DErivBase: A derivational morphology resource for German

Britta D. Zeller*, Jan Šnajder†, Sebastian Padó*

*Institute of Computational Linguistics, Heidelberg University
†Faculty of Electrical Engineering and Computing, University of Zagreb

The 51st Annual Meeting of the
Association for Computational Linguistics
August 6, 2013

EXCITEMENT

hrzz
Hrvatska zaklada za znanost

A derivational resource – what is that?

- Derivation: a morphological process of word formation
- Derivational resource groups content words into derivational families:
  to $sleep_V$ – $sleepy_A$ – $sleepless_A$ – $sleep_N$ – . . .
  ⇒ Concept for a set of morphologically related words across POSes
- Resource provides information of morphological relatedness
  ↔ frequently implies semantic relatedness
- Degree of similarity depends on idiosyncrasies:
  $book_N$ – $bookish_A$
- Most previous research in computational morphology is about inflection normalisation, although derivational information is valuable

A derivational resource – what for?

Accounts for semantic relationships across POS boundaries:

- Extension of semantic roles resources [Green et al., 2004]:
  Extend lexical unit inventory of FrameNet [Baker et al., 1998]:
  to $ornament_V$ – $ornamentation_N$
- Improvement of text fluency:
  Reformulation in Natural Language Generation
  [Thadani and McKeown, 2011]:
  Ferrero is mainly a candy $producer_N$. $\rightarrow$ *Ferrero* $produces_V$ *candies.*
- Textual Entailment [Szpektor and Dagan, 2008]:
  Knowledge of derivations provides information for inference rules,
  e.g. noun modifiers which act as predicate:
  the $running_A$ $X \leftrightarrow X$ $runs_V$

Related Work

- Manually constructed morphological analyzers: two-level approach, replacement rules in finite state technology [Koskenniemi, 1983], [Karttunen and Beesley, 2005]
- Unsupervised morphology learning with statistical and data-driven methods [Déjean, 1998, Schone and Jurafsky, 2000, Hammarström and Borin, 2011]
  - No distinction between different morphological processes
  - We aim at more fine-grained control over precision and recall
- Derivational resource for English: CatVar [Habash and Dorr, 2003]
  - Builds on resources available only for English

Morphology for German

- Related resources and their shortcomings:
  - CELEX [Baayen et al., 1996]: Limited coverage
  - IMSLEX [Fitschen, 2004]: Not publicly available
  - SMOR [Schmid et al., 2004], Morphix [Finkler and Neumann, 1988]: No distinction between inflection, compounding, and derivation

- DErivBase:
  - Publicly available
  - Contains morphologically related derivational families from a corpus
  - Covers over 280,000 German verbs, nouns, and adjectives
  - Rule-based approach → high precision

A rule-based approach

Motivation:

- German derivational processes are quite regular
- Small number of generic processes; can be freely combined
- Rules based on preexisting linguistic knowledge

Examples for derivational processes:

- Suffix derivation: *to edit*$_V$ – *edition*$_N$
  "append 'ion' to the end of the stem"
- Stem change: *to sing*$_V$ – *song*$_N$
  "replace 'i' by 'o' "
- Combinations: *to perceive*$_V$ – *perception*$_N$
  "alter stem 'eive' into 'ept', append 'ion' to the end of the stem"

Application of rule-based framework

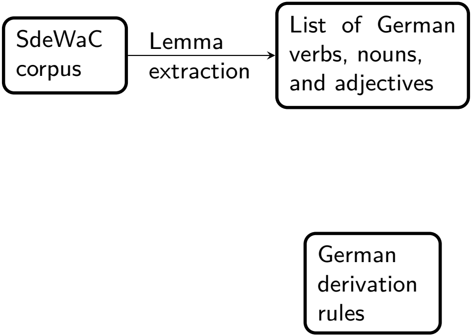
German derivation rules

Application of rule-based framework

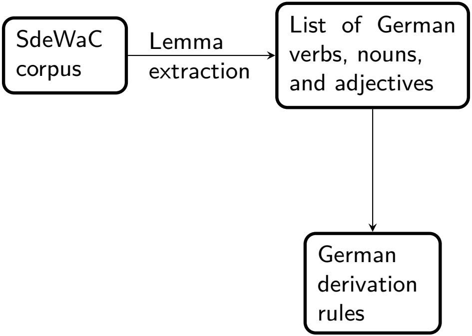
List of German verbs, nouns, and adjectives

