Aspects of Tree-Based Statistical Machine Translation

Marcello Federico

Human Language Technology
FBK

2014

Outline

Tree-based translation models:

- Synchronous context free grammars
- Hierarchical phrase-based model

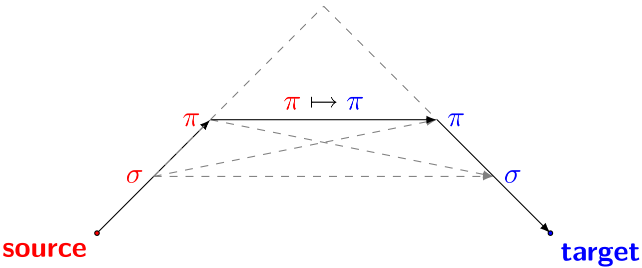
Decoding with SCFGs:

- Translation as Parsing
- DP-based chart decoding
- Integration of language model scores

Credits: adapted from slides by Gabriele Musillo.

Tree-Based Translation Models

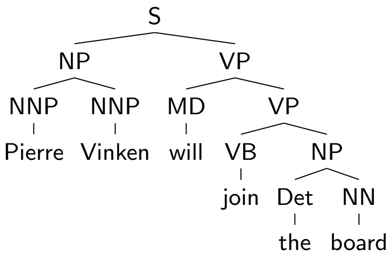
Levels of Representation in Machine Translation:



- $\pi \mapsto \sigma$: tree-to-string
- $\sigma \mapsto \pi$: string-to-tree
- $\pi \mapsto \pi$: **tree-to-tree**

? Appropriate Levels of Representation ?

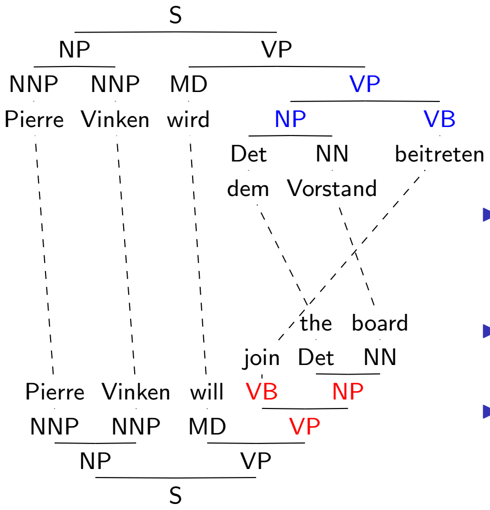
Tree Structures



Syntactic Structures:

- **rooted ordered** trees
- internal nodes labeled with **syntactic categories**
- leaf nodes labeled with words
- **linear** and **hierarchical** relations between nodes

Tree-to-Tree Translation Models



- syntactic **generalizations** over pairs of languages: **isomorphic** trees
- syntactically informed **unbounded reordering**
- formalized as derivations in **synchronous grammars**

? Adequacy of Isomorphism Assumption ?

Context-Free Grammars

CFG (Chomsky, 1956):

- formal model of languages
- more expressive than Finite State Automata and Regular Expressions
- first used in linguistics to describe **embedded** and **recursive** structures

CFG Rules:

- **left-hand side nonterminal symbol**
- **right-hand side string of nonterminal or terminal symbols**
- distinguished **start** nonterminal symbol

$$\begin{cases} S \rightarrow 0S1 & S \text{ rewrites as } 0S1 \\ S \rightarrow \epsilon & S \text{ rewrites as } \epsilon \end{cases}$$

CFG Examples

$\mathbf{G_1}$:

$$
\begin{aligned}
R = \{ & S \rightarrow NP\ VP, \\
& NP \rightarrow N | DET\ N | N\ PP, \\
& VP \rightarrow V\ NP | V\ NP\ PP, \\
& PP \rightarrow P\ NP, \\
& DET \rightarrow the | a, \\
& N \rightarrow Alice | Bob | trumpet, \\
& V \rightarrow chased, \\
& P \rightarrow with \}
\end{aligned}
$$

