Incremental Tree Substitution Grammar for Parsing and Sentence Prediction

Federico Sangati and Frank Keller
<federico.sangati@gmail.com> <keller@inf.ed.ac.uk>
School of Informatics, University of Edinburgh

Introduction

Incrementality and Sentence Prediction

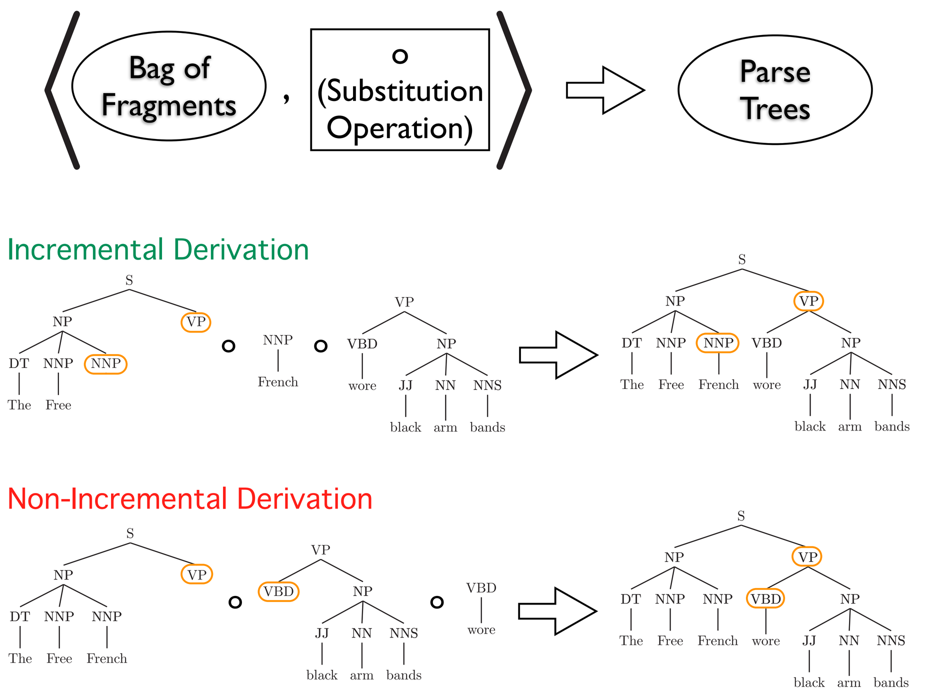
- Human sentence processing is incremental (Demberg and Keller, 2008).
- Predicts upcoming words (Grabski and Scheffer, 2004).
- Assumes strongly lexicalized parsing model.

Incremental TSG Model (ITSG)

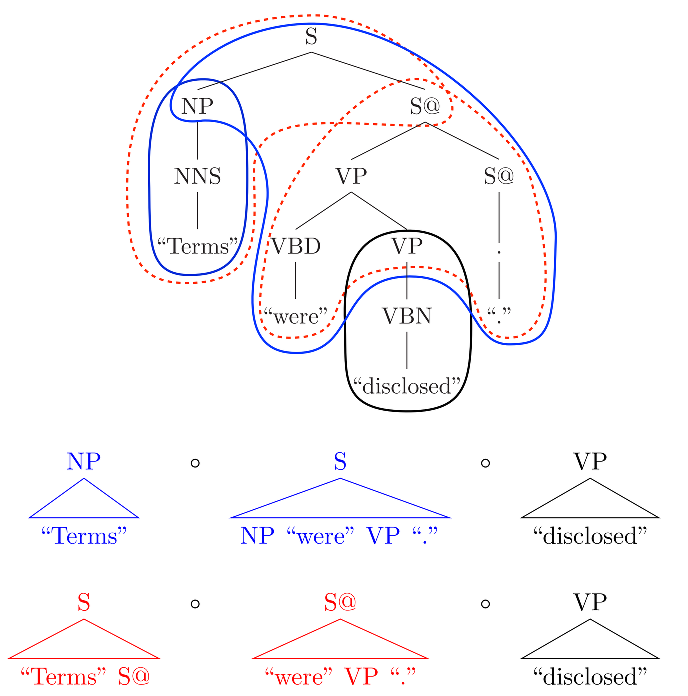
Incremental TSG Generative Process

- Arbitrarily large fragments as in TAG/TSG (Schabes, 1990; Bod et al., 2003).
- Fully connected incremental structure.
- Left-Right generative process (each step must extend the prefix).
- More constraints on the fragments (lexical anchor in first or second position).