German derivation rules

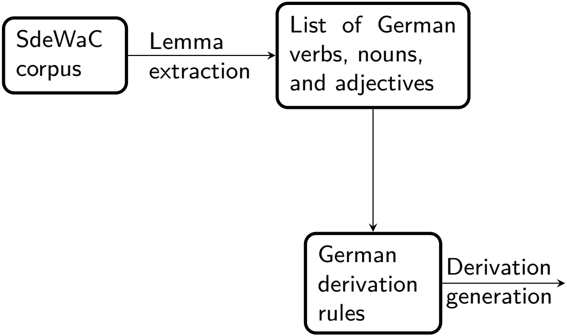
Application of rule-based framework



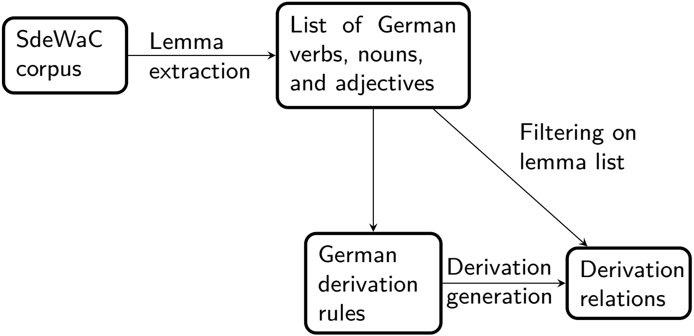
Application of rule-based framework



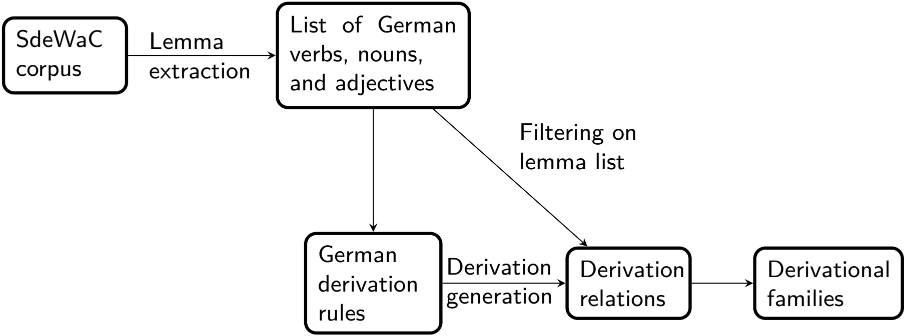
Application of rule-based framework



Application of rule-based framework



Application of rule-based framework



Definition of rule-based framework

- Modeling framework by [Šnajder and Dalbelo Bašić, 2010]
- Core of the framework:
  - Transformation function t: Maps a *basis* lemma into a *derived* lemma:
    Input: *to manage*$_V$
    Function: *sfx('ment')*
    Output: *management*$_N$
  - Inflectional paradigms $\mathcal{P}_1$, $\mathcal{P}_2$: POS and gender information for basis/derived lemma
  - Derivational rules d: Derivation of derived lemma from basis lemma

$$d = (t, \mathcal{P}_1, \mathcal{P}_2) \quad (1)$$

Transformation functions

- Atomic string edit operations, e.g., *sfx*('*ment*')
- Can be composed into higher-order functions:

$$d = ((sfx(`ness`) \circ try(rsfx(`y`, `i`))), \mathcal{A}, \mathcal{N}) \qquad (2)$$

→ $kind_A$ – $kindness_N$
→ $happy_A$ – $happiness_N$

- Rule induction: Derivation rules in traditional grammar books
  Total implemented rules: 158
  Amount of work: ∼ 22 person-hours

Induction of derivational families

- Input: Set $\mathcal{L}$ of lemma-paradigm pairs *l-p* from lemmatised, POS-tagged SdeWaC with gender information [Schmid, 1994, Faaß et al., 2010, Bohnet, 2010]:
  to respect-V
- Generate possible derivations with derivational rules *d*:
  respect-N, to disrespect-V, respected-A
- Avoid overgeneration: Remove derivations which occur less than 3 times in $\mathcal{L}$:
  ** respectation-N*
- Building the derivational family:
  Transitive closure of all pairs connected by derivation relations

Evaluation setting

- Induction of derivational families: clustering problem
- Similar to semantic class induction [im Walde and Brew, 2002] or coreference resolution [Cardie and Wagstaff, 1999]
  - Several evaluation techniques proposed
  - Our choice: Evaluation of Precision and Recall for pairs of lemmas

Evaluation sampling

- Skewed class distribution: Almost all pairs in $\mathcal{L}$ derivationally unrelated
  $\rightarrow$ Random sampling of pairs is problematic
- Preselection through String Similarity clustering based on Levenshtein distance $\leftrightarrow$ Baseline
- Assumption: Preselection contains all true positive lemma pairs (all lemmas of derivational families):
  cut_N, $to\ cut_V$, $cutting_A$, $cutlery_N$, $cuttlefish_N$, $cute_A$ ...
- Sampling: Draw a pair of lemmas from the same cluster, and compute Precision and Recall
  Total: 2,000 pairs

- N.B.: Due to methodological caution, we carried out a more complex sampling; details in the paper

Sample annotation

- Binary annotation for each lemma pair: derivationally related or not?
  - Positive annotations: semantically and/or morphologically related
  - Negative annotations: no morphological relation, lemmatization errors, compound words
- Inter-Annotator Agreement:
  - Agreement: 0.85
  - Cohen's κ: 0.79

Results I

Method	Precision	Recall
DErivBase	**0.83**	**0.71**
Stemming	0.66	0.07
String distance	0.36	0.20

