Profiling the XDG Constraint Solver

Ralph Debusmann

Programming Systems Lab, Saarbrücken

PS-Lab Colloquium, March 23th, 2005

Overview

1. Introduction
2. Profiling
3. Numbers
4. Interpretation
5. Future Work

Overview

1 Introduction
2 Profiling
3 Numbers
4 Interpretation
5 Future Work

Extensible Dependency Grammar (XDG)

- grammar formalism for natural language (Debusmann, Duchier, Koller, Kuhlmann, Smolka, Thater 2004)
- parsing/realization as finite domain/finite set constraint solving in Mozart/Oz (Duchier 1999, Duchier 2003)
- system: XDG Development Kit (XDK) (Debusmann, Duchier, Niehren 2004)

Very expressive

- German word order (Duchier and Debusmann 2001)
- relational syntax-semantics interface (Debusmann, Duchier, Koller, Kuhlmann, Smolka, Thater 2004)
- phonology-semantics interface (information structure) (Debusmann, Postolache, Traat 2005)

Complexity

- precursor of XDG: Topological Dependency Grammar (TDG) is NP-complete (Koller, Striegnitz 2002)
- XDG is also NP-complete (Duchier p.c.), diss
- NP-hardness proof by reduction of HAMILTONIAN-PATH

Efficiency

- efficient for smaller-scale grammars (mostly handcrafted)
- does not yet scale up to large-scale grammars (induced from treebanks):

Examples

- Czech, Prague Dependency Treebank (Bojar 2004)
- English, Penn Treebank (Dienes, Koller, Kuhlmann 2003), (Narendranath 2004)
- German, TIGER Treebank (Korthals 2003), (Möhl 2004)

Overview

1. Introduction
2. Profiling
3. Numbers
4. Interpretation
5. Future Work

Profiling

- idea: find out reasons for efficiency breakdowns
- new XDK functionality:
  - count constraint variables (finite domain, finite set)
  - count propagators
  - average lexical ambiguity (entries per word)
  - more statistics

First Results

- handcrafted grammars: can be very efficient if propagation is complete
- induced grammars: different sources for efficiency breakdowns:

Reasons

- English, (Dienes, Koller, Kuhlmann 2003), (Narendranath 2004): too unrestricted (valency, word order)
- Czech, (Bojar 2004): too many lexical entries, therefore too many constraint variables and propagators
- German, (Korthals 2003), (Möhl 2004): not yet profiled

- 64 words barrier (Mozart)

Overview

1. Introduction
2. Profiling
3. Numbers
4. Interpretation
5. Future Work

Smaller-scale grammar (2D) 1

- German, (Debusmann 2001), min, max, average:

Words	*Time (s)*	*Sols/Fails*	*Lex. Amb.*	*Vars*	*Props*
3	0.030	0/0	1	603	2276
64	4.950	6/2	9	14633	339121
7.9	0.169	1.14/0.36	2.12	1929	14673

Smaller-scale grammar (2D) 2

- German, (Debusmann 2001), specific sentences:

Words	*Time (s)*	*Vars*	*Props*
11	0.21	2515	15148
20	0.46	4549	40834
29	0.99	6583	78508
38	1.49	8617	128170
50	2.00	11329	213034
63	4.63	14592	329345

Smaller-scale grammar (5D)

- English, (Debusmann, Duchier, Koller, Kuhlmann, Smolka, Thater 2004), min, max, average:

Words	*Time (s)*	*Sols/Fails*	*Lex. Amb.*	*Vars*	*Props*
5	0.270	0/0	1	12842	41058
15	4.840	42/46	44	79680	285656
7.2	0.714	5.04/2.14	3.08	25985.8	82683.2

Large-scale grammar (2D)

- Czech, (Debusmann, Duchier, Koller, Kuhlmann, Smolka, Thater 2004), min, max, average:

