Language Technology for Polish in Practice

Morpho-syntactic processing of Polish

Maciej Piasecki
Wrocław University of Science and Technology
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl
2017-01-17

CLARIN-PL
Common Language Resources and Technology Infrastructure

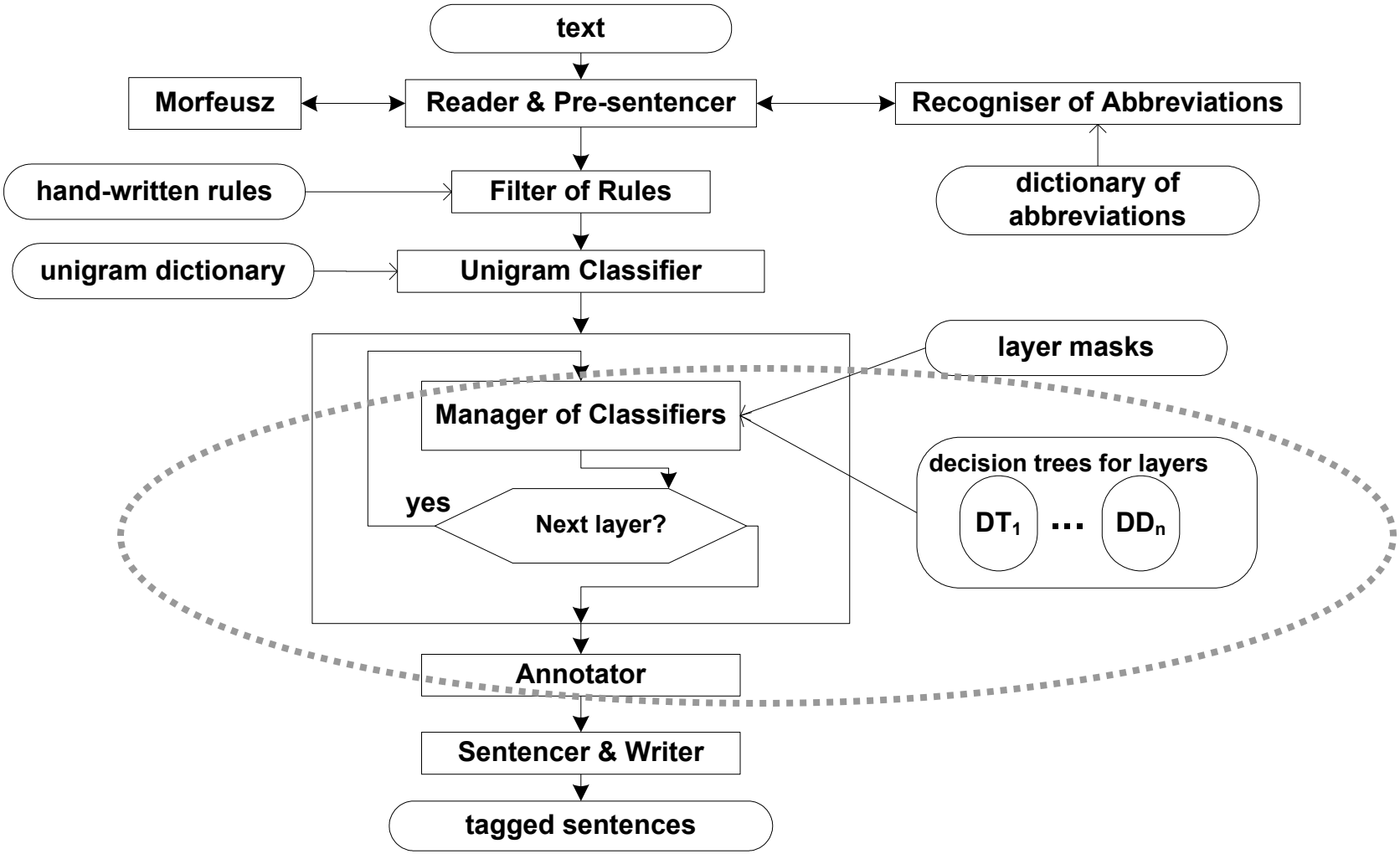
Tools

- Morphological analysis:
  - Morfeusz – IPI PAN (Woliński, 2014)
- Morpho-syntactic analysis
  - TaKIPI – PWr., old but still used (Piasecki, 2007)
  - WCRFT 1 and WCRFT 2 – PWr. (Radziszewski, 2013)
    - Available via CLARIN-PL Web Services
  - Pantera – IPI PAN (Acedański, 2010)
  - Concraft – IPI PAN (Waszczuk, 2012)
- Shallow parsing and chunking
  - Iobber – PWr., trained on KPWr (Radziszewski & Pawlaczek, 2013)
  - Spejd – IPI PAN (Przepiórkowski, 2008)