? **derivations** of
Alice chased Bob with the trumpet

$\mathbf{G_3}$:

$$
\begin{aligned}
R = \{ & NP \rightarrow NP\ CONJ\ NP | NP\ PP | DET\ N, \\
& PP \rightarrow P\ NP, P \rightarrow of, \\
& DET \rightarrow the | two | three, \\
& N \rightarrow mother | pianists | singers, \\
& CONJ \rightarrow and \}
\end{aligned}
$$

? **derivations** of
the mother of three pianists and two singers

- same parse tree can be derived in different ways ($\neq$ order of rules)
- same sentence can have different parse trees ($\neq$ choice of rules)

Transduction Grammars aka Synchronous Grammars

TG (Lewis and Stearns, 1968; Aho and Ullman, 1969):

- **two or more strings derived simultaneously**
- more powerful than FSTs
- used in NLP to model **alignments**, unbounded **reordering**, and mappings from surface forms to logical forms

Synchronous Rules:

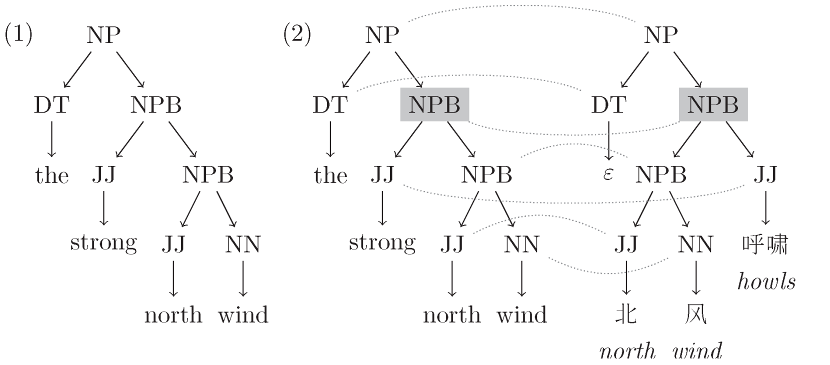
- left-hand side nonterminal symbol associated with **source** and **target** right-hand sides
- **bijection** $_{[]}$ mapping nonterminals in source and target of right-hand sides

$$\begin{cases} E \rightarrow E_{[1]} + E_{[2]} \;/\; + \; E_{[1]} \; E_{[2]} & \text{infix to Polish notation} \\ E \rightarrow E_{[1]} * E_{[2]} \;/\; * \; E_{[1]} \; E_{[2]} & \\ E \rightarrow n \;/\; n & n \in N \end{cases}$$

Synchronous CFG

$$\begin{aligned}
\text{NP} &\longrightarrow \text{DT}_{\boxed{1}}\text{NPB}_{\boxed{2}} \;/\; \text{DT}_{\boxed{1}}\text{NPB}_{\boxed{2}} \\
\text{NPB} &\longrightarrow \text{JJ}_{\boxed{1}}\text{NN}_{\boxed{2}} \;/\; \text{JJ}_{\boxed{1}}\text{NN}_{\boxed{2}} \\
\text{NPB} &\longrightarrow \text{NPB}_{\boxed{1}}\text{JJ}_{\boxed{2}} \;/\; \text{JJ}_{\boxed{2}}\text{NPB}_{\boxed{1}} \\
\text{DT} &\longrightarrow \text{the} \;/\; \varepsilon \\
\text{JJ} &\longrightarrow \text{strong} \;/\; \text{呼啸} \\
\text{JJ} &\longrightarrow \text{north} \;/\; \text{北} \\
\text{NN} &\longrightarrow \text{wind} \;/\; \text{风}
\end{aligned}$$

- **1-to-1 correspondence** between nodes
- **isomorphic** derivation trees
- uniquely determined **word alignment**



Hierarchical Phrase-Based Models

HPBM (Chiang, 2007):

- **formalized as SCFG**
- first tree-to-tree approach to perform better than phrase-based systems in large-scale evaluations
- **discontinuous phrases**, i.e. phrases with gaps
- **long-range reordering rules**
- no syntactic rules: **only two non-terminal symbols**

Example

Chinese-English: original, transliteration, glosses, and translation

澳洲	是	与	北韩	有	邦交	的	少数	国家	之一	。
Aozhou	shi	yu	Beihan	you	bangjiao	de	shaoshu	guojia	zhiyi	.
Australia	is	with	North Korea	have	dipl. rels.	that	few	countries	one of	.

