Cross-Lingual Constituency Parsing for Middle High German: A Delexicalized Approach

Ercong Nie [1,2]   Helmut Schmid[1]   Hinrich Schütze[1,2]

[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany
nie@cis.lmu.de

September 8, 2023
ALP 2023 @ Varna, Bulgaria

Outline

1. Introduction

2. Delexicalization Parsing for Middle High German

3. Experimental Setup and Results

4. Conclusion

Parsing for Historical Languages

- **Syntactically annotated corpora of historical languages**
  - form the foundation for **linguistic analysis** (language change, contact and variation, linguistic evolution of morphology, syntax, etc.).
  - serve as a building block for **NLP applications**.
  - enrich **interdisciplinary** cultural, literature and historical studies.

Parsing for Historical Languages

- **Syntactically annotated corpora of historical languages**
  - form the foundation for **linguistic analysis** (language change, contact and variation, linguistic evolution of morphology, syntax, etc.).
  - serve as a building block for **NLP applications**.
  - enrich **interdisciplinary** cultural, literature and historical studies.
- **Difficulties** in constructing parsed corpora for historical languages:
  - Scarcity of digital text resources
  - High demand of linguistic expertise
  - Large manual effort

Parsing for Historical Languages

- **Syntactically annotated corpora of historical languages**
  - form the foundation for **linguistic analysis** (language change, contact and variation, linguistic evolution of morphology, syntax, etc.).
  - serve as a building block for **NLP applications**.
  - enrich **interdisciplinary** cultural, literature and historical studies.

- **Difficulties** in constructing parsed corpora for historical languages:
  - Scarcity of digital text resources
  - High demand of linguistic expertise
  - Large manual effort

→ **Solution**: Training an automatic syntactic analysis system using cross-lingual transfer techniques.

Historical German Language Resources

- **POS-Tagged Corpora**
  - German Reference Corpus[1]

Referenzkorpus Mittelniederdeutsch / Niederrheinisch

Universität Hamburg

2,3 Mio. Texttoken
1200–1650

Referenzkorpus Altdeutsch

Humboldt-Universität zu Berlin

1/2 Mio. Texttoken
750–1050

Referenzkorpus Frühneuhochdeutsch

Ruhr-Universität Bochum

3,5 Mio. Texttoken
1350–1650

Referenzkorpus Mittelhochdeutsch

Ruhr-Universität Bochum
Universität Bonn

2,5 Mio. Texttoken
1050–1350

[1] https://www.deutschdiachrondigital.de/

[2] https://korpling.german.hu-berlin.de/ddb-doku/index.htm

[3] https://ipchg.iu.edu/index.html

[4] https://www.chlg.ugent.be/

Historical German Language Resources

- **POS-Tagged Corpora**
  - German Reference Corpus[1]

Referenzkorpus Mittelniederdeutsch / Niederrheinisch

Universität Hamburg

2,3 Mio. Texttoken
1200–1650

ReN

Referenzkorpus Altdeutsch

Humboldt-Universität zu Berlin

1/2 Mio. Texttoken
750–1050

ReA

Referenzkorpus Frühneuhochdeutsch

Ruhr-Universität Bochum

3,5 Mio. Texttoken
1350–1650

ReF

Referenzkorpus Mittelhochdeutsch

Ruhr-Universität Bochum
Universität Bonn

2,5 Mio. Texttoken
1050–1350

ReM

- **Parsed Corpora**

Id.	Name	Languages	Style	Size
DDB[2]	German Diachronic Treebank	OHG, MHG, ENHG	Tiger	8,580
IPCHG[3]	Indiana Parsed Corpus of Historical (High) German	OHG, MHG, ENHG	PTB	∼10,000
CHLG[4]	Corpus of Historical Low German	MLG, OLG	PTB	∼200,000

[1] https://www.deutschdiachrondigital.de/

[2] https://korpling.german.hu-berlin.de/ddb-doku/index.htm

[3] https://ipchg.iu.edu/index.html

[4] https://www.chlg.ugent.be/

Outline

1. Introduction

2. Delexicalization Parsing for Middle High German

3. Experimental Setup and Results

4. Conclusion

Motivation

Our work focused on the constituency parsing of **Middle High German** (MHG):

- a historical stage of the German language that was spoken between 1050 and 1350.
- the linguistic predecessor of Modern German (MG).

Motivation

Our work focused on the constituency parsing of **Middle High German** (MHG):

- a historical stage of the German language that was spoken between 1050 and 1350.
- the linguistic predecessor of Modern German (MG).

Motivation of the Delexicalization Method:

- The continuity in the process of language evolution gives rise to **linguistic similarities** between **MG** and **MHG**.
  - Similar sentence structure
  - Similar word order
- **Rich resources of MG** texts with syntactic annotations.
  - Tiger Corpus (Smith, 2003)

Delexicalization Parsing System for MHG

The delexicalization parsing system for MHG comprises three modules:

- **POS Tagger**
  - Annotates a sequence of MHG tokens with POS and morphological tags.
  - Trained on the ReM corpus using RNNTagger (Schmid, 2019).

