analysis, Definition Generation, Machine Translation.
- Long papers: up to 8 pages (minimaly 5), excluding references, in ACL Rolling Review style.
- Submitted papers must be anonymous: peer reviewing is double-blind.
- Replicability: Submission backed by code repository (area chairs only).

Schedule

By April 26	Declare choice of track (and team composition)
Week 19	Track specific joint mentoring meeting
May 16	(Strict) Submission deadline for scientific papers
May 18–25	Reviewing period: Each student reviews two papers
May 26	Area Chairs make and announce acceptance decisions
June 2	Camera-ready manuscripts due, with requested revisions
June 6	Oral presentations and awards at the workshop

WNNLP 2023: Programme Committee

General Chair

- Erik Velldal

Track Chairs

- Targeted Sentiment Analysis: Egil Rønningstad
- Definition generation: Andrey Kutuzov
- Neural Machine Translation: David Samuel

Peer Reviewers

- All students who have submitted a scientific paper

- Sentiment Analysis:
    - identifying subjective content in text, and
    - measuring positive/negative polarity.
    - different granularities: Document-level, sentence-level, sub-sentence-level

Track 1: Targeted Sentiment Analysis

- Sentiment Analysis:
  - identifying subjective content in text, and
  - measuring positive/negative polarity.
  - different granularities: Document-level, sentence-level, sub-sentence-level
- Fine-grained sentiment analysis at the sub-sentence level
  - what is the target of sentiment?
  - what is the polarity of sentiment directed at the target?

1. Denne disken$_{POS}$ er svært stillegående
   'This disk runs very quietly'

En juleklassiker er utsprungen

Album jul

Guren Hagen med Sølvguttene og Ole Edvard Antonsen
Julekveld hos oss

En hjertevarm, personlig juleplate full av herlig musikalsk variasjon.

Guren Hagen er en av våre fineste viseartister, og med *Julekveld hos oss* har han skapt ei stemningsfull juleplate med stor troverdighet. Spesielt hans egenkomponerte julesanger gjør dette til julens store stemningsspreder. Hans varme stemme og østerdalsdialekt drar deg inn hans verden, mens Sølvguttenes vakre sang og Ole Edvard Antonsens trompet gir ekstra krydder til en av årets mest vellykkede juleplater. Guren Hagen har på utsøkt vis i tittelmelodien klart å formidle hvordan han opplevde barndommens østerdalsjul slik at vi alle vil være en del av denne kosen. Samtidig gjør han Alf Prøysens «Julekveldsvisa», med hawaiigitar og lapsteel til en sann fryd. I spennet mellom disse to sangene gir han oss troen tilbake på at det finnes overraskende grep også for denne musikktradisjonen. Selv kjente travere som «Det lyser i stille grender» og «Kimer i klokker», sistnevnte i en flott duett med Lise Backstrøm, besitter en nyvunnen nærhet. Dette er blitt en vidunderlig juleplate full av musikalsk overskudd og variasjon. Musikken bærer i seg like porsjoner glede, nærhet og undring.

Svein Andersen

- Newly released dataset for fine-grained SA of Norwegian
- `https://github.com/ltgoslo/norec_fine`

	# Examples				
	Train	Dev.	Test	Total	Avg. len.
Sents.	8634	1531	1272	11437	16.8
Targets	5044	877	735	6656	2.0

Table: Number of sentences and annotated targets across the data splits.

A Fine-Grained Sentiment Dataset for Norwegian

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, Erik Velldal
University of Oslo
Department of Informatics
{liljao,pettemae,jeremycb,erikve}@ifi.uio.no

Abstract

We introduce NoReC$_{fine}$, a dataset for fine-grained sentiment analysis in Norwegian, annotated with respect to polar expressions, targets and holders of opinion. The underlying texts are taken from a corpus of professionally authored reviews from multiple news-sources and across a wide variety of domains, including literature, games, music, products, movies and more. We here present a detailed description of this annotation effort. We provide an overview of the developed annotation guidelines, illustrated with examples, and present an analysis of inter-annotator agreement. We also report the first experimental results on the dataset, intended as a preliminary benchmark for further experiments.

Keywords: Sentiment analysis, opinion mining, Norwegian

1. Introduction

In this work, we describe the annotation of a fine-grained sentiment dataset for Norwegian, analysing opinions in terms of their polar expressions, targets, and holders. The dataset, including the annotation guidelines, is made publicly available[1] and is the first of its kind for Norwegian.

