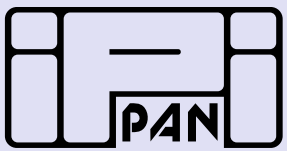
Towards a Weighted Induction Method of Dependency Annotation

Alina Wróblewska & Adam Przepiórkowski



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

PolTAL
Warsaw, 18 September 2014

Outline

Introduction

Supervised dependency parsing

high performance, but manual annotation of training data is a very time-consuming and expensive process

Unsupervised dependency parsing

low performance and high complexity

Cross-lingual dependency projection methods

outperform unsupervised methods of dependency parsing (McDonald et al., 2011)

Cross-lingual dependency projection

Cross-lingual dependency projection

- is a method of annotating sentences with dependency trees in less researched languages,
- builds on the assumption that a dependency tree encoding the predicate-argument structure of a sentence largely carries over to its translation.

Idea behind the dependency projection

- parsing of source sentences in a parallel corpus,
- projecting acquired dependency trees to equivalent target sentences via word alignment links,
- projected dependencies constitute valid dependency structures of target sentences (the ideal case),
- some additional smoothing techniques and aggressive filtering methods are necessary (the practical case).

Outline

Weighted induction method (1)

Method of annotating Polish sentences with dependency trees

- this method is set within the mainstream of the study on dependency projection,
- it builds on the idea of weighted projection (Wróblewska and Przepiórkowski, 2012),
- a weighting factor is involved not only in projecting dependency relations but also in acquiring dependency structures from projected sets of dependency relations.

Weighted induction method

1. weighted projection of dependency relations,
2. induction of unlabelled dependency trees from projected directed graphs (digraphs).

Weighted induction method (2)

Weighted projection

- using a parallel corpus, its English side is automatically annotated with a syntactic parser,
- resulting dependency relations are transferred to equivalent Polish sentences via an extended set of word alignment links – *complete bipartite alignment graph* (BG),
- projected relations constitute digraphs with initially weighted arcs.

Weighted induction

- initial weights are recalculated with the EM selection algorithm,
- maximum spanning trees fulfilling properties of well-formed dependency structures are selected from EM-scored directed graphs.

Complete bipartite alignment graph (BG)

- BG consists of links from different automatic word alignments and some additional links,
- vertices in BG are decomposed into two disjoint sets:
  - a set of English tokens with the ROOT node,
  - a set of Polish tokens with the ROOT node,
- every pair of vertices from these sets is adjacent,
- bipartite edges are weighted,
- an edge weight corresponds to the number of occurrences of the edge in automatic alignment sets,
- word alignment sets:
  - two unidirectional word alignments,
  - a set of bidirectional alignment links symmetrised with the *grow-diag-final-and* heuristic (Koehn, 2010),
- the edge between root nodes is scored with 1.

Weighted projection

Projection procedure

- input: a BG, an English dependency tree and a Polish sentence,
- iterative projection of English arcs to the Polish sentence,
- restriction on projection: it is not possible to project arcs via bipartite edges which are both weighted with 0,
- output: a Polish digraph.