Incremental TSG Derivations



Multiple Derivations

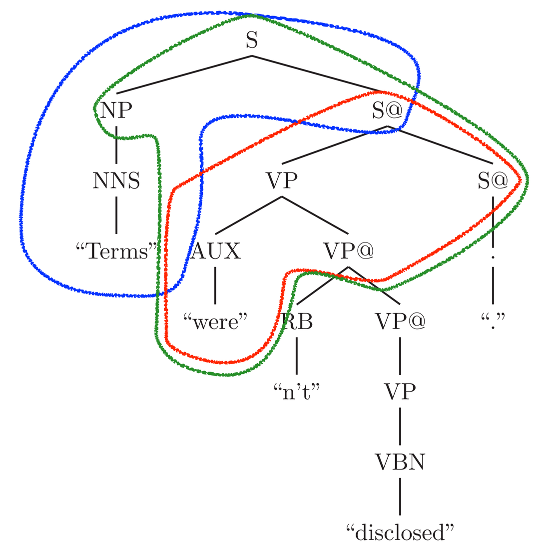


Fragment Types

3 fragment types:

- *Initial*: lexical anchor in the first position (sentence initial).
- *Lex-First*: lexical anchor in first position (non sentence-initial).
- *Sub-First*: lexical anchor in second position, and a substitution site in first.

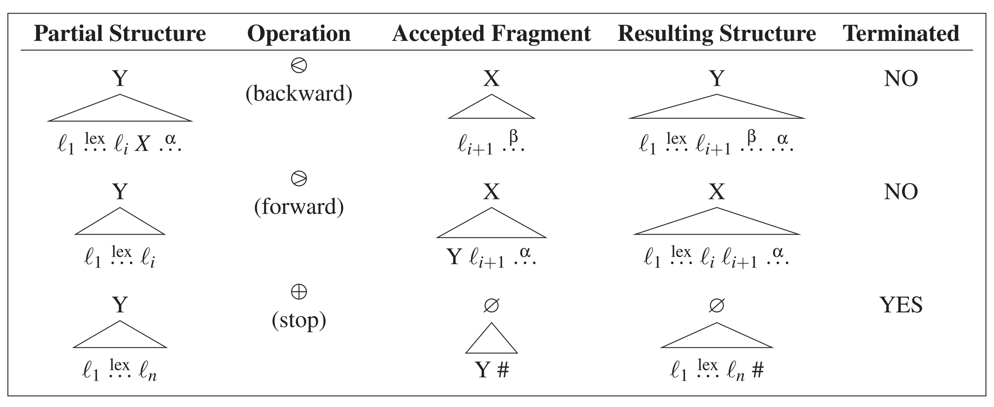
Type	Fragment	Horizontal Notation
Initial	S / "Terms" S@	S ◁ "Terms" S@
Lex-First	S@ / "were" VP@ S@	S ◁ "were" VP@ S@
Sub-First	S / NP "were" VP@ S@	S ◁ "were" VP@ S@



Combining Operations

4 operations (+ START, STOP)

- Scan
- Backward Substitution
- Forward Substitution
- Complete



Probabilistic Chart Parser

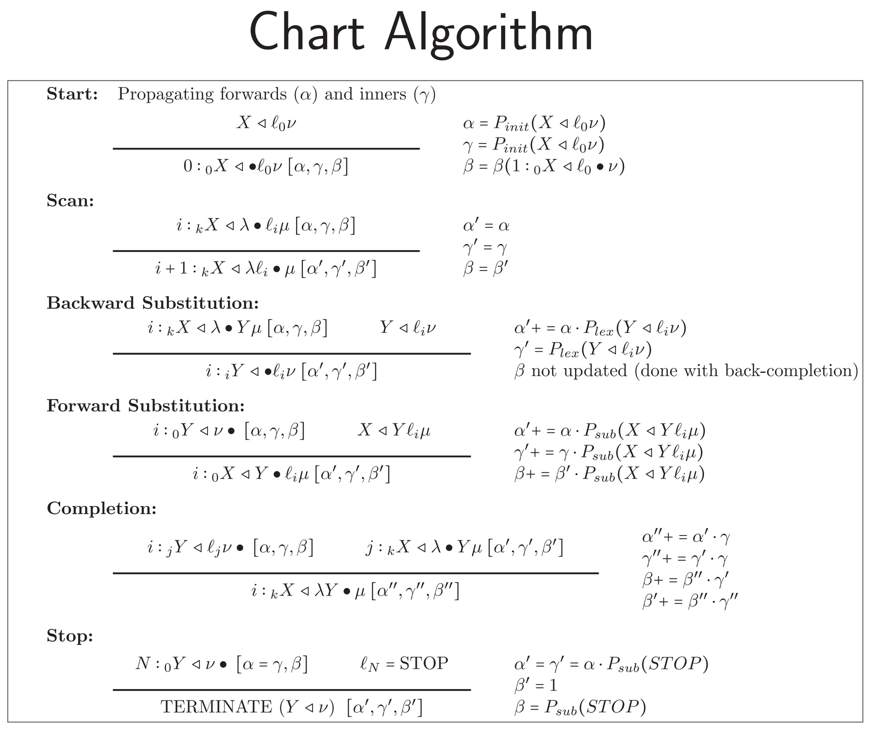
Earley based (Earley, 1970; Stolcke, 1995). No cycles.

