

The UPV Translation System for IWSLT 2009

Authors: *Guillem Gascó i Mora*
Joan Andreu Sánchez Peiró



ITI

INSTITUTO TECNOLÓGICO
DE INFORMÁTICA



MIPRCV
CONSOLIDER INGENIO 2010
Multimodal Interaction in Pattern Recognition and Computer Vision



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

1-2 December 2009

Index

- 1 *Introduction* ▷ 1
- 2 Decoding ▷ 3
- 3 SITG Training ▷ 7
- 4 Experiments ▷ 16
- 5 Conclusions ▷ 19

Motivation and Related Work

► Motivation

- PBT systems tend to be weak on target language fluency.
- Long-range dependencies or reorderings cannot be controlled by the n-gram language models.
- Extending PBT systems with syntactic information is difficult.
- Syntactic MT systems usually solve such problems but have a low sentence coverage.
- Solution proposed: Fill the gap between both approaches presenting a hybrid system that uses Stochastic Inversion Transduction Grammars.

► Related Work

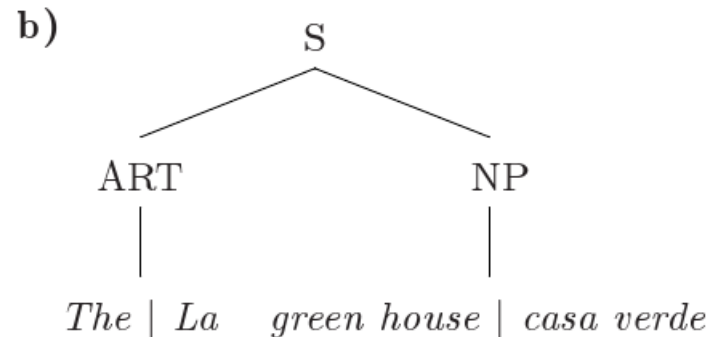
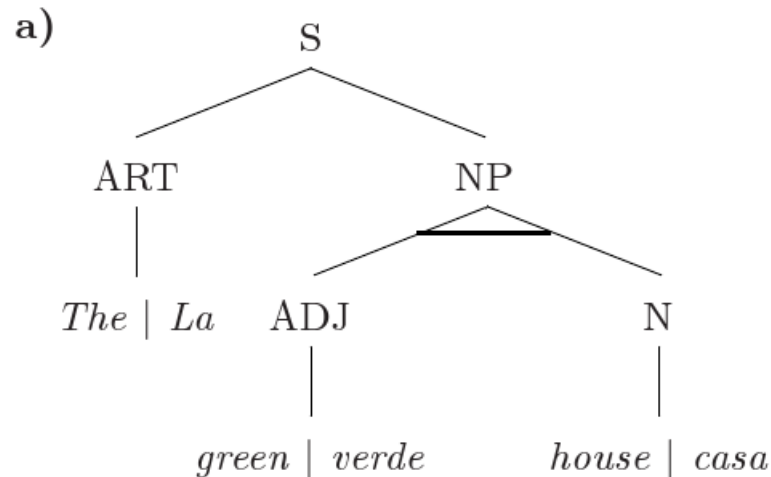
- Maximum entropy models for BTG statistical MT [Xiong06] [Xiong08].
- Hierarchical MT systems [Chiang 05]
- Syntax Augmented MT [Venugopal 06]

Index

- 1 Introduction ▷ 1
- 2 *Decoding* ▷ 3
- 3 SITG Training ▷ 7
- 4 Experiments ▷ 16
- 5 Conclusions ▷ 19

Theoretical Framework: Stochastic Phrasal ITG

- ▶ Phrasal Inversion Transduction Grammars $(\mathcal{N}, \Sigma, \Delta, S, \mathcal{R})$:
 - Direct Syntactic rule: $A \rightarrow [BC]$ where $A, B, C \in \mathcal{N}$
 - Inverse Syntactic rule: $A \rightarrow \langle BC \rangle$ where $A, B, C \in \mathcal{N}$
 - Lexical rule: $A \rightarrow x/y$ where $x \in \Sigma^*$ and $y \in \Delta^*$