Lex. Amb.	*Vars*	*Props*
1	10085	11092
1472	734995	1005552
36.32	171259	218557

Large-scale grammar (2D)

- Czech, (Debusmann, Duchier, Koller, Kuhlmann, Smolka, Thater 2004), specific sentences:

Words	*Lex. Amb.*	*Vars*	*Props*
7	21.43	43125	50697
10	20.9	60211	74506
11	28.4	89335	106335
17	56.06	270668	308770
21	62.3	371632	428442
33	54.36	511277	646483
45	24.9	326388	573398

Overview

1. Introduction
2. Profiling
3. Numbers
4. Interpretation
5. Future Work

Problems

- too many lexical entries
- too many constraint variables
- constraint variables/propagators ratio
- unconstrained induced grammars
- incomplete propagation

Possible Solutions

- optimized propagation
- polynomial fragments of XDG
- modelling: Treebank to XDG, XDG to Mozart/Oz
- search: distribution strategy, guided search (Dienes, Koller, Kuhlmann 2003), (Narendranath 2004)
- supertagging
- Gecode (Schulte, Tack)

Overview

1. Introduction
2. Profiling
3. Numbers
4. Interpretation
5. Future Work

Future Work

- too many lexical entries: supertagging, modeling (Treebank to XDG)
- too many constraint variables: modelings (Treebank to XDG, XDG to Mozart/Oz)
- constraint variables/propagators ratio: modelings
- unconstrained induced grammars: search (distribution strategy, guided search)
- incomplete propagation: optimized propagation, polynomial fragments of XDG, modelings, search, Gecode

References

Ondrej Bojar.
Problems of Inducing Large Coverage Constraint-Based Dependency Grammar.
In *Proceedings of the International Workshop on Constraint Solving and Language Processing*, Roskilde/DK, 2004.

Ralph Debusmann, Denys Duchier, Alexander Koller, Marco Kuhlmann, Gert Smolka, and Stefan Thater.
A Relational Syntax-Semantics Interface Based on Dependency Grammar.
In *Proceedings of COLING 2004*, Geneva/CH, 2004.

References

- Ralph Debusmann, Denys Duchier, and Joachim Niehren.
The XDG Grammar Development Kit.
In *Proceedings of the MOZ04 Conference*, volume 3389 of *Lecture Notes in Computer Science*, pages 190–201, Charleroi/BE, 2004. Springer.

- Ralph Debusmann, Oana Postolache, and Maarika Traat.
A Modular Account of Information Structure in Extensible Dependency Grammar.
In *Proceedings of the CICLING 2005 Conference*, Mexico City/MEX, 2005. Springer.

References

- Peter Dienes, Alexander Koller, and Marco Kuhlmann.
  Statistical A* Dependency Parsing.
  In *Prospects and Advances in the Syntax/Semantics Interface*, Nancy/FR, 2003.

- Denys Duchier.
  Axiomatizing Dependency Parsing Using Set Constraints.
  In *Proceedings of MOL 6*, Orlando/US, 1999.

- Denys Duchier.
  Configuration of Labeled Trees under Lexicalized Constraints and Principles.
  Research on Language and Computation, 1(3–4):307–336, 2003.

References

- Denys Duchier and Ralph Debusmann.
  Topological Dependency Trees: A Constraint-Based Account of Linear Precedence.
  In *Proceedings of ACL 2001*, Toulouse/FR, 2001.
- Alexander Koller and Kristina Striegnitz.
  Generation as Dependency Parsing.
  In *Proceedings of ACL 2002*, Philadelphia/US, 2002.
- Mathias Möhl.
  Modellierung natürlicher sprache mit hilfe von topologischer dependenzgrammatik.
  Technical report, Saarland University, 2004.
  Forschungsprojekt.

References

Renjini Narendranath.
Evaluation of the Stochastic Extension of a Constraint-Based Dependency Parser.
Technical report, Saarland University, 2004.
Bachelorarbeit.

Thank you!