Prob. State $i : {}_k X \triangleleft \lambda \bullet \mu \quad [\alpha, \gamma, \beta, \leftarrow]$

- α : forward probability $P(\ell_0^{i-1}, i : {}_k X \triangleleft \lambda \bullet \mu)$
- γ : inner probability $P(\ell_k^{i-1}, i : {}_k X \triangleleft \lambda \bullet \mu)$
- β : outer probability $P(\ell_0^{k-1}, \ell_i^N, i : {}_k X \triangleleft \lambda \bullet \mu)$
- $\leftarrow$: viterbi best previous state

Frag. Type	Fragment	Symbol	Probability
Initial	X / $\ell_0 \ldots$	π_{init}	$\frac{f(\pi_{init})}{\sum_{\pi'_{init}} f(\pi'_{init})}$
Lex-First	X / $\ell_i \ldots$	$\pi_{lex(X)}$	$\frac{f(\pi_{lex(X)})}{\sum_{\pi'_{lex(X)}} f(\pi'_{lex(X)})}$
Sub-First	X / Y $\ell_i \ldots$	$\pi_{sub(Y)}$	$\frac{f(\pi_{sub(Y)})}{\sum_{\pi'_{sub(Y)}} f(\pi'_{sub(Y)})}$

Chart Algorithm



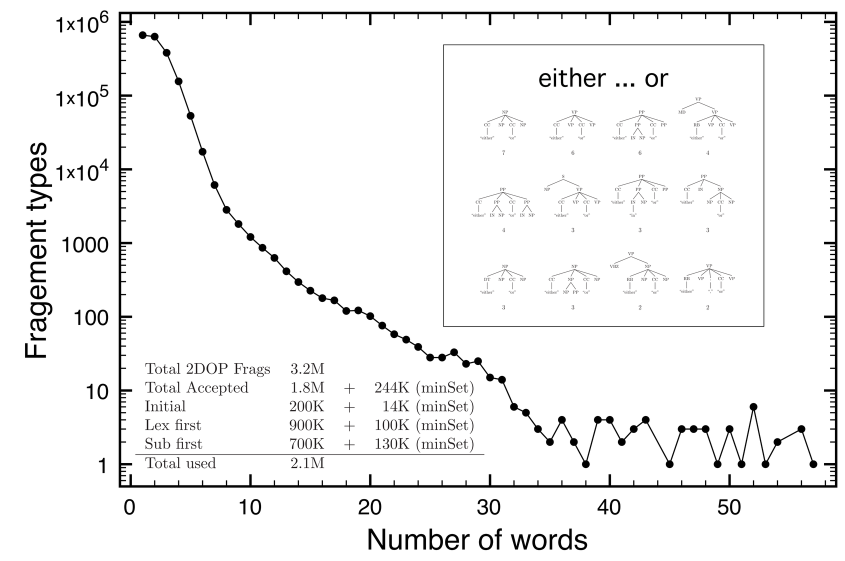
Experimental Setup

Corpus Setup

- Penn WSJ Treebank (Marcus et al., 1993).
- Removing traces and functional tags.
- Apply right binarization (Klein and Manning, 2003), with no horizontal and vertical conditioning (H0V1).
- Replace words appearing < 5 times in the train with lexical features (Petrov, 2009).

Grammar Extraction

- Use *FragmentSeeker* (Sangati et al., 2010).
- Remove all non valid frags (e.g., no lexical items).
- Add all one-word fragments (minSet).
- Count frag frequency in the training corpus.



Parsing

- MPD: Maximum Probable (partial) Derivation.
- MPP: Maximum Probable Parse (approximated).
- MRP: Minimum Risk Parse (Goodman, 1996).

Evaluation

Metrics

- Standard Parsing Evaluation (full sentences).
- Incremental Parsing Evaluation: for each prefix of the input sentence we compute the parsing accuracy on the minimal structure spanning that prefix.
- Sentence Prediction (for every prefix).
  - Word prediction PRD(m): whether the m predicted words are correct.
  - Word presence PRS(m): whether the m predicted words are present in the same order.
  - Longest common subsequence LCS: computes the LCS between the sequence of predicted words and the words following the prefix in the original sentence.

