A one-pass valency-oriented chunker for German

LTC'13

Adrien Barbaresi

ICAR Lab / ENS Lyon

Poznań – December 7, 2013

Outline

Use of finite-state automata to approximate a grammar

Early 90s

After a few decades of work on more powerful grammars due to the "persuasiveness of syntactic structures"[a]

Work of Pereira (1990): "Finite-state approximations of grammars"

Notion of chunk parsing (Abney 1991)

[a] Karttunen, Lauri, 2001. Applications of Finite-State Transducers in Natural Language Processing. In S. Yu and A. Paun (eds.), CIAA 2000, LNCS 2088. Heidelberg: Springer, pp. 34–46.

The application side: Information extraction

Automata which do not yield full parses but rather a series of indications obtained faster.

FASTUS (Hobbs et al. 1997)
Also the case concerning German (Neumann et al. 1997)

Transducers for German

Kermes and Evert (2002) as well as Schiehlen (2003) use several levels of parsing to achieve a better precision.

> *"It turns out that topological fields together with chunked phrases provide a solid basis for a robust analysis of German sentence structure*[1]*"*

Complete overview in the doctoral thesis of Müller (2007): finite-state parsers are quite efficient, although they do not perform well on certain types of clauses.

[1] Hinrichs, E. W., 2005. Finite-State Parsing of German. In Antti Arppe and et al. (eds.), *Inquiries into Words, Constraints and Contexts*. Stanford: CSLI Publications, pp. 35–44.

Interest of a valency-oriented tool

Hypothesis: use the strengths of the FST and exploiting the irregularities in the output from NLP tools in order to detect linguistic phenomena

- Readability and text quality assessment: isolation of difficult parts of a text, syntactical complexity (simulate "parse tree depth features")
- Non-standard text analysis: learner or web corpora
- Detection of irregularities: quality assessment of quality of POS-tagger output, creation of selective benchmarks for tools

Part of annotation techniques designed to help qualify texts, provide a "reasonable" image of text complexity

Paper: Approximation de la complexité perçue, méthode d'analyse.

In *Actes TALN'2011/RECITAL*.

State of the art of this particular processing step

FASTUS approach

Analysis of basic phrases: sentences are segmented into noun groups, verb groups, and particles, + complex noun and verb groups are identified

Sundance approach

Segmentation part of the Sundance shallow parser
(Riloff and Phillips 2004)

Voss (2005)

Shallow parsing seen as the detection of indicators of phrase structure without necessarily constructing that full structure

Characteristics of valency-oriented phrase chunking

Grouping into possibly relevant chunks enables a valency detection for each verb based on topological fields (Reis 1980).

- intra-propositional side: syntactic complexity of the groups (and possibly grammatically relevant phrases)
- propositional side: complementation of the verbs and topological nature of a phrase

Objective: yield various kinds of linguistic information useful to the language researcher

Characteristics of one-pass processing

- Aims at robustness
- Linear approach
- Fine-tuning and hand-crafted rules: chunker limited to German
- Pattern-based matching of POS-tags using regular expressions (which are themselves finite-state automata)
  Uses the STTS tagset (Schiller et al. 1995)
- Ecosystem with POS-tagger: decisions in common situations are (at least statistically) known

Implementation (NP and PP)