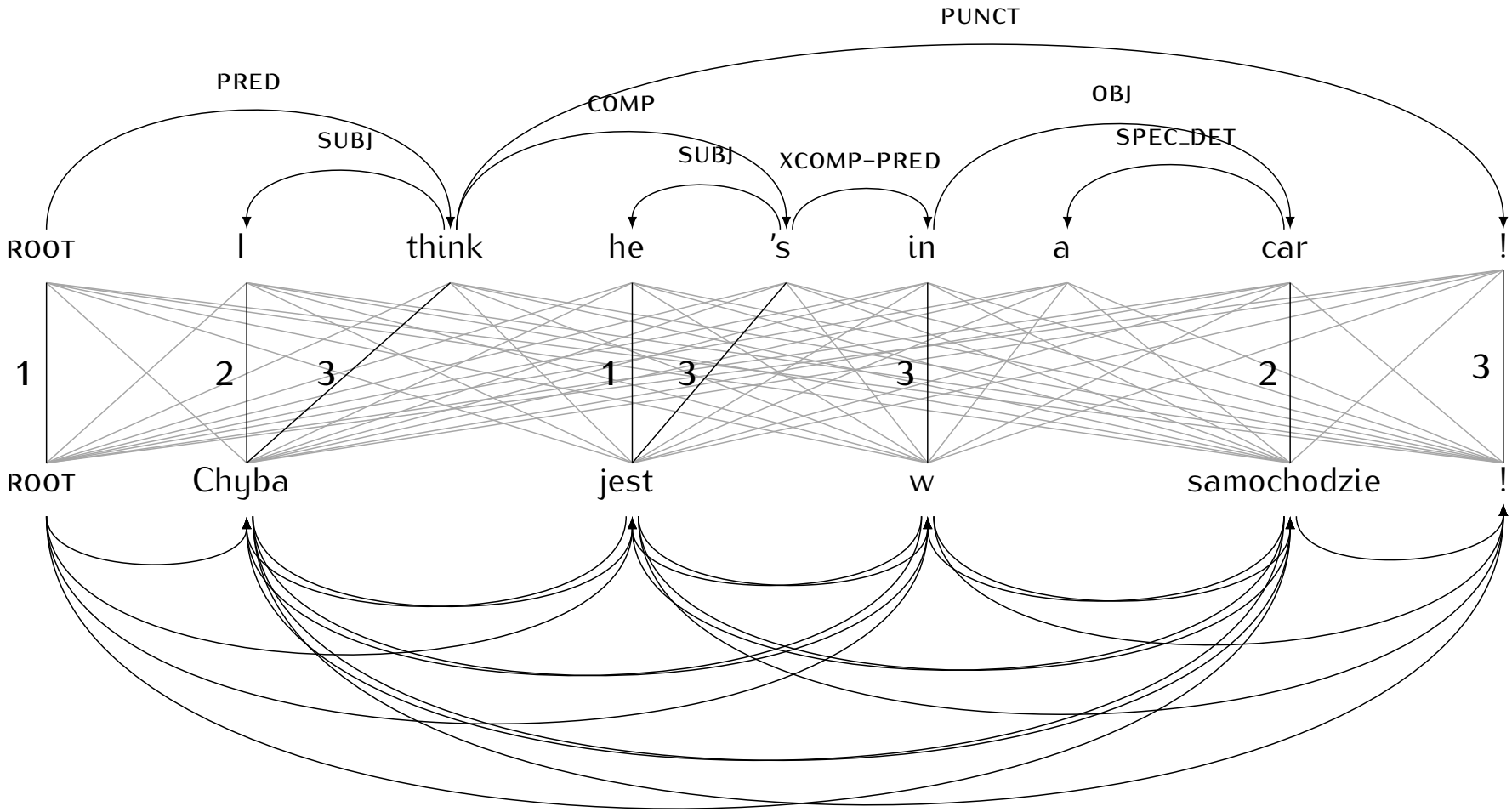
Intuitive weighting of projected arcs

- intuitively, an arc between two tokens might be more important than arcs between other tokens if it is projected via bipartite edges with higher scores,
- an initial weight of a projected arc is estimated:

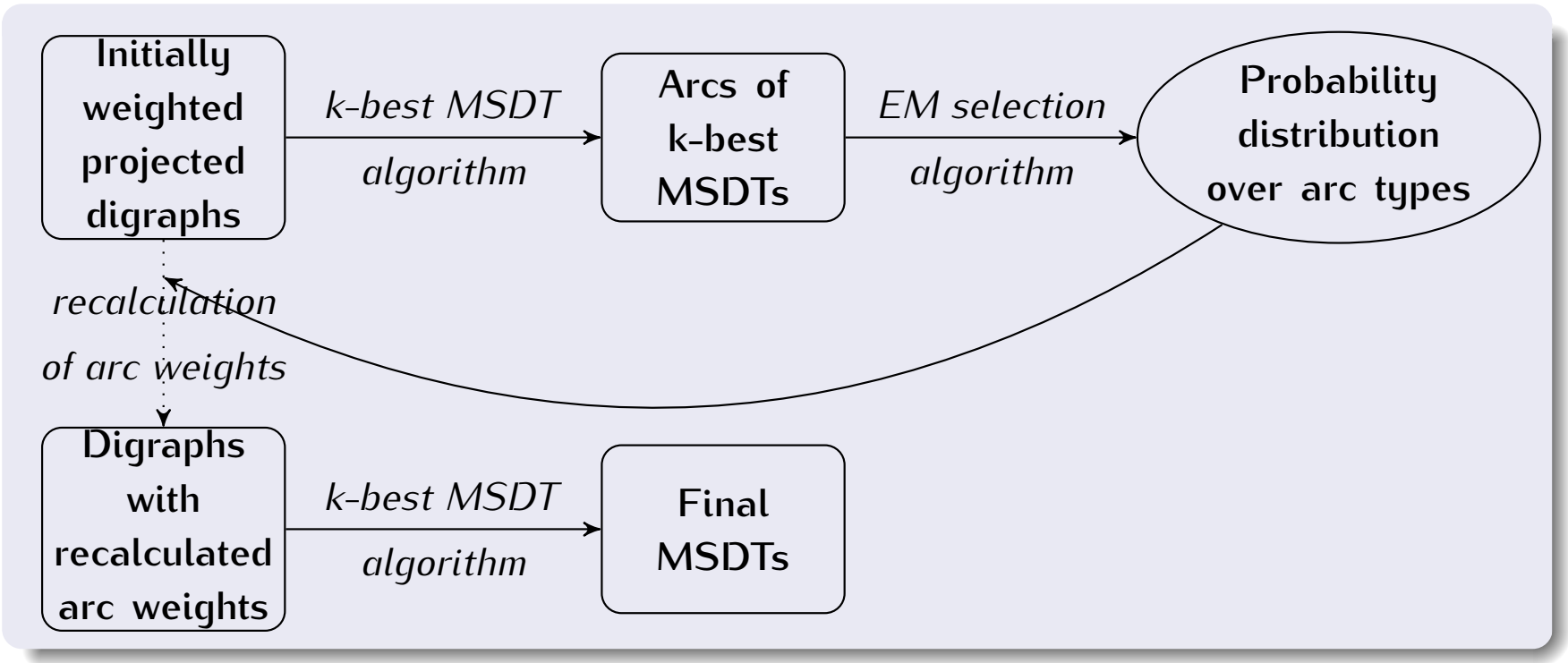
$$s = w_d + w_g + 2w_d w_g f$$

w_d – the dependent link value, w_g – the governor link value, f – the projection frequency.

Schema of the weighted projection procedure



Schema of the weighted induction procedure



Weighted induction

k-best maximum spanning dependency trees (MSDTs)

- k-best MSDTs are selected from projected digraphs using a slightly modified version of the k-best MSTs selection algorithm by Camerini et al. (1980),
- modification: additional conditions are imposed on candidate MSTs so that they meet properties of MSDTs.

Probability distribution over arc types

- we use a version of the EM algorithm defined by Dębowski (2009),
- in its last iteration, the EM selection algorithm estimates the probability distribution $\left(p_j^{(T)}\right)$ over feature representations j of arcs in k-best MSDTs.

Recalculation of arc weights in projected digraphs

- the new weight of an arc with the feature representation j and the initial weight s is calculated as

$$s^* = \sqrt{s} \times p_j$$

- if an arc is not present in any of k-best MSDTs (i.e., its p_j is equal to 0), its new score is calculated as

$$s^* = \sqrt{s} \times \min_j p_j \times \alpha, \qquad \text{for some } 0 < \alpha < 1$$

- arcs with the probability greater than zero are rewarded and other arcs are penalised,
- arcs with higher weights are more likely to be selected as part of final dependency trees.

Outline

Data and preprocessing

Polish–English parallel corpus

the experiment is conducted on a large collection of bitexts from *Europarl* (Koehn, 2005), *DGT-Translation Memory* (Steinberger et al., 2012), *OPUS* (Tiedemann, 2012) and *Pelcra Parallel Corpus* (Pęzik et al., 2011).

Word alignment

automatic word alignment links are generated with the statistical machine translation system MOSES (Koehn et al., 2007).

English parser

the English side of the corpus is parsed with the handcrafted wide-coverage English LFG (Dalrymple, 2001; Bresnan, 2001), using XLE as a processing platform.

Automatic induction of Polish dependency trees

Projection module

- input: 3 sets of word alignment links for each sentence pair, English dependency trees, and Polish sentences,
- output: almost 5 million initially weighted Polish digraphs.

Induction module

- extracts over 4.6 million sets of k-best MSDTs (for $k = 10$),
- estimates the probability distribution over arc types in these k-best MSDTs,
- recalculates initial arc weights in projected digraphs,
- acquires final MSDTs from these digraphs.

Labelling module

- assigns Polish dependency labels to arcs in induced trees (almost 4 million) in a rule-based fashion.

Evaluation experiment

Evaluation

- an extrinsic evaluation shows to what extent induced trees affect performance of a parser trained on them,
- the *Mate* dependency parsing system (Bohnet, 2010),
- evaluation metrics: *unlabelled attachment score* (UAS) and *labelled attachment score* (LAS).

