

ACCURATE AND ROBUST LFG-BASED GENERATION FOR CHINESE

Yuqing Guo¹ Haifeng Wang² Josef van Genabith¹
presented by Jennifer Foster¹

¹NCLT, School of Computing
Dublin City University

²Research & Development Center
Toshiba (China) Co. Ltd.



TOSHIBA
Leading Innovation >>>

INLG, June 2008

OUTLINE

1 INTRODUCTION

- Surface Realisation
- Lexical-Functional Grammar

2 LFG-BASED GENERATION

- PCFG Generation Models
- Smoothing Algorithms for Chart Generator

3 EXPERIMENTS AND RESULTS

OUTLINE

1 INTRODUCTION

- Surface Realisation
- Lexical-Functional Grammar

2 LFG-BASED GENERATION

- PCFG Generation Models
- Smoothing Algorithms for Chart Generator

3 EXPERIMENTS AND RESULTS

WHAT IS SURFACE REALISATION?

DEFINITION

A subtask of NLG to produce syntactically, morphologically, and orthographically correct sentences from a given semantic/syntactic representation.

INPUT: lexicalised semantic or syntactic representation of linguistic content

OUTPUT: meaningful, grammatical and fluent natural language string

MOTIVATION

- Unification-based formalisms, e.g. LFG, HPSG, CCG, TAG
 - Handcrafted rules: FUF/SURGE, Lingo/LKB, OpenCCG, XLE etc.
 - Wide-coverage grammars automatically extracted from treebanks: Nakanishi (2005), Cahill (2006), White (2007) etc.
 - ? How *automatically acquired grammars* perform in generation
- An inverse process of parsing
 - *Grammar transforms* and *lexicalisation* improve PCFG parsing
 - ? How parsing techniques affect the performance of generation
- Robustness, Reusability
 - Language- and Domain-independent
 - ? How the techniques and grammars perform on *Chinese* data

MOTIVATION

- Unification-based formalisms, e.g. LFG, HPSG, CCG, TAG
 - Handcrafted rules: FUF/SURGE, Lingo/LKB, OpenCCG, XLE etc.
 - Wide-coverage grammars automatically extracted from treebanks: Nakanishi (2005), Cahill (2006), White (2007) etc.
 - ? How *automatically acquired grammars* perform in generation
- An inverse process of parsing
 - *Grammar transforms* and *lexicalisation* improve PCFG parsing
 - ? How parsing techniques affect the performance of generation
- Robustness, Reusability
 - Language- and Domain-independent
 - ? How the techniques and grammars perform on *Chinese* data

MOTIVATION

- Unification-based formalisms, e.g. LFG, HPSG, CCG, TAG
 - Handcrafted rules: FUF/SURGE, Lingo/LKB, OpenCCG, XLE etc.
 - Wide-coverage grammars automatically extracted from treebanks: Nakanishi (2005), Cahill (2006), White (2007) etc.
 - ? How *automatically acquired grammars* perform in generation
- An inverse process of parsing
 - *Grammar transforms* and *lexicalisation* improve PCFG parsing
 - ? How parsing techniques affect the performance of generation
- Robustness, Reusability
 - Language- and Domain-independent
 - ? How the techniques and grammars perform on *Chinese* data

OUTLINE

1 INTRODUCTION

- Surface Realisation
- **Lexical-Functional Grammar**

2 LFG-BASED GENERATION

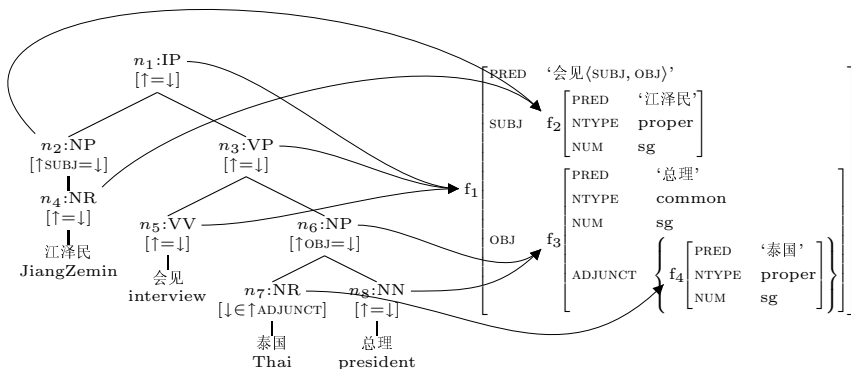
- PCFG Generation Models
- Smoothing Algorithms for Chart Generator