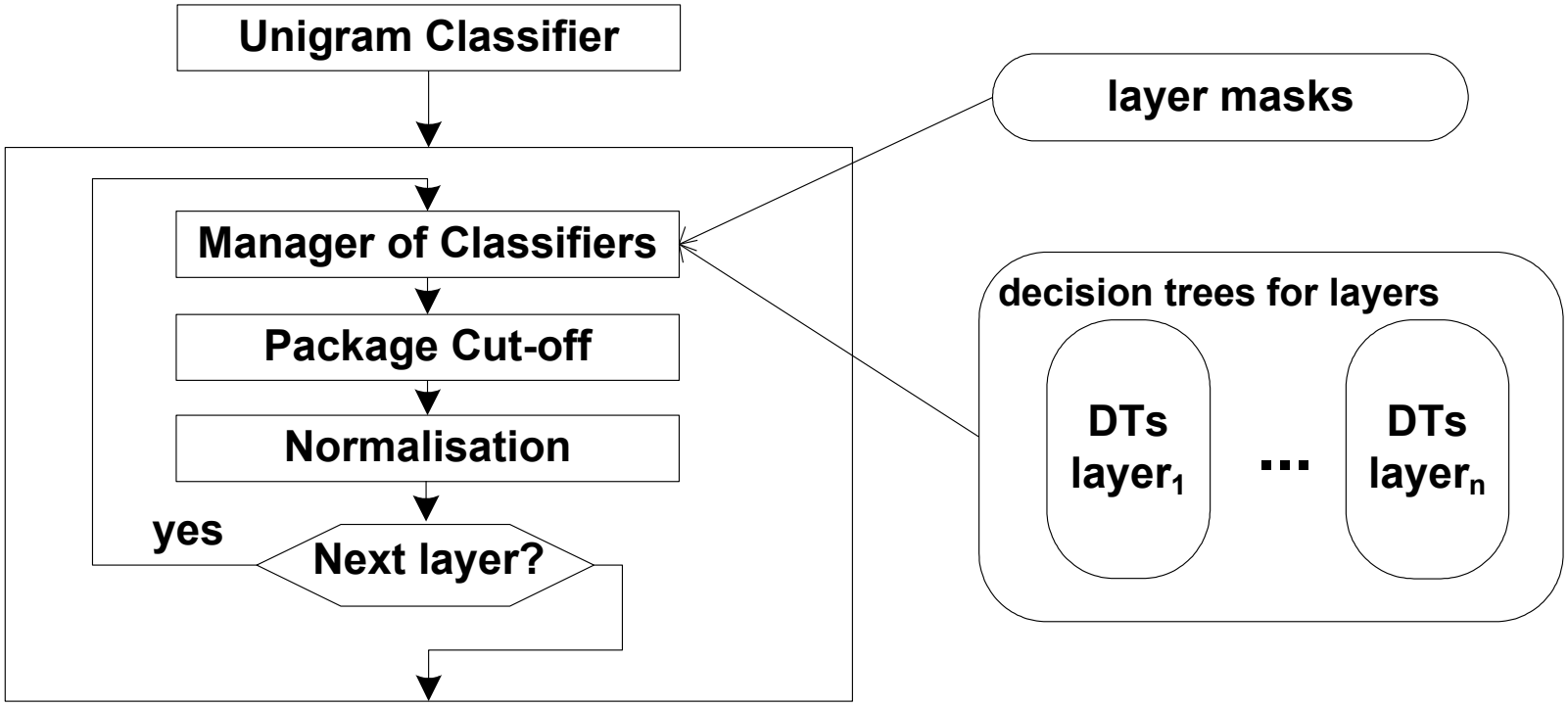
TaKIPI – the first tagger for Polish

- TaKIPI (Piasecki & Godlewski, 2006)
  - name comes from: `a tagger for the IPI PAN Corpus'
  - the first morpho-syntactic tagger for Polish publicly available and in wider use
  - quite fast and easy to install, so still in use
  - partial disambiguation (in some cases more than one tag per word is left), combined with simple morphological guesser
  - ~93% of weak accuracy
- Construction
  - tiered tagging: gradual disambiguation during several phases
  - heterogonous: limited set of rules combined with decision trees

TaKIPI: tiered tagging



TaKIPI: tiered tagging



WCRFT tagger

- WCRFT comes from `Wrocław CRF-based Tagger' (Radziszewski 2013)
  - morpho-syntactic tagger for Polish
  - all words, tags for unknown guessed , but not their lemmas
- Construction
  - tiered tagging in several phases
  - based on supervised learning and NKJP (Polish National Corpus)
    - CRF algorithm (Conditional Random Fields)
    - 1 mln word part of NKJP used for training

Radziszewski, Adam (2013) A tiered CRF tagger for Polish. In R. Bembenik et al. (eds.) Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions. Springer.