- ▶ SPhITG: Stochastic extension of PhITG.

$$\sum_{B, C \in \mathcal{N}} (\Pr(A \rightarrow [B C]) + \Pr(A \rightarrow \langle B C \rangle)) + \sum_{\substack{x \in \Sigma^* \\ y \in \Delta^*}} (\Pr(A \rightarrow x/y)) = 1$$

Translation Model

- ▶ Translation goal:

$$(t^*, D^*) = \operatorname{argmax}_{t, D} \Pr(S \xrightarrow{D} s/t)$$

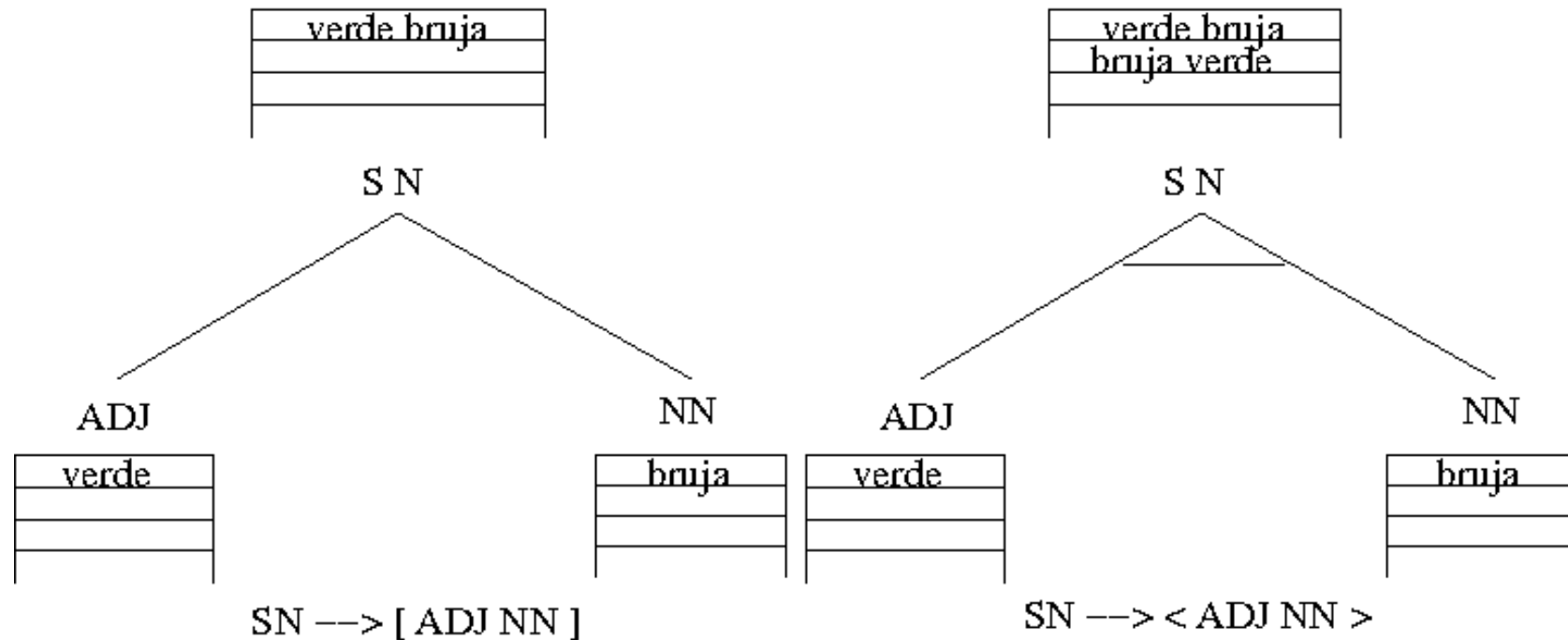
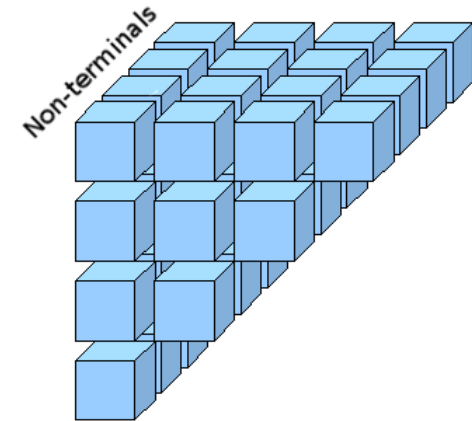
- ▶ Log-linear model over the derivations:

$$\Pr(D) = \prod_i h_i(D)^{\lambda_i}$$

- ▶ Models used: Usual models of PBT systems and the syntactic model (probability of the syntactic SITG rules used in D).

Decoding Algorithm

- ▶ Without n-gram language model: **CKY-like** chart decoding.
- ▶ Using the n-gram language model: A Hypotheses stack in each cell of the chart.
- ▶ Optimization strategies: Recombination of hypotheses, beam pruning and histogram pruning.



Index

- 1 Introduction ▷ 1
- 2 Decoding ▷ 3
- 3 *SITG Training* ▷ 7
- 4 Experiments ▷ 16
- 5 Conclusions ▷ 19

SITG Training

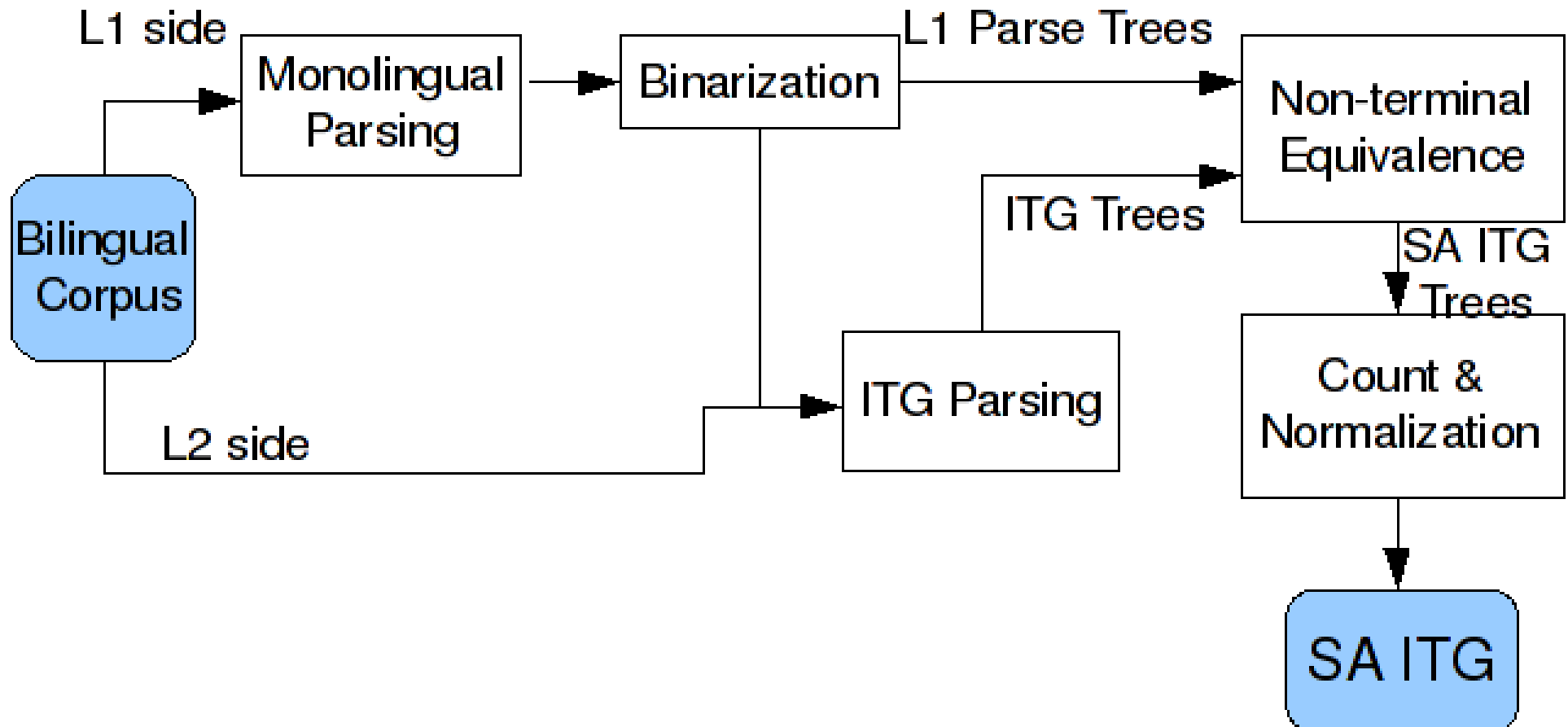
- ▶ Initial ITG:
 - Lexical rules from alignments, all with the same non-terminal.
 - All the possible syntactical rules using 4 non-terminals (from NT1 to NT4) with a random probability.