3 EXPERIMENTS AND RESULTS

C- AND F-STRUCTURES

$$\phi : C \rightarrow F$$

$$\phi(n_1)=\phi(n_3)=\phi(n_5)=f_1 \quad \phi(n_2)=\phi(n_4)=f_2 \quad \phi(n_6)=\phi(n_8)=f_3 \quad \phi(n_7)=f_4$$



C(onstituent) Structure

F(unctional) Structure

LFG GRAMMARS

- (1) IP \longrightarrow NP VP
 [↑SUBJ=↓] [↑=↓]
- (2) VP \longrightarrow VV NP
 [↑=↓] [↑OBJ=↓]
- (3) NP \longrightarrow NR NN
 [↑ADJ=↓] [↑=↓]
- (4) NP \longrightarrow NR
 [↑=↓]

(1) 会见
interview

(2) 总理
president

(3) 泰国
Thai

(4) 江泽民
JiangZemin

VV (↑PRED)='会见<SUBJ, OBJ>'

NN (↑PRED)='总理'
 (↑NTYPE)=common
 (↑NUM)=sg

NR (↑PRED)='泰国'
 (↑NTYPE)=proper
 (↑NUM)=sg

NR (↑PRED)='江泽民'
 (↑NTYPE)=proper
 (↑NUM)=sg

f-structure annotated CFG rules

lexical entries

OUTLINE

1 INTRODUCTION

- Surface Realisation
- Lexical-Functional Grammar

2 LFG-BASED GENERATION

- PCFG Generation Models
- Smoothing Algorithms for Chart Generator

3 EXPERIMENTS AND RESULTS

LFG GENERATES CONTEXT-FREE LANGUAGE

DEFINITION

Generation in LFG is to determine the strings of a language that correspond to a specified f-structure, given a particular grammar.

INPUT: f-structure with unordered grammatical functions

OUTPUT: corresponding functional annotated c-structure tree
(specifies a surface realisation)

KAPLAN AND WEDEKIND (2000)

The set of strings generated from fully specified f-structures according to the LFG grammar is a *context-free* language.

LFG GENERATES CONTEXT-FREE LANGUAGE

DEFINITION

Generation in LFG is to determine the strings of a language that correspond to a specified f-structure, given a particular grammar.

INPUT: f-structure with unordered grammatical functions

OUTPUT: corresponding functional annotated c-structure tree
(specifies a surface realisation)

KAPLAN AND WEDEKIND (2000)

The set of strings generated from fully specified f-structures according to the LFG grammar is a *context-free* language.

BASIC PCFG MODEL

CAHILL AND VAN GENABITH (2006)

A PCFG-based chart generator using wide-coverage LFG approximations automatically extracted from the Penn-II treebank.

- F = an f-structure
- T = an annotated c-structure corresponding to the f-structure
- A statistical generation model defines $P(T|F)$
- The best realisation is $T_{best} = \operatorname{argmax}_T P(T|F)$
- Decomposition:

$$P(T|F) = \prod_{X \rightarrow Y \text{ in } T} P(X \rightarrow Y | X, Feats)$$

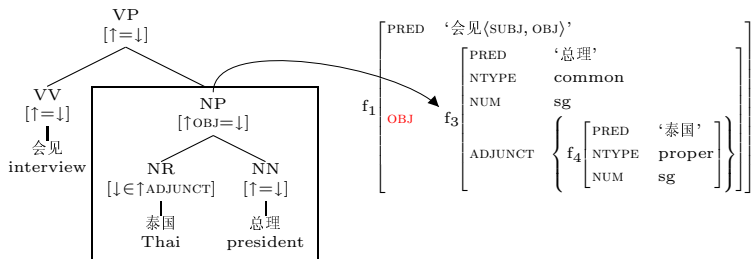
$$Feats = \{a_i | a_i \in \phi(X)\}$$

INDEPENDENCE ASSUMPTION IN PCFGs

- Missing *structural dependencies* between rules
- Lack of sensitivity to *lexical dependencies*