Sets of test trees

- *manual test* – 822 dependency trees taken from the Polish dependency treebank,
- *automatic test* – 822 trees from previous test with automatically generated POS tags and morphological features,
- *additional test* – 100 relatively complex trees.

Evaluation results

model	data	manual test		automatic test		additional test	
		UAS	LAS	UAS	LAS	UAS	LAS
induced	3.9M	73.7	–	72.8	–	63.5	–
labelled	3.9M	74.6	69.4	74.0	68.1	63.7	58.3
modified	3.9M	85.1	79.2	84.0	77.3	74.3	68.5
filtered	2.3M	86.0	80.5	84.7	78.3	**76.1**	**70.3**
supervised	7.4K	**92.7**	**87.2**	**88.4**	**81.0**	76.0	69.5

Sets of Polish dependency trees used in Mate training:

- *induced* – acquired with the weighted induction method,
- *labelled* – induced and labelled,
- *modified* – labelled and modified with correction rules,
- *filtered* – labelled, modified and filtered,
- *supervised* – trees from the Polish dependency treebank.

Outline

Conclusion

- we presented a novel weakly supervised method of obtaining Polish dependency structures,
- results are mostly a little below the performance of the supervised parser, when tested on treebank data,
- a test against real data shows that a projection-based parser may exceed the supervised upper bound,
- as this projection-based result was achieved with much less manual work than in the supervised scenario, the method described here rivals the supervised scenario.

Conclusion

- we presented a novel weakly supervised method of obtaining Polish dependency structures,
- results are mostly a little below the performance of the supervised parser, when tested on treebank data,
- a test against real data shows that a projection-based parser may exceed the supervised upper bound,
- as this projection-based result was achieved with much less manual work than in the supervised scenario, the method described here rivals the supervised scenario.

Thank you for your attention!

References

- Bohnet, B.: Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 89–97 (2010)
- Bresnan, J.: Lexical-Functional Syntax. Blackwell, Oxford (2001)
- Camerini, P.M., Fratta, L., Maffioli, F.: The K Best Spanning Arborescences of a Network. Networks 10, 91–110 (1980)
- Dalrymple, M.: Lexical-Functional Grammar. Syntax and Semantics, vol. 34. Academic Press (2001)
- Dębowski, Ł.: Valence extraction using EM selection and co-occurrence matrices. Language Resources and Evaluation 43(4), 301–327 (2009)
- Ganchev, K., Gillenwater, J., Taskar, B.: Dependency Grammar Induction via Bitext Projection Constraints. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1. pp. 369–377 (2009)
- Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., Kolak, O.: Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering 11(3), 311–325 (2005)
- Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit Conference. pp. 79–86 (2005)
- Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 177–180 (2007)
- McDonald, R., Petrov, S., Hall, K.B.: Multi-Source Transfer of Delexicalized Dependency Parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 63–72 (2011)
- Pęzik, P., Ogrodniczuk, M., Przepiórkowski, A.: Parallel and spoken corpora in an open repository of Polish language resources. In: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. pp. 511–515 (2011)
- Smith,D.A.,Eisner,J.: Parser Adaptation and Projection with Quasi-Synchronous Grammar Features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 822–831 (2009)

References (1)

- Søgaard, A.: Data point selection for cross-language adaptation of dependency parsers. In: Proceedings of the 49th Annual Meeting of the Association for Com- putational Linguistics: Human Language Technologies: Short Papers – Volume 2. pp. 682–686 (2011)
- Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: DGT-TM: A freely Available Translation Memory in 22 Languages. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. pp. 454–459 (2012)
- Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. pp. 2214–2218 (2012)
- Wróblewska, A.: Polish Dependency Bank. Linguistic Issues in Language Technology 7(1), 1–15 (2012)
- Wróblewska, A., Przepiórkowski, A.: Induction of Dependency Structures Based on Weighted Projection. In: Proceedings of the 4th International Conference on Computational Collective Intelligence Technologies and Applications, Part I. LNAI, vol. 7653, pp. 364–374. Springer-Verlag, Berlin (2012)
- Wróblewska, A., Przepiórkowski, A.: Projection-based Annotation of a Polish Dependency Treebank. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. pp. 2306–2312. ELRA, Reykjavík, Iceland (2014)
- Zeman, D., Resnik, P.: Cross-Language Parser Adaptation between Related Languages. In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. pp. 35–42 (2008)