BOKMÅL

Summary

The Norwegian UD treebank is based on the Bokmål section of the Norwegian Dependency Treebank (NDT), which is a syntactic treebank of Norwegian. The current version of NDT has been automatically converted to the UD scheme by Ingerid Løyning Dale, Per Erik Solberg and Andre Kåsen at the Norwegian Language Bank at the National Library of Norway. This conversion builds to a large extent on previous conversions by Lilja Øvrelid at the University of Oslo.

Introduction

NDT was developed 2011-2014 at the National Library of Norway in collaboration with the Text Laboratory and the Department of Informatics at the University of Oslo. NDT contains around 300,000 tokens taken from a variety of genres. The treebank texts have been manually annotated for morphosyntactic information. The morphological annotation mainly follows mainly the Oslo-Bergen Tagger (http://tekstlab.uio.no/obt-ny/). The syntactic annotation follows, to a large extent, the Norwegian Reference Grammar, as well as a dependency annotation scheme formulated at the outset of the annotation project and iteratively refined throughout the construction of the treebank. For more information, see the references below.

Data splits

In creating the data splits, care has been taken to preserve contiguous texts in the different splits and also to keep a fair balance of genres in each of the splits. Petter Hohle created the splits for the Norwegian UD treebank. The splits were created by concatenating the following files (available with the distribution of NDT):

Training data (15.696 sentences, 180 individual files):

- ap001_0000 -- ap012_0002 (53 files)
- bt001_0000 -- bt005_0001 (28 files)
- db001a_0000 -- db013_0004 (42 files)
- kk001_0000 -- kk005_0001 (10 files)
- sp-bm001_0000 -- sp-bm001_0008 (9 files)
- vg001_0000 -- vg002_0003 (8 files)
- blogg-bm001_0000 -- blogg-bm003_0000 (9 files)
- nou001_0000 -- nou004_0000 (10 files)
- st001_0000 -- st005_0000 (11 files)

Development data (2.410 sentences, 26 individual files):

- ap012_0003 -- ap014_0002 (7 files)
- bt005_0002 -- bt005_0005 (4 files)
- db013_0005 -- db014_0002 (5 files)
- kk006_00001 -- kk007_0000 (2 files)
- sp-bm002_0000 -- sp-bm002_0001 (2 files)
- vg002_0004 (1 file)
- blogg-bm003_0001 -- blogg-bm003_0002 (2 files)
- nou004_0001 (1 file)
- st005_0001 -- st005_0002 (2 files)

Test data (1.939 sentences, 26 individual files):

- ap014_0003 -- ap015_0002 (7 files)
- bt005_0006 -- bt006_0001 (4 files)
- db014_0003 -- db014_0007 (5 files)
- kk007_0001 -- kk008_0000 (2 files)
- sp-bm003_0000 -- sp-bm003_0001 (2 files)
- vg002_0005 (1 file)
- blogg-bm003_0003 -- blogg-bm003_0004 (2 files)
- nou004_0002 (1 file)
- st005_0003 -- st005_0004 (2 files)

Basic statistics

- Tree count: 20.045
- Word count: 311.277
- Token count: 311.277
- Dep. relations: 35, of which 2 language specific
- POS tags: 17
- Category=value feature pairs: 31

Tokenization

White space always indicates a token boundary and punctuation constitute separate tokens, except:

- numbers with periods, commas or colons, e.g. *1.3*, *0,6*, *10:13*
- abbreviations, e.g. *f.eks.*, *Carl J. Hambro*
- URLs, e.g. *http://www.ifi.uio.no*

The treebank does not contain multiword tokens.

Morphology

The PoS-tags follow the universal tag set and does not add any language-specific PoS-tags. The morphological features follow the Oslo-Bergen Tagger scheme (Hagen et. al., 2000). PoS-tags and morphological features were converted automatically to the UD scheme.

Syntax

The syntactic annotation in the Norwegian UD treebank conforms to the UD guidelines, adding language-specific relations for relative clauses (`acl:relcl`) and verb particles (`compound:prt`). The annotation has been automatically converted to UD from the original dependency scheme described in Solberg et. al. (2014) and further described in the NDT guidelines (Kinn et. al.).