Delexicalization Parsing System for MHG

The delexicalization parsing system for MHG comprises three modules:

- **POS Tagger**
  - Annotates a sequence of MHG tokens with POS and morphological tags.
  - Trained on the ReM corpus using RNNTagger (Schmid, 2019).
- **Tag Mapper**
  - Mapping tags from the HiTS tag set (used for ReM) to STTS tag set (used for MG treebanks).

Delexicalization Parsing System for MHG

The delexicalization parsing system for MHG comprises three modules:

- **POS Tagger**
  - Annotates a sequence of MHG tokens with POS and morphological tags.
  - Trained on the ReM corpus using RNNTagger (Schmid, 2019).
- **Tag Mapper**
  - Mapping tags from the HiTS tag set (used for ReM) to STTS tag set (used for MG treebanks).
- **Delexicalized Parser**
  - Based on the Berkeley Neural Parser (Benepar) (Kitaev and Klein, 2018)
  - Trained on the Tiger Treebank (50,474 MG parse trees)

Delexicalization Parsing System for MHG

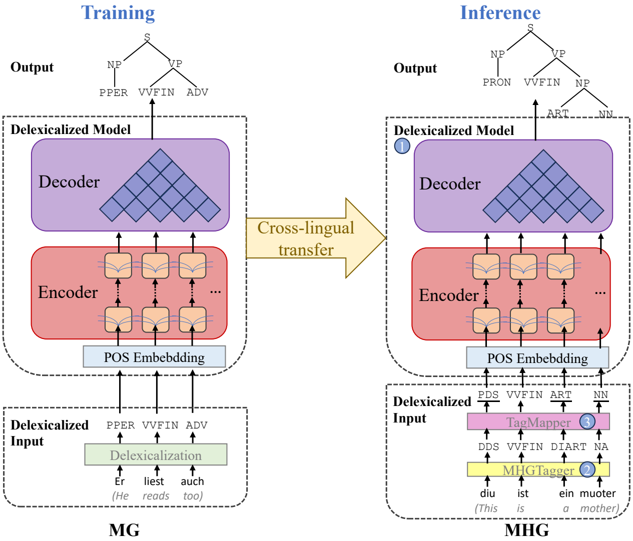

Figure: Overview of the cross-lingual delexicalized parsing system for MHG

Outline

1. Introduction

2. Delexicalization Parsing for Middle High German

3. Experimental Setup and Results

4. Conclusion

Baselines

- **Vanilla Benepar**: performing a vanilla zero-shot cross-lingual transfer, training a Benepar model directly on MG treebanks without the delexicalization.

- **Tetra-Tagging with PLMs**: a technique reducing constituency parsing to sequence labeling (Kitaev and Klein, 2020)
  - **gBERT**: Tetra-Tagging with the German BERT model (Chan et al., 2020)
  - **mBERT**: Tetra-Tagging with the multilingual BERT model (Devlin et al., 2019)

Main results

	Recall		**Precision**		**FScore**		**CM**	
	MG	MHG	MG	MHG	MG	MHG	MG	MHG
Baselines								
Vanilla Benepar	84.18	34.41	87.57	44.40	85.84	38.77	45.80	0.00
Tetra-gBERT	**86.31**	23.20	**88.19**	29.53	**87.24**	25.98	**51.70**	3.12
Tetra-mBERT	60.68	19.69	65.61	23.25	63.15	21.32	21.35	0.00
Our proposed method								
Dexparser	81.39	**64.72**	84.89	**70.19**	83.10	**67.34**	39.03	**12.50**

Table: Parsing performance of different cross-lingual transfer methods. **CM** refers to “complete match” The best value of each column is indicated in **bold**.

- Dexparser demonstrates substantial advantages in parsing MHG compared to other baselines.
- Dexparser also achieves comparable results on MG.

Ablation Study

	Recall	**Precision**	**FScore**	**CM**
Delexicalized parser using gold tags	**66.18**	**71.17**	**68.59**	**14.58**
- *using predicted tags*	64.72	70.19	67.34	12.50
- *without mapping*	59.16	68.82	63.63	7.29
- *without morphological information*	48.66	65.38	55.8	9.28

Table: The MHG parsing results with delexicalized parser in the ablation study.

- **Quality of POS annotation**, **tag set mapping** and **annotation of morphological information** collectively contribute to the performance of the delexicalization parser on MHG.

Outline

1 Introduction

2 Delexicalization Parsing for Middle High German

3 Experimental Setup and Results

4 Conclusion

Conclusion

1. We present an effective cross-lingual constituency parsing approach by using the delexicalization.

2. We utilize the linguistic similarities between MHG and Modern German (MG) to develop an automatic syntactic annotation system for Middle High German (MHG) based on the rich treebank resources of MG.

3. Our work provides a solution for the parsing of historical and ancient languages facing similar situations:
   a. having relevant (modern) languages with rich treebank resources,
   b. having rich POS-tagged text data.

Thanks for your attention!

References

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2020. Tetra-tagging: Word-synchronous parsing with linear-time inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6255–6261, Online. Association for Computational Linguistics.

Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, pages 133–137.

George Smith. 2003. A brief introduction to the tiger treebank, version 1. Technical report, Universität Potsdam.