WCRFT: feature types

1. word form of a token,
2. possible values of the grammatical class of a token,
3. possible values of grammatical number,
4. possible values of gender,
5. possible values of grammatical case,
6. a predicate checking if there holds a grammatical agreement of the current and the next token with respect to number, gender and case,
7. a similar predicate that checks the agreement of the previous, current and the next tokens (-1, 0, 1),
8. if the current token's orthographic form starts with an upper-case letter,
9. if it starts with lower-case letter.

WCRFT: feature realisation

1. Word forms for *p*(ositions) ∈ {−2,−1,0,1,2}
2. Word form bigrams: (−1, 0) and (0, 1)
3. Grammatical class for $p \in \{-2, -1, 0, 1, 2\}$
4. Class bigrams: (−2, −1), (−1, 0), (0, 1), (1, 2)
5. Class trigrams: $(p - 1, p, p + 1)$ for $p \in \{-2, 0, 1\}$
6. Case: for $p \in \{-2,-1,0,1,2\}$
7. Gender: for $p \in \{-2,-1,0,1,2\}$
8. Number: for $p \in \{-2,-1,0,1,2\}$
9. Agreement: num_gen_cas(−1, 0), num_gen_cas(−1 …1)
10. Agreement: num_gen_cas(−1 …1)
11. Upper case letter: p=0

WCRFT: implementation in brief

- Problems with NKJP (Polish National Corpus) as training source
  - … permanent, to be honest
  - some inconsistences in the morphological annotation
  - Increasing differences between Morfeusz and NKJP
    - output format – from the very beginning
    - tagset, evolving, e.g. new grammatical classes in Morfeusz
    - sets of tags for word forms – increasing lack of synchronisation
- Attempt to decrease their negative influence in WCRFT
  - if the tag assigned manually in NKJP is generated by Morfeusz, it is used
  - otherwise the word is marked as unknown (`ign`) and assigned the tag from NKJP

WCRFT: implementation in brief

- Training
  - Tag sets for unknown words are collected from NKJP
  - performed in tiers (phase): grammatical class, number, case, … (other categories)
    - one attribute per tier (phase)
    - partial disambiguation after each tier: all tags inconsistent with the choice eliminated - ideal decision from NKJP
  - feature templates are used to generate characteristic functions for CRF++
    - WCCL language is used to define feature templates
    - Corpus2 library is used for reading corpus/text
- Tagging
  - in tiers, but partial disambiguation on the basis of the tagger results

WCRFT evaluation

Tagger	Acc_{lower}	Acc_{upper}	Acc^{K}_{lower}	Acc^{U}_{lower}
WMBT	87.50%	87.82%	89.78%	13.57%
PANTERA	88.79%	89.09%	91.08%	14.70%
WMBT+u	89.71%	90.04%	91.20%	41.45%
WCRFT	90.34%	90.67%	91.89%	40.13%

- Evaluated on the manually annotated part of NKJP (average from 10 fold cross-validation)

Bibliography

- Piasecki, M. (2007) Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly, 2007, 11, 151-167. https://pdfs.semanticscholar.org/ec4d/0a95f75d2b9aa883cae3a32444442495254d5.pdf
- Piasecki, M. and Godlewski, G. (2006). Effective architecture of the Polish tagger. In *Text, Speech and Dialogue*, volume 4188, pages 213–220, Brno, Czechy. Springer.
- Radziszewski, Adam (2013) A tiered CRF tagger for Polish. In R. Bembenik et al. Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions, Springer Verlag http://nlp.pwr.wroc.pl/files/publications/wcrft.pdf

WCRFT Project web page:
http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki

- Woliński, Marcin. (2014) Morfeusz Reloaded. In Nicoletta Calzolari (Conference Chair) Khalid Choukri, Thierry Declerck Hrafn Loftsson Bente Maegaard Joseph Mariani Asuncion Moreno Jan Odijk Stelios Piperidis (Ed.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1106–1111, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, ISBN: 978-2-9517408-8-4. http://nlp.ipipan.waw.pl/Bib/wol:14.pdf

Bibliography

- Jakub Waszczuk. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In: Proceedings of COLING 2012, Mumbai, India. http://zil.ipipan.waw.pl/Concraft?action=AttachFile&do=view&target=coling2012.pdf
- S. Acedański (2010) A Morphosyntactic Brill Tagger for Inflectional Languages," in Advances in Natural Language Processing, pp. 3-14. http://ripper.dasie.mimuw.edu.pl/~accek/homepage/wp-content/papercite-data/pdf/ace10.pdf
- Adam Radziszewski, Adam Pawlaczek, (2013) „Incorporating head recognition into a CRF chunker". In: IIS 2013, Warsaw, Poland, June 17-18, 2013. http://nlp.pwr.wroc.pl/pl/publikacje/137/show/publication
- Resource: http://hdl.handle.net/11321/15
- Adam Przepiórkowski. (2008). Powierzchniowe przetwarzanie języka polskiego. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Thank you very much for your attention!

www.clarin-pl.eu

CLARIN-PL
Common Language Resources and Technology Infrastructure

Supported by the Polish Ministry of Science and Higher Education [CLARIN-PL]