2. Related Work

Fine-grained approaches to sentiment analysis include opinion annotations as in (Wiebe et al., 2005), aspect-based sentiment (Hu and Liu, 2004), and targeted sentiment (Vo and Zhang, 2015). Whereas document- and sentence-level sentiment analysis make the simplifying assumption that all

Task specifics

- Data format: BIO (target + polarity)

  \# sent_id = 501595-13-04

Munken	B-targ-Positive
Bistro	I-targ-Positive
er	O
en	O
hyggelig	O
nabolagsrestaurant	O
for	O
hverdagslige	O
og	O
uformelle	O
anledninger	O
.	O

Task specifics

- Data format: BIO (target + polarity)

  # sent_id = 501595-13-04

Munken	B-targ-Positive
Bistro	I-targ-Positive
er	O
en	O
hyggelig	O
nabolagsrestaurant	O
for	O
hverdagslige	O
og	O
uformelle	O
anledninger	O
.	O

- Baseline system: PyTorch pre-code for BiLSTM
- Evaluation code provided

Possible directions

1. **Experiment with alternative hyperparameters and pretrained language models.** (If this is all the experimenting you do, the analysis part of your paper needs to be very good.)
2. **Error analysis:** Confusion matrix, the most common errors, target length vs errors, most common words missed or wrongly classified, etc.
3. **Cross-domain performance**
4. **Experiment with finer-grained sentiment annotations**

Track 2: Definition Modeling for Norwegian

Definition modeling for Norwegian with Encoder-Decoder Language Models

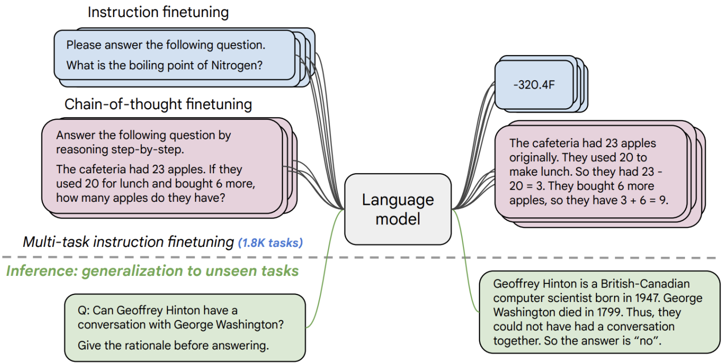
Definition modeling is automatically generating definitions for individual word usages and word senses.

Generated definitions

Usage example	Word	Definition
'about half of the soldiers in our rifle platoons were **draftees** whom we had trained for about six weeks'	**draftee**	*a person who is enlisted in the army or navy under compulsory draft.*

We need a proper prompt for a generative language model. For example, the usage example + 'What is the definition of TARGET_WORD?'
'about half of the soldiers in our rifle platoons were **draftees** whom we had trained for about six weeks. what is the definition of **draftee**?

Track 2: Definition Modeling for Norwegian



Definition is a text and the usage example is a text.
Let's conditionally generate definitions from usage examples!

Pre-code: `https://github.uio.no/in5550/2023/tree/main/exam/definition_modeling`

Track 2: Definition Modeling for Norwegian

- **Definition generation** is a well-established NLG field [Mickus et al., 2019, Huang et al., 2021, Kong et al., 2022]:
  - Task formulation 1: Generate a definition given a target word alone
  - Task formulation 2: Generate a definition given a target word and an example usage
  - SOTA for task 2: Fine-tuned encoder-decoder models (BART, T5, etc).
- This project: generate definitions for Norwegian words (including polysemous) with the corresponding sets of example usages.
- For this, we will use fine-tuned FLAN-T5 [Chung et al., 2022] and norT5 (by LTG) models.

Track 2: Definition Modeling for Norwegian

Some impressive examples from English

The word *'word'* in three senses.

1. 'There are people out there who have never heard of the Father, Son and Holy Spirit, let alone the **Word** of God.'
2. 'Good News Bible Before the world was created, the **Word** already existed; he was with God, and he was the same as God.'
3. 'It was in that basement that I learned the skills necessary to succeed in the difficult thespian world-specifically, get up on stage, say my **words**, get off the stage-skills...'

Definitions generated by a fine-tuned FLAN-T5-XL

1. 'THE BIBLE'
2. '(CHRISTIANITY) THE SECOND PERSON OF THE TRINITY ; JE'
3. 'THE DIALOGUE OF A PLAY.'