- (1) a. 这 趟 列车 开 往北京
 this CLS train run to Beijing
 ‘The train is bound for Beijing.’
- b. 这 趟 列车 往北京 开
 this CLS train to Beijing run
- (2) a. 泰国 总理 对中国 进行 访问
 Thai president to China make visit
 ‘The Thai president paid a visit to China.’
- b. * 泰国 总理 进行 访问 对中国
 Thai president make visit to China

HISTORY-BASED MODEL



Grammar Rule	Conditions
NP [↑OBJ=↓] → NR [↓∈↑ADJUNCT] NN [↑=↓]	NP [↑OBJ=↓] {PRED, NTYPE, NUM, ADJUNCT} OBJ

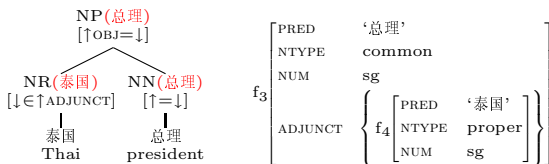
$$P(T|F) = \prod P(X \rightarrow Y|X, Feats, \mathbf{GF})$$

$$X \rightarrow Y \text{ in } T$$

$$Feats = \{a_i | a_i \in \phi(X)\}$$

$$\exists f (f \mathbf{GF}) = \phi(X)$$

LEXICALISED MODEL



Grammar Rule			Conditions
NP(总理) [↑OBJ=↓]	→	NR(泰国) NN(总理) [↓∈↑ADJUNCT] [↑=↓]	NP(总理) {PRED, NTYPE, NUM, ADJUNCT} [↑OBJ=↓]

$$P(T|F) = \prod_{X \rightarrow Y \text{ in } T} P(X(\mathbf{x}) \rightarrow Y(\mathbf{y}) | X(\mathbf{x}), Feats)$$

$$Feats = \{a_i | a_i \in \phi(X)\}$$

OUTLINE

1 INTRODUCTION

- Surface Realisation
- Lexical-Functional Grammar

2 LFG-BASED GENERATION

- PCFG Generation Models
- Smoothing Algorithms for Chart Generator

3 EXPERIMENTS AND RESULTS

LEXICAL SMOOTHING

- 1 Extracting lexical macros from lexical entries

	POS	lexical features (f)
泰国	NR	$(\uparrow\text{PRED})=\text{'泰国'}, (\uparrow\text{NTYPE})=\text{proper}, (\uparrow\text{NUM})=\text{sg}$
江泽民	NR	$(\uparrow\text{PRED})=\text{'江泽民'}, (\uparrow\text{NTYPE})=\text{proper}, (\uparrow\text{NUM})=\text{sg}$
lexical macro	NR	$(\uparrow\text{NTYPE})=\text{proper}, (\uparrow\text{NUM})=\text{sg}$

- 2 Predicting POS of unknown words by lexical macros

$$P(\text{POS}|f) = \frac{\text{count}(\text{POS}, f)}{\sum_{i=1}^n \text{count}(\text{POS}_i, f)}$$

RULE SMOOTHING

FEATURES SMOOTH: reducing the conditioning f-structure features

PARTIAL MATCH: reducing components of the CFG rule

EXAMPLES

Nonsmooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$
Feature smooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$
Partial match	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$

RULE SMOOTHING

FEATURES SMOOTH: reducing the conditioning f-structure features

PARTIAL MATCH: reducing components of the CFG rule

EXAMPLES

Nonsmooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$
Feature smooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow],$
Partial match	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$