Other Models

- Demberg et al. (2014) [Standard Parsing]
- Schuler et al. (2010) [Standard Parsing]
- Roark (2001); Roark et al. (2009) [Standard Parsing, Incremental Parsing]
- 3-gram model using SRILM (Stolcke, 2002) [Sentence Prediction]

Results

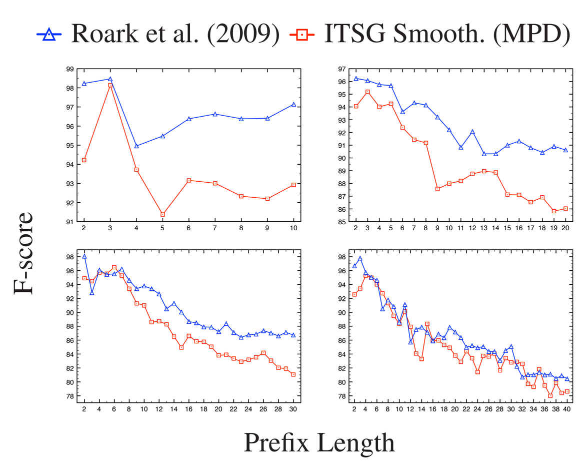
Standard Parsing

	R	P	F1
Demberg et al. (2014)	79.4	79.4	79.4
Schuler et al. (2010)	83.4	83.7	83.5
Roark (2001)	86.6	86.5	86.5
Roark et al. (2009)	**87.7**	**87.5**	**87.6**
ITSG (MPD)	81.5	83.5	82.5
ITSG (MPP)	81.6	83.6	82.6
ITSG (MRP)	82.6	85.8	84.1
ITSG Smoothing (MPD)	83.0	83.5	83.2
ITSG Smoothing (MPP)	83.2	83.6	83.4
ITSG Smoothing (MRP)	**83.9**	**85.6**	**84.8**

Sentence Prediction

	ITSG			3-gram LM (SRILM)		
	Correct	R	P	Correct	R	P
PRD(1)	4,637	8.7	12.5	11,430	**21.5**	**21.6**
PRD(2)	864	1.7	**13.9**	2,686	**5.3**	5.7
PRD(3)	414	0.9	**20.9**	911	**1.9**	2.1
PRD(4)	236	0.5	**23.4**	387	**0.8**	1.0
PRS(1)	34,831	**65.4**	**93.9**	21,954	41.2	41.5
PRS(2)	4,062	8.0	**65.3**	5,726	**11.3**	12.2
PRS(3)	1,066	2.2	**53.7**	1,636	**3.4**	3.8
PRS(4)	541	1.2	**53.7**	654	**1.4**	1.7
LCS	44,454	5.9	**89.4**	92,587	**12.2**	18.4

Incremental Parsing



Prefix	Shares of UAL , the parent	PRD(3)	PRS(3)
ITSG	company of United Airlines ,	–	–
SRILM	company , which is the	–	–
Goldstd	of United Airlines , were extremely active all day Friday .		
Prefix	PSE said it expects to report earnings of $ 1.3 million to $ 1.7 million , or 14		
ITSG	cents a share ,	–	+
SRILM	% to $ UNK	–	–
Goldstd	cents to 18 cents a share .		

Conclusions

- First incremental TSG parser.
- Competitive results on both full sentence and sentence prefix F-score.
- Outperforms standard n-gram LM in predicting more than one upcoming word.

References

Rens Bod, Khalil Sima'an, and Remko Scha. 2003. *Data-Oriented Parsing*. University of Chicago Press, Chicago, IL, USA.
Ciprian Chelba and Frederick Jelinek. 2000. Structured language modeling. *Computer Speech and Language*, 14:283–332.
Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 101(2):193–210.
Vera Demberg, Frank Keller, and Alexander Koller. 2014. Parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 40(1).
Jay Earley. 1970. An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102.
Joshua Goodman. 1996. Parsing algorithms and metrics. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 177–183. Association for Computational Linguistics, Morristown, NJ, USA.
Korinna Grabski and Tobias Scheffer. 2004. Sentence completion. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 433–439.
Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics, Morristown, NJ, USA.
Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
Slav Petrov. 2009. *Coarse-to-Fine Natural Language Processing*. Ph.D. thesis, University of California at Bekeley, Berkeley, CA, USA.
Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Comput. Linguist.*, 27:249–276.
Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the EMNLP 2009*, pages 324–333. Association for Computat. Linguist., Stroudsburg, PA, USA.
Federico Sangati, Willem Zuidema, and Rens Bod. 2010. Efficiently Extract Recurring Tree Fragments from Large Treebanks. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.
Yves Schabes. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage parsing using human-like memory constraints. *Comput. Linguist.*, 36(1):1–30.
Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Comput. Linguist.*, 21(2):165–201.
Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proc. International Conference on Spoken Language Processing*, pages 257–286.