The conversion has not been manually checked. There are a few known discrepancies from UD:

- no mwe analysis in the treebank. This is also information that is not present in the original data.

References

Kristin Hagen, Janne Bondi Johannessen and Anders Nøklestad: "A Constraint-based Tagger for Norwegian". 2000. Proceedings of the 17th Scandinavian Conference in Linguistics.

Kari Kinn, Per Erik Solberg and Pål Kristian Eriksen. "NDT Guidelines for Morphological Annotation". National Library Tech Report.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen and Janne Bondi Johannessen. 2014."The Norwegian Dependency Treebank", Proceedings of LREC 2014, Reykjavik

Lilja Øvrelid & Petter Hohle (2016). "Universal Dependencies for Norwegian" (http://www.lrec-conf.org/proceedings/lrec2016/pdf/462_Paper.pdf), In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'16)

Acknowledgements

NDT has been automatically converted to the UD scheme by Ingerid Løyning Dale, Per Erik Solberg and Andre Kåsen at the Norwegian Language Bank at the National Library of Norway.

This conversion builds to a large extent on previous conversions by Lilja Øvrelid at the University of Oslo. Petter Hohle created the data splits and Fredrik Jørgensen aligned the treebank to the original texts.

We thank the annotators of the original NDT: Pål Kristian Eriksen, Kari Kinn and Per Erik Solberg.

Changelog

--> UD 2.12

- The conversion is completely rewritten using Grew (https://grew.fr/) by the Norwegian Language Bank at the National Library of Norway. The conversion is to a large
- extent based on the guidelines of the previous version.
- *som* in relative clauses is not longer treated as pronouns, but complementizers with the postag SCONJ and the label mark.
- There is no longer an explicit analysis of verbal particles. The postag has changed from ADP to ADV and the label is advmod.
- The changes in 2.10 and 2.12 (https://universaldependencies.org/changes.html) are implemented.

UD 1.3 --> UD 1.4

- Added SpaceAfter annotation describing tokenization
- Removed tokens that were introduced during treebanking but are not present in original texts: 1) tokens introduced for paragraph boundaries (|), and 2) extra punctuation introduced following sentence-final abbreviations.

BOKMÅL

UD 1.2 --> UD 1.3

- In order to improve consistency between the Germanic languages, a list of auxiliary verbs were agreed on. For Norwegian these are: *bli* 'become', *burde* 'should', *få* 'get', *ha* 'have', *kunne* 'can', *måtte* 'must', *skulle* 'should', *tørre* 'dare', *ville* 'will', *være* 'be'.
- Appositions (`appos`) now exclusively follow their head, in compliance with UD guidelines
- Earlier `nsubj` in cleft constructions are now `dislocated`
- Adpositions marking a subordinate clause are now `mark` as opposed to earlier `case`
- In v1.2 a right-headed analysis was used for combinations of proper and common nouns such as *Premier League-spillere* 'Premier League players'. In v1.3 a `compound` analysis was chosen for these.
- PronType feature was added for all PRON or DET, and VerbType for all VERB and AUX
- General bug fixing

Summary

The Norwegian UD treebank is based on the Nynorsk section of the Norwegian Dependency Treebank (NDT), which is a syntactic treebank of Norwegian. NDT has been automatically converted to the UD scheme by Lilja Øvrelid at the University of Oslo.

Introduction

NDT was developed 2011-2014 at the National Library of Norway in collaboration with the Text Laboratory and the Department of Informatics at the University of Oslo.

NDT contains around 300,000 tokens taken from a variety of genres. The treebank texts have been manually annotated for morphosyntactic information. The morphological annotation mainly follows the Oslo-Bergen Tagger http://tekstlab.uio.no/obt-ny/. The syntactic annotation follows, to a large extent, the Norwegian Reference Grammar, as well as a dependency annotation scheme formulated at the outset of the annotation project and iteratively refined throughout the construction of the treebank. For more information, see the references below.