Track 2: Definition Modeling for Norwegian

Reference-based evaluation

Typical NLG and MT metrics: BLEU, NIST, METEOR, ROUGE, SACREBLEU, etc

Track 2: Definition Modeling for Norwegian

Reference-based evaluation

Typical NLG and MT metrics: BLEU, NIST, METEOR, ROUGE, SACREBLEU, etc

But you will evaluate the models under different setups:

- In-distribution: Fine-tune an LM on a Norwegian definition dataset and test it on a held-out subset of the same corpus.
- Language shift: Fine-tune an LM on a corpus of definitions in L1 and test it on L2 (English →Norwegian, Spanish → Norwegian, etc)
- Task shift (zero-shot): Directly test an LM on a Norwegian definition dataset, without any fine-tuning.

Track 2: Definition Modeling for Norwegian

Data

Main statistics of the datasets of definitions for English. Ratio is the *sense/lemma* ratio: the number of entries over the number of lemmas.

Dataset	Entries	Lemmas	Ratio	Usage length	Definition length
WordNet	15,657	8,938	1.75	4.80 ± 3.43	6.64 ± 3.77
Oxford	122,318	36,767	3.33	16.73 ± 9.53	11.01 ± 6.96
CoDWoE	63,596	36,068	2.44	24.04 ± 21.05	11.78 ± 8.03

Track 2: Definition Modeling for Norwegian

Data

Main statistics of the datasets of definitions for English. Ratio is the *sense/lemma* ratio: the number of entries over the number of lemmas.

Dataset	Entries	Lemmas	Ratio	Usage length	Definition length
WordNet	15,657	8,938	1.75	4.80 ± 3.43	6.64 ± 3.77
Oxford	122,318	36,767	3.33	16.73 ± 9.53	11.01 ± 6.96
CoDWoE	63,596	36,068	2.44	24.04 ± 21.05	11.78 ± 8.03

A typical example (Oxford)

- Synset: orphanage%oxford.0
- Usage example: *his early orphanage shaped his character as an adult*
- Definition: 'THE CONDITION OF BEING A CHILD WITHOUT LIVING PARENTS'

Track 2: Definition Modeling for Norwegian

Data

Main statistics of the datasets of definitions for English. Ratio is the *sense/lemma* ratio: the number of entries over the number of lemmas.

Dataset	Entries	Lemmas	Ratio	Usage length	Definition length
WordNet	15,657	8,938	1.75	4.80 ± 3.43	6.64 ± 3.77
Oxford	122,318	36,767	3.33	16.73 ± 9.53	11.01 ± 6.96
CoDWoE	63,596	36,068	2.44	24.04 ± 21.05	11.78 ± 8.03

A typical example (Oxford)

- Synset: orphanage%oxford.0
- Usage example: *his early orphanage shaped his character as an adult*
- Definition: 'THE CONDITION OF BEING A CHILD WITHOUT LIVING PARENTS'

CoDWoE dataset [Mickus et al., 2022] also features definitions for Spanish, Italian and Russian. Can be useful for cross-lingual transfer!

Track 2: Definition Modeling for Norwegian

- There are no definition modeling datasets for Norwegian.
- ...or at least I did not find any.
- But we have *Det Norske Akademis Ordbok* ...

Track 2: Definition Modeling for Norwegian

- There are no definition modeling datasets for Norwegian.
- ...or at least I did not find any.
- But we have *Det Norske Akademis Ordbok* ...
- `https://naob.no/ordbok`

hund

hund substantiv

BØYNING en; hunden, hunder
UTTALE [hun:]
ETYMOLOGI av norrønt *hundr*

BETYDNING OG BRUK
INNHOLDSFORTEGNELSE

1 temmet rovdyr i hundefamilien | vitenskapelig navn *Canis familiaris*

EKSEMPLER
- *ha, holde hund*
- *en herreløs hund*
- *ville og tamme hunder*
- *en klynkende hund*

SITATER
- *hun så på ham, sådan som en hund ser på sin herre* (Bjørnstjerne Bjørnsons fortællinger 312)
- *dyrenes karakteristik synes aabenbart at være: min hund og andres bikjer* (Morgenbladet 1930/25/6/4)
- *en skitten hund smatt vekk gjennom et hull i gjerdet* (Finn Carling *Gjenskinn* LBK 1994)
- *– Følger du meg? – Som en trofast hund* (John Ege *Dominoklubben* LBK 1995)
- *– Er bikkja på beina? spurte Gøran. ... – Hunden, rettet han* (Karin Fossum *Elskede Poona* 226 2009)
- *[han løfter] to fingre til munnen, plystrer hardt og en hund kommer løpende med en pinne i munnen* (Eivind Hofstad Evjemo *Det siste du skal se er et ansikt av kjærlighet* LBK 2012)

Track 2: Definition Modeling for Norwegian

Encoder-decoder models to try

- FLAN-T5 (https://huggingface.co/google/flan-t5-base)
- T5 (https://huggingface.co/t5-base)
- Multilingual T5 (https://huggingface.co/google/mt5-base)
- norT5 (https://huggingface.co/ltg/nort5-base)
- ...