Australia is one of the few countries that have diplomatic relations with North Korea.

HPBM: Motivations

Typical Phrase-Based Chinese-English Translation:

[Aozhou] [shi]$_1$ [yu Beihan]$_2$ [you] [bangjiao] [de shaoshu guojia zhiyi] [.]

[Australia] [has] [dipl. rels.] [with North Korea]$_2$ [is]$_1$ [one of the few countries] [.]

- Chinese VPs follow PPs / English VPs precede PPs

  $yu\ X_1\ you\ X_2$ / $have\ X_2\ with\ X_1$

- Chinese NPs follow RCs / English NPs precede RCs

  $X_1\ de\ X_2$ / $the\ X_2\ that\ X_1$

- translation of *zhiyi* construct in English word order

  $X_1\ zhiyi$ / $one\ of\ X_1$

HPBM: Example Rules

$$S \rightarrow X_1 \;/\; X_1 \tag{1}$$

$$S \rightarrow S_1\ X_2 \;/\; S_1\ X_2 \tag{2}$$

$$X \rightarrow yu\ X_1\ you\ X_2 \;/\; have\ X_2\ with\ X_1 \tag{3}$$

$$X \rightarrow X_1\ de\ X_2 \;/\; the\ X_2\ that\ X_1 \tag{4}$$

$$X \rightarrow X_1\ zhiyi \;/\; one\ of\ X_1 \tag{5}$$

$$X \rightarrow Aozhou \;/\; Australia \tag{6}$$

$$X \rightarrow Beihan \;/\; N.\ Korea \tag{7}$$

$$X \rightarrow she \;/\; is \tag{8}$$

$$X \rightarrow bangjiao \;/\; dipl.rels. \tag{9}$$

$$X \rightarrow shaoshu\ guojia \;/\; few\ countries \tag{10}$$

Summary

Synchronous Context-Free Grammars:

- Context-Free Grammars
- HPB recursive reordering model

Next topics:

- Decoding SCFGs: Translation as Parsing
- DP-based chart decoding
- Integration of language model scores

Synchronous Context-Free Grammars

SCFGs:

- CFGs in **two dimensions**
- **synchronous** derivation of **isomorphic**[a] **trees**
- **unbounded reordering** preserving **hierarchy**

[a] excluding leafs

$$\ldots$$

$$VB \rightarrow PRP_1\ VB1_2\ VB2_3\ /\ PRP_1\ VB2_3\ VB1_2$$

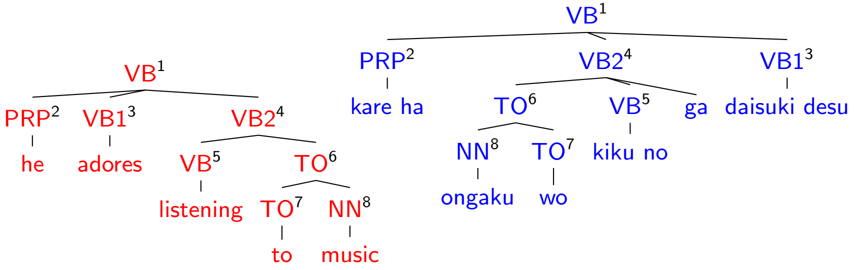
$$VB2 \rightarrow VB_1\ TO_2\ /\ TO_2\ VB_1\ ga$$

$$TO \rightarrow TO_1\ NN_2\ /\ NN_2\ TO_1$$

$$PRP \rightarrow he\ /\ kare\ ha$$

$$VB \rightarrow listening\ /\ daisuki\ desu$$

$$\ldots$$



Weighted SCFGs

- rules $A \rightarrow \alpha \ / \ \beta$ associated with positive weights $\mathbf{w}_{A\rightarrow\alpha/\beta}$
- derivation trees $\pi = \langle \pi_1, \pi_2 \rangle$ weighted as

$$\mathbf{W}(\pi) = \prod_{A\rightarrow\alpha/\beta\in G} \mathbf{w}_{A\rightarrow\alpha/\beta}^{c(A\rightarrow\alpha/\beta;\pi)}$$