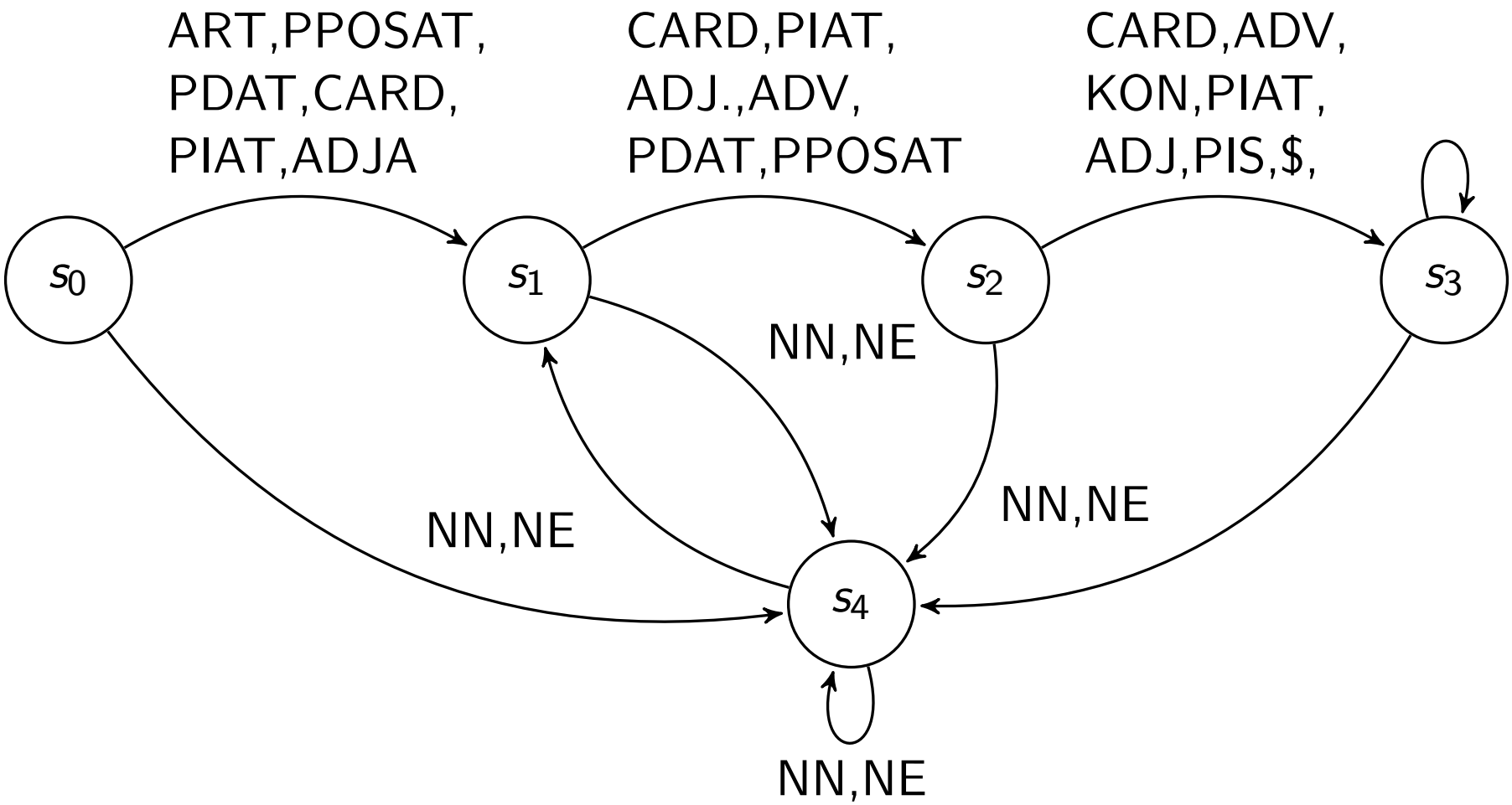
Automaton

Uses POS-tags.
The transducer can go through several states and add tokens to the chunk according to certain transition rules.
Final step: common or proper noun (*NN* or *NE*) that is not followed by a word which could be possibly linked to the chunk, like another noun or a tag which leads to the first state.

Right attachments

The head of the phrase is supposedly on the right of the group.
The pattern is greedy: everything that fits under a predefined composition of a phrase counts.

Simplified pattern used for detecting noun phrases



NB: APPRART and APPR tags required to initiate detection of prepositional phrases

Valency estimation

Gives an estimation of the number of arguments that may be syntactically connected to a given verb.

$\rightarrow$ Find the boundaries of the clauses (case of German, importance of commas).

$\rightarrow$ Each head of a chunk found in a given clause increments the actual valency variable.

Example

Überfüllte	Einzimmerbehausungen	,	moderne	Apartments	oder	Kolonialvillen
NP0	NP3		NP0	NP3	NP3-R	NP3-R
	1			*1*		1

im	französischen	Viertel	–	der	Fotokünstler	Hu	Yang	versucht
PP0-R	PP1-R	PP3-R		NP0	NP3	NP3-R	NP3-R	VP
				2			*2*/1	

mit	seinen	Bildern	,	möglichst	viele	Facetten	seiner	Heimatstadt	einzufangen	.
PP0	PP1	PP3			NP0	NP3	NP1-R	NP3-R	VP	
		3/2				**1**				

Level 1: Sentence text, phrases underlined
Level 2: Chunker output
Level 3: Valency counter (*black*), and gold standard (**correct**/mistake)

NP, PP and VP are phrase types, the numbers are states. "R" → extension on the right detected.

Statistical evaluation

Corpus

2,416 recent online articles, German version of the *Geo* magazine[a]
838,790 tokens

[a] http://www.geo.de

Output statistics

469,655 non-verbal tokens
234,120 verbal tokens (verbs + modifiers)
92,680 punctuation tokens
About 6 % of the tokens are potentially words without possible connections
547,686 non-verbal tokens in total had a chance to be analyzed
14 % missing: this information could be used to detect difficulties

Evaluation in detail

3 different samples of 1,000 tokens in a row extracted from the corpus

Output	Errors	Missed	Precision	Recall
831	95	87	.886	.894

The majority of errors are linked to tokenization and tagging artifacts.

Conclusion

- Linear approach, uses a bottom-up linguistic model implemented using finite-state automata.
- The trade-off seems to be justifiable.
- A possible application is what both metrics do not show, what it could not integrate or analyze successfully: focus on complex phrases or sentences, and on irregularities in a corpus.
- Future work:
  $\rightarrow$ Metrics for actual valency detection and error analysis.
  $\rightarrow$ Integration of more precise morphosyntactic information.

References

Abney, S. P., 1991. Parsing by chunks. *Principle-based parsing*, 44:257–278.

Barbaresi, A., 2011. Approximation de la complexité perçue, méthode d'analyse. In *Actes TALN'2011/RECITAL*

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M., 1997. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *Finite-State Language Processing*:383–406.

Kermes, H. and Evert, S. 2002. YAC – A Recursive Chunker for Unrestricted German Text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, vol. 5.

Neumann, G., Backofen R., Baur J., Becker M., and Braun C., 1997. An Information Extraction Core System for Real World German Text Processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics.

Pereira, F., 1990. Finite-state approximations of grammars. In *Proceedings of the Annual Meeting of the ACL*.

Riloff, E. and Phillips, W., 2004. An Introduction to the Sundance and AutoSlog Systems. Technical report, School of Computing, University of Utah.

Schiehlen, M., 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th conference of the EACL*, vol. 2.

Voss, M. J., 2005. Determining syntactic complexity using very shallow parsing. Master's thesis, CASPR, Artificial Intelligence Center, University of Georgia.

Contact: adrien.barbaresi@ens-lyon.fr

http://perso.ens-lyon.fr/adrien.barbaresi/

https://github.com/adbar/valency-oriented-chunker