- DErivBase achieves good precision and substantial recall
- Stemming leads to overclustering $\rightarrow$ low recall
- String similarity achieves more balanced but still poor results

Results II

- Manual analysis: Reliability of the derivational rules
- Three groups of rules:
  - L3 – very reliable
  - L2 – generally reliable
  - L1 – less reliable

Method	Precision	Recall
DErivBase-L123	0.83	**0.71**
DErivBase-L23	0.88	0.61
DErivBase-L3	**0.93**	0.35

Conclusions

- Derivational resources provide knowledge across POSes which is helpful for various NLP tasks
- DErivBase is the first broad-coverage German derivational resource, and publicly available
- Combination of rule-based framework and corpus evidence allows for high accuracy and solid coverage

Thank you for your attention.

Download DErivBase from:
`http://www.cl.uni-heidelberg.de/~zeller/res/derivbase/`

Don't miss our talk **today at 17:05** in **Hall 7**:
Application of **DErivBase for smoothing distributional semantics**

Baayen, H. R., Piepenbrock, R., and Gulikers, L. (1996).
The CELEX Lexical Database. Release 2. LDC96L14.
Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998).
The berkeley framenet project.
In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bohnet, B. (2010).
Top accuracy and fast dependency parsing is not a contradiction.
In Proceedings of the 23rd International Conference on Computational Linguistics, pages 89–97.

Cardie, C. and Wagstaff, K. (1999).
Noun phrase coreference as clustering.
In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 82–89, University of Maryland, MD. Association for Computational Linguistics.

Déjean, H. (1998).
Morphemes as necessary concept for structures discovery from untagged corpora.
In Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, pages 295–298, Sydney, Australia.

Faaß, G., Heid, U., and Schmid, H. (2010).
Design and application of a gold standard for morphological analysis: SMOR in validation.
In Proceedings of the Seventh International Conference on Language Resources and Evaluation, pages 803–810.

Finkler, W. and Neumann, G. (1988).
Morphix - a fast realization of a classification-based approach to morphology.
In Proceedings of 4th Austrian Conference of Artificial Intelligence, pages 11–19, Vienna, Austria.

Fitschen, A. (2004).
Ein computerlinguistisches Lexikon als komplexes System.
PhD thesis, IMS, Universität Stuttgart.

Green, R., Dorr, B. J., and Resnik, P. (2004).
Inducing frame semantic verb classes from wordnet and ldoce.
In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 375–382, Barcelona, Spain.

Habash, N. and Dorr, B. (2003).
A categorial variation database for English.
In Proceedings of the Anuual Meeting of the North American Association for Computational Linguistics, pages 96–102, Edmonton, Canada.

Hammarström, H. and Borin, L. (2011).
Unsupervised learning of morphology.
Computational Linguistics, 37(2):309–350.

im Walde, S. S. and Brew, C. (2002).
Inducing german semantic verb classes from purely syntactic subcategorisation information.
In In Proceedings of the 40th Annual Meeting of the ACL, pages 223–230.

Karttunen, L. and Beesley, K. R. (2005).
Twenty-five years of finite-state morphology.
In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H., and Yli-Jyr, A., editors, Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday, pages 71–83. CSLI Publications, Stanford, California.

Koskenniemi, K. (1983).
Two-level Morphology: A General Computational Model for Word-Form Recognition and Production.
PhD thesis, University of Helsinki.

Schmid, H. (1994).
Probabilistic part-of-speech tagging using decision trees.
In Proceedings of ICNLP, Manchester, UK.

Schmid, H., Fitschen, A., and Heid, U. (2004).
Smor: A German computational morphology covering derivation, composition and inflection.
In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Schone, P. and Jurafsky, D. (2000).
Knowledge-free induction of morphology using latent semantic analysis.
In Proceedings of the Conference on Natural Language Learning, pages 67–72, Lisbon, Portugal.

Šnajder, J. and Dalbelo Bašić, B. (2010).
A computational model of Croatian derivational morphology.
In Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages, pages 109–118, Dubrovnik, Croatia.

Szpektor, I. and Dagan, I. (2008).
Learning Entailment Rules for Unary Templates.
In Proceedings of the 22nd International Conference on Computational Linguistics, pages 849–856, Manchester, UK.

Thadani, K. and McKeown, K. (2011).
Towards strict sentence intersection: Decoding and evaluation strategies.
In Proceedings of the ACL Workshop on Monolingual Text-To-Text Generation, pages 43–53, Portland, Oregon.

Addendum

Statistics of the implemented rules

- 79 noun derivations, 33 verb derivations, 46 adjective derivations
- 6 zero derivations, 106 suffixations, 35 prefixations, 9 stem changes, 2 circumfixations

Addendum

Statistics of the performance of DErivBase

- Total coverage: 280,336 lemmas
- Grouped into 239,680 derivational families:
  17,799 non-singletons covering 58,455 lemmas
  Many singletons are compound nouns
- Biggest 100% precision family: 40 lemmas