Data splits

In creating the data splits, care has been taken to preserve contiguous texts in the different splits and also to keep a fair balance of genres in each of the splits. Petter Hohle created the splits for the Norwegian UD treebank. The splits were created by concatenating the following files (available with the distribution of NDT):

Training data (245.330 tokens, 14.174 sentences)

- blogg-nn001_0000 -- blogg-nn003_0001
- dot001_0000 -- dot014_0007
- firda-nn001_0000 -- firda-nn004_0005
- kk-nn001_0000 -- kk-nn006_0002
- mom001_0000 -- mom003_0004
- st-nn001_0000 -- st-nn002_0000
- vtb-nn001_0000 -- vtb-nn006_0004

Development data (31.250 tokens, 1.890 sentences)

- blogg-nn003_0002
- dot014_0008 -- dot015_0002
- firda-nn004_0006 -- firda-nn005_0002
- kk-nn006_0003 -- kk-nn007_0002
- mom003_0005
- st-nn002_0001
- vtb-nn006_0005 -- vtb-nn007_0001

Test data (24.773 tokens, 1.511 sentences)

- blogg-nn003_0003
- dot015_0003 -- dot016_0004
- firda-nn005_0003 -- firda-nn005_0006
- kk-nn007_0003 -- kk-nn007_0006
- mom003_0006
- st-nn002_0002
- vtb-nn007_0002 -- vtb-nn007_0004

Basic statistics

- Tree count: 17.575
- Word count: 301.353
- Token count: 301.353

Tokenization

White space always indicates a token boundary and punctuation constitute separate tokens, except:

- numbers with periods, commas or colons, e.g. *1.3*, *0,6*, *10:13*
- abbreviations, e.g. *f.eks.*, *Carl J. Hambro*
- URLs, e.g. *http://www.ifi.uio.no*

The treebank does not contain multiword tokens.

Morphology

The PoS-tags follow the universal tag set and does not add any language-specific PoS-tags. The morphological features follow the Oslo-Bergen Tagger scheme (Hagen et. al., 2000). PoS-tags and morphological features were converted automatically to the UD scheme.

Syntax

The syntactic annotation in the Norwegian UD treebank conforms to the UD guidelines, adding a language-specific relation for relative clauses (`acl:relcl`). The annotation has been automatically converted to UD from the original dependency scheme described in Solberg et. al. (2014) and further described in the NDT guidelines (Kinn et. al.).

The conversion has not been manually checked. There are a few known discrepancies from UD:

- no mwe analysis in the treebank. This is also information that is not present in the original data.

References

Kristin Hagen, Janne Bondi Johannessen and Anders Nøklestad: "A Constraint-based Tagger for Norwegian". 2000. Proceedings of the 17th Scandinavian Conference in Linguistics.

Kari Kinn, Per Erik Solberg and Pål Kristian Eriksen. "NDT Guidelines for Morphological Annotation". National Library Tech Report.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen and Janne Bondi Johannessen. 2014."The Norwegian Dependency Treebank", Proceedings of LREC 2014, Reykjavik

Velldal, Erik; Øvrelid, Lilja & Hohle, Petter (2017). Joint UD Parsing of Norwegian Bokmål and Nynorsk, In Jörg Tiedemann (ed.), Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa). Linköping University Electronic Press. ISBN 978-91-7685-601-7. Article no. 001. s 1 - 10

Acknowledgements

NDT has been automatically converted to the UD scheme by Lilja Øvrelid at the University of Oslo. Petter Hohle created the data splits and Fredrik Jørgensen aligned the treebank to the original texts.

We thank the annotators of the original NDT: Pål Kristian Eriksen, Kari Kinn and Per Erik Solberg.

Changelog

--> UD 2.12

- The conversion is completely rewritten using [Grew] (https://grew.fr/) by the Norwegian Language Bank at the National Library of Norway. The conversion is to a large extent based on the guidelines of the previous version.
- *som* in relative clauses is no longer treated as pronouns, but complementizers with the postag SCONJ and the label mark.
- There is no longer an explicit analysis of verbal particles. The postag has changed from ADP to ADV and the label is advmod.
- [The changes in 2.10 and 2.12] (https://universaldependencies.org/changes.html) are implemented.