- **probabilistic SCFGs** if the following conditions hold

$$\mathbf{w}_{A\rightarrow\alpha/\beta} \in [0,1] \text{ and } \sum_{\alpha,\beta} \mathbf{W}_{A\rightarrow\alpha/\beta} = 1$$

- notice: SCFGs might well include rules of type

$$A \rightarrow \alpha/\beta_1 \ldots A \rightarrow \alpha/\beta_k$$

MAP Translation Problem

Maximum A Posterior Translation:

$$e^\star = \operatorname*{argmax}_e p(e|f)$$
$$= \operatorname*{argmax}_e \sum_{\pi \in \Pi(f,e)} p(e, \pi | f)$$

$\Pi(f, e)$ *is the set of synchronous derivation trees yielding* $\langle f, e \rangle$

- Exact MAP decoding is NP-hard (Simaan, 1996; Satta and Peserico, 2005)

Viterbi Approximation

Tractable Approximate Decoding:

$$
\begin{aligned}
e^\star &= \operatorname*{argmax}_{e} \sum_{\pi \in \Pi(f,e)} p(e, \pi | f) \\
&\simeq \operatorname*{argmax}_{e} \max_{\pi \in \Pi(f,e)} p(e, \pi | f) \\
&= E(\operatorname*{argmax}_{\pi \in \Pi(f)} p(\pi))
\end{aligned}
$$

$\Pi(f)$ *is the set of synchronous derivations yielding* f

$E(\pi)$ *is the target string resulting from the synchronous derivation* π

Translation as Parsing

$$\pi^\star = \underset{\pi \in \Pi(f)}{\operatorname{argmax}}\, p(\pi)$$

Parsing Solution:

1. compute the **most probable derivation tree** that generates f using the **source dimension** of the WSCFG
2. build the **translation string** e by applying the **target dimension** of the rules used in the most probable derivation

- most probable derivation computed in $O(n^3)$ using **dynamic programming** algorithms for parsing **weighted CFGs**
- transfer of decoding algorithms developed for CFG to SMT

Weighted CFGs in Chomsky Normal Form

WCFGs:

- rules $A \rightarrow \alpha$ associated with positive weights $\mathbf{w}_{A\rightarrow\alpha}$
- derivation trees π weighted as

$$\mathbf{W}(\pi) = \prod_{A\rightarrow\alpha\in G} \mathbf{w}_{A\rightarrow\alpha}^{c(A\rightarrow\alpha;\pi)}$$

- probabilistic CFGs if the following conditions hold

$$\mathbf{w}_{A\rightarrow\alpha} \in [0,1] \text{ and } \sum_{\alpha} \mathbf{w}_{A\rightarrow\alpha} = 1$$

WCFGs in CNF:

- rules in CFGs in Chomsky Normal Form: $\mathbf{A} \rightarrow \mathbf{BC}$ or $\mathbf{A} \rightarrow \mathbf{a}$
- **equivalence** between WCFGs and WCFGs in CNF
- no analogous equivalence holds for weighted SCFGs

Weighted CKY Parsing

Dynamic Programming:

- **recursive division of problems into subproblems**
- **optimal solutions compose optimal sub-solutions (Bellman's Principle)**
- **tabulation of subproblems and their solutions**

CKY Parsing:

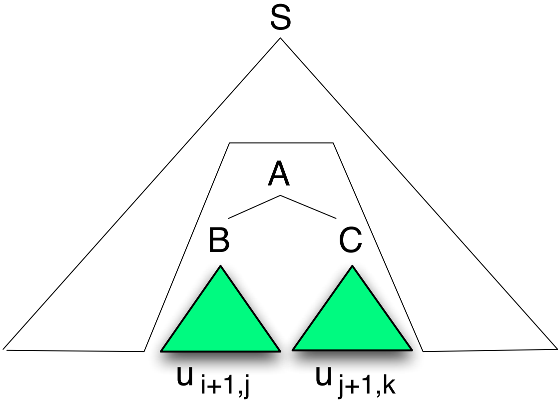
- subproblems: **parsing substrings of the input string**

$$u_1 \ldots u_n$$

- **bottom up** algorithm starting with derivation of terminals
- solutions to subproblems tabulated using a **chart**
- $O(n^3|G|)$ time complexity

Weighted CKY Parsing

$$Q(A, i, k) = \max_{B, C, i<j<k} \{\mathbf{w}_{A \rightarrow B\ C} \times Q(B, i, j) \times Q(C, j, k)\}$$



Parsing SCFG and Language Modelling

Viterbi Decoding of WSCFGs:

- focus on **most probable** derivation of source (ignoring different target sides associated with the same source side)
- **derivation weights** do not include **language models scores**

? HOW TO EFFICIENTLY COMPUTE TARGET LANGUAGE MODEL SCORES FOR POSSIBLE DERIVATIONS ?

Approaches:

1. **online**: integrate target *m*-gram LM scores into dynamic programming parsing
2. **cube pruning** (Huang and Chiang, 2007): rescore *k*-best sub-translations at each node of the parse forest

Online Translation

Bàowēier	*yǔ*	*Shālóng*	*jǔxíng*	*le*	*huìtán*
Powell	with	Sharon	hold	[*past*]	meeting

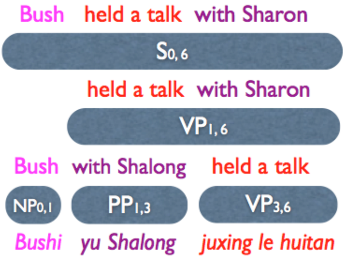
"Powell held a meeting with Sharon"

S	→	$NP^{(1)}\ VP^{(2)}$,	$NP^{(1)}\ VP^{(2)}$
VP	→	$PP^{(1)}\ VP^{(2)}$,	$VP^{(2)}\ PP^{(1)}$
NP	→	*Bàowēier*,	Powell
VP	→	*jǔxíng le huìtán*,	held a meeting
PP	→	*yǔ Shālóng*,	with Sharon

Online Translation: parsing of the source string and building of the corresponding subtranslations **in parallel**

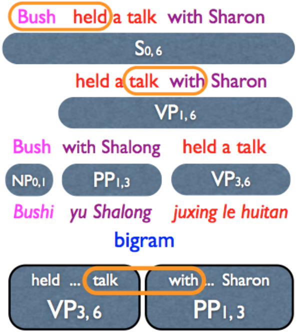
$$\frac{PP_{1,3} : (w_1, t_1)\ \ VP_{3,6} : (w_2, t_2)}{VP_{1,6} : (w \times w_1 \times w_2, t_2 t_1)}$$

- w_1, w_2: weights of the two antecedents
- w: weight of the synchronous rule
- t_1, t_2: translations



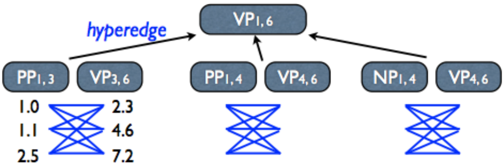
LM Online Integration (Wu, 1996)

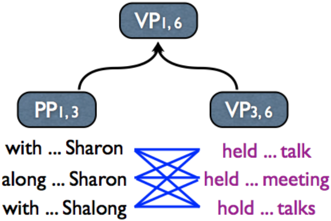
$$\frac{PP_{1,3}^{with*Sharon} : (w_1, t_1) \quad VP_{3,6}^{held*talk} : (w_2, t_2)}{VP_{1,6}^{held*Sharon} : (w \times w_1 \times w_2 \times p_{LM}(with|talk), t_2 t_1)}$$



- Integrate LM information in the state: $Q(A, i, j, pfx, sfx)$
- $O(n^3|E|^{4(m-1)})$: recombine 4 prefixes/suffixes of (m-1) words