The models come in different sizes and are available locally on Fox.

HuggingFace Transformers has good support for conditional generation with T5-like models.
https://huggingface.co/docs/transformers/en/model_doc/auto#transformers.AutoModelForSeq2SeqLM

Track 2: Definition Modeling for Norwegian

Workflow

1. Collect a reasonably-sized (≈ 100 instances) definition dataset from *Det Norske Akademis Ordbok*
2. Split it into train-dev-test in whatever way you see fit.
3. Try to generate Norwegian definitions zero-shot from multilingual and Norwegian T5 models
4. Evaluate the results qualitatively and quantitatively
5. Fine-tune the language models:
   - on Norwegian data
   - on data from other languages
   - on everything
   - ....
6. Evaluate the results as well.
7. Play with hyperparameters.
8. Discuss your findings.

Track 2: Definition Modeling for Norwegian

Possible research directions

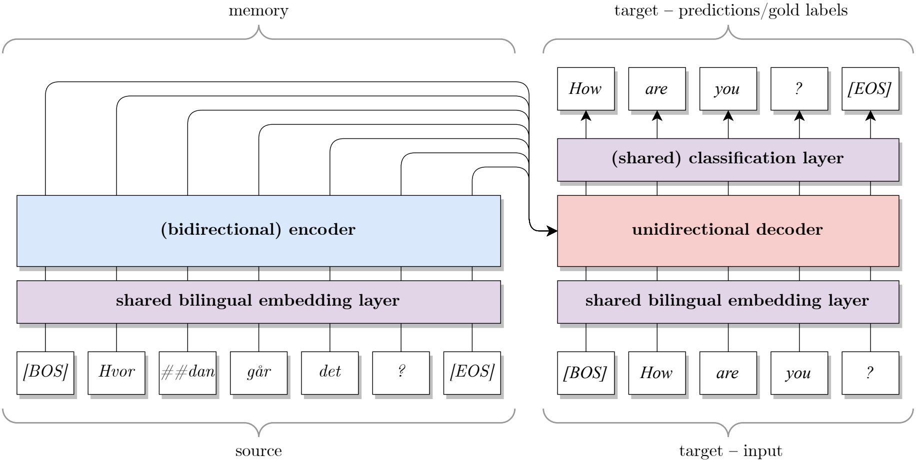
1. What prompts to use for Norwegian? Does it depend on the fine-tuning data
2. Does it actually help to fine-tune on a toy-sized definition dataset?
3. Is it better to always generate one top definition, or sample n most probable predictions?
4. Can we use the information about parts of speech to improve definition modeling?
5. Do multilingual models possess zero-shot abilities regarding Norwegian?
6. How the quality of definition modeling increases with the scale of the models? Does fine-tuning compensate for it or not?
7. ...

Pre-code: `https://github.uio.no/in5550/2023/tree/main/exam/definition_modeling`

Track 3: Neural Machine Translation

See more in the detailed task description:

`https://github.uio.no/in5550/2023/tree/main/exam/nmt`



WNLP 2023: What Makes a Good Scientific Paper?

Empirical (Experimental)

- Motivate architecture choice(s) and hyper-parameters;
- systematic exploration of relevant parameter space;
- comparison to reasonable baseline or previous work.

Replicable (Reproducible)

- Everything relevant to run and reproduce in GitHub.

Analytical (Reflective)

- Identify and relate to previous work;
- explain choice of baseline or points of comparison;
- meaningful, precise discussion of results;
- ‘negative’ results can be interesting too;
- look at the data: discuss some examples;
- error analysis: identify remaining challenges.

References I

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models.

Huang, H., Kajiwara, T., and Arase, Y. (2021). Definition modelling for appropriate specificity. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

References II

Kong, C., Chen, Y., Zhang, H., Yang, L., and Yang, E. (2022).
Multitasking framework for unsupervised simple definition generation.
In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Mickus, T., Paperno, D., and Constant, M. (2019).
Mark my word: A sequence-to-sequence approach to definition modeling.
In Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, pages 1–11, Turku, Finland. Linköping University Electronic Press.

References III

Mickus, T., Van Deemter, K., Constant, M., and Paperno, D. (2022). Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings.
In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1–14, Seattle, United States. Association for Computational Linguistics.