RULE SMOOTHING

FEATURES SMOOTH: reducing the conditioning f-structure features

PARTIAL MATCH: reducing components of the CFG rule

EXAMPLES

Nonsmooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$
Feature smooth	$VP[\uparrow=\downarrow] \rightarrow VV[\uparrow=\downarrow] NP[\uparrow OBJ=\downarrow],$
Partial match	$VP \rightarrow VV \quad [\uparrow OBJ=\downarrow], \{SUBJ, OBJ, PRED\}$

CHART GENERATION ALGORITHM

- Each (sub-)f-structure indexes a (sub-)chart
- Each local chart generates the most probable trees for the corresponding f-structure:
 - 1 generating lexical edges by PRED and special atomic features representing function words;
 - 2 applying unary rules and binary rules to generate new edges until no more new edges can be generated;
 - 3 if no edge in the local chart can cover the current f-structure, applying smoothed unary rules and binary rules to generate new edges;
 - 4 propagating compatible edges to the upper-level chart.

EXPERIMENTAL DATA

- Penn Chinese Treebank (CTB6)
 - Data source: newswire from Xinhua, Sinorama, HKSAR and transcripts from broadcast news
 - Training Set: 15,663 sentences
 - Test Set: 500 sentences (5 words \leq length \leq 80 words)
 - Development Set: 500 sentences (5 words \leq length \leq 80 words)
- F-structures automatically converted from CTB (Guo et al., 2007)
- Annotated CFG Rules extracted from the training set

	associated with f-structures	without f-structures
PCFG	22,372	8,548
HB-PCFG	28,487	11,969
LEX-PCFG	325,094	286,468

EVALUATION METRICS

COVERAGE: The percentage of input f-structures for which the generator produces a (complete or partial) sentence.

EXACT MATCH: The percentage of generated sentences that exactly match the reference.

BLEU SCORE: A geometric average of n-gram accuracy, adjusted by length penalty LP

$$BLEU = \exp \left(\sum_{n=1}^N w_n \log p_n \right) \times LP$$

SIMPLE STRING ACCURACY:

$$SSA = 1 - \frac{I + D + S}{R}$$

RESULTS WITHOUT SMOOTHING

All Sentences	Coverage	ExMatch	BLEU	SSA
PCFG	100%	7.2%	0.5401	0.6261
HB-PCFG	100%	8.60%	0.5474	0.6281
LEX-PCFG	100%	9.40%	0.5687	0.6537

TABLE: Results for all input f-structures

Complete Sentences	Coverage	ExMatch	BLEU	SSA
PCFG	36.40%	19.78%	0.7101	0.7687
HB-PCFG	34.80%	24.71%	0.7513	0.8092
LEX-PCFG	37.00%	25.41%	0.7431	0.8024

TABLE: Results for f-structures producing a complete sentence

RESULTS WITH LEXICAL AND RULE SMOOTHING

	Complete Sentences			
Partial match	Coverage	ExMatch	BLEU	SSA
PCFG	97.20%	11.32%	0.7022	0.7356
HB-PCFG	96.20%	12.27%	0.7263	0.7458
LEX-PCFG	97.80%	14.31%	0.7265	0.7696
Feature smooth	Coverage	ExMatch	BLEU	SSA
PCFG	100%	11.20%	0.7021	0.7330
HB-PCFG	100%	12.00%	0.7245	0.7413
LEX-PCFG	100%	14.20%	0.7265	0.7675

TABLE: Results for f-structures producing a complete sentence

SUMMARY

- PCFG-based models for *Chinese* generation.
- Using *treebank-based, automatically acquired* LFG resources.
- Weakening inappropriate independence assumptions by *including lexical dependencies* and *enriching conditioning context with parent f-structure features* in PCFG models.

- Outlook
 - Integrating a language model to rank equivalent realisations.
 - Comparing PCFG- and Dependency-based generation models.

THANKS & QUESTIONS!

TOSHIBA
Leading Innovation >>>