Cube Pruning (Huang and Chiang, 2007)





Beam Search:

- at each step in the derivation, keep at most *k* items integrating target subtranslations in a beam
- enumerate all possible combinations of LM items
- extract the *k*-best combinations

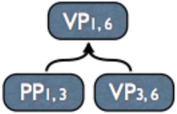
Cube Pruning:

- get *k*-best LM items **without computing all combinations**
- approximate search: in practice negligible search errors

Cube Pruning

Heuristic Assumption:

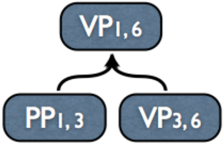
- margin scores are -log-probs of the left/right spans
- **best adjacent items** lie towards the **upper-left corner**
- part of the grid can be pruned **without computing its cells**



non-monotonic grid due to LM combo costs

		$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{along} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Shalong}})$
		1.0	3.0	8.0
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{meeting}})$	1.0	2.5	9.0	9.5
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{talk}})$	1.1	2.4	9.5	9.4
$(\mathrm{VP}_{3,6}^{\text{hold} \star \text{conference}})$	3.5	5.1	17.0	12.1

Cube Pruning: Example



bigram (meeting, with)

non-monotonic grid
due to LM combo costs

		(PP $_{1,3}^{\text{with} \star \text{Sharon}}$) 1.0	(PP $_{1,3}^{\text{along} \star \text{Sharon}}$) 3.0	(PP $_{1,3}^{\text{with} \star \text{Shalong}}$) 8.0
(VP $_{3,6}^{\text{held} \star \text{meeting}}$)	1.0	2.0 + 0.5	4.0 + 5.0	9.0 + 0.5
(VP $_{3,6}^{\text{held} \star \text{talk}}$)	1.1	2.1 + 0.3	4.1 + 5.4	9.1 + 0.3
(VP $_{3,6}^{\text{hold} \star \text{conference}}$)	3.5	4.5 + 0.6	6.5 +10.5	11.5 + 0.6

Cube Pruning: Example

***k*-best parsing**
(Huang and Chiang, 2005)

- a priority queue of candidates
- extract the best candidate

		$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{along} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Shalong}})$
		1.0	3.0	8.0
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{meeting}})$	1.0	2.5		
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{talk}})$	1.1			
$(\mathrm{VP}_{3,6}^{\text{hold} \star \text{conference}})$	3.5			

Cube Pruning: Example

k-best parsing
(Huang and Chiang, 2005)

- a priority queue of candidates
- extract the best candidate
- push the two successors

	$(\text{PP}_{1,3}^{\text{with}} \star \text{Sharon})$ 1.0	$(\text{PP}_{1,3}^{\text{along}} \star \text{Sharon})$ 3.0	$(\text{PP}_{1,3}^{\text{with}} \star \text{Shalong})$ 8.0
$(\text{VP}_{3,6}^{\text{held}} \star \text{meeting})$ 1.0	2.5	9.0	
$(\text{VP}_{3,6}^{\text{held}} \star \text{talk})$ 1.1	2.4		
$(\text{VP}_{3,6}^{\text{hold}} \star \text{conference})$ 3.5			

Cube Pruning: Example

***k*-best parsing**
(Huang and Chiang, 2005)

- a priority queue of candidates
- extract the best candidate
- push the two successors

		$(\text{PP}^{\text{with} \star \text{Sharon}}_{1,3})$	$(\text{PP}^{\text{along} \star \text{Sharon}}_{1,3})$	$(\text{PP}^{\text{with} \star \text{Shalong}}_{1,3})$
		1.0	3.0	8.0
$(\text{VP}^{\text{held} \star \text{meeting}}_{3,6})$	1.0	2.5	9.0	
$(\text{VP}^{\text{held} \star \text{talk}}_{3,6})$	1.1	2.4	9.5	
$(\text{VP}^{\text{hold} \star \text{conference}}_{3,6})$	3.5	5.1		

Summary

Translation As Parsing:

- Viterbi Approximation
- Weighted CKY Parsing
- Online LM Integration and Cube Pruning