$$NT1 \rightarrow [NT1 NT1] \dots NT4 \rightarrow \langle NT4 NT4 \rangle$$

- ▶ Reestimation of the SITG using the Viterbi reestimation algorithm:
 - Get the most likely ITG parse trees.
 - Estimate probabilities by counting productions and normalizing.

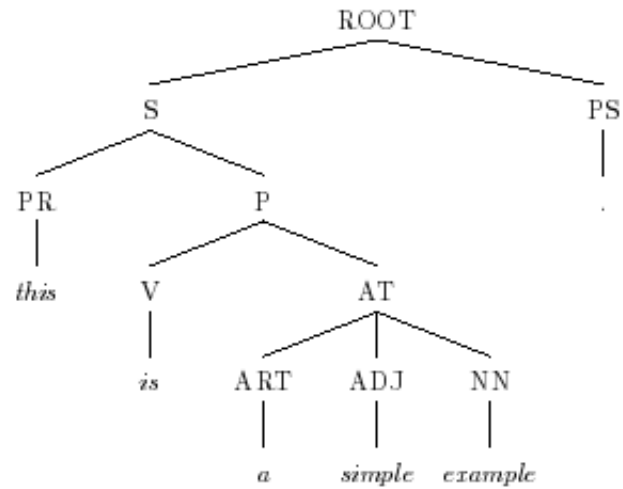
SITG Training

- ▶ Association of linguistic meaning (from the input) to the SITG non-terminals.



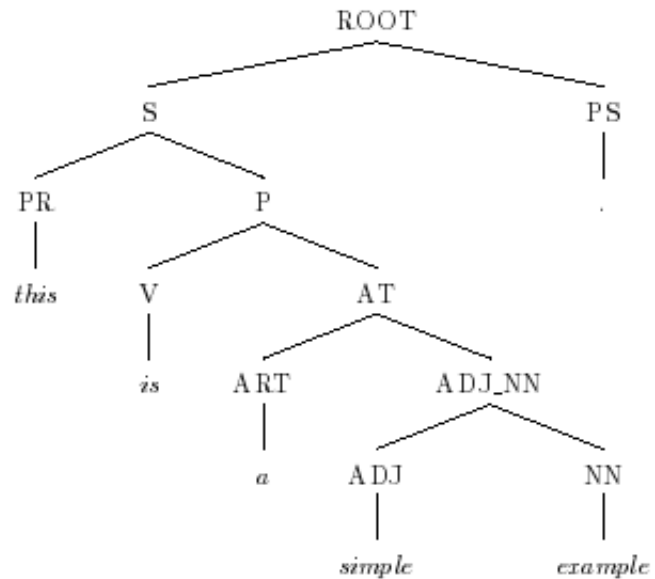
SITG Training

this is a simple example . # esto es un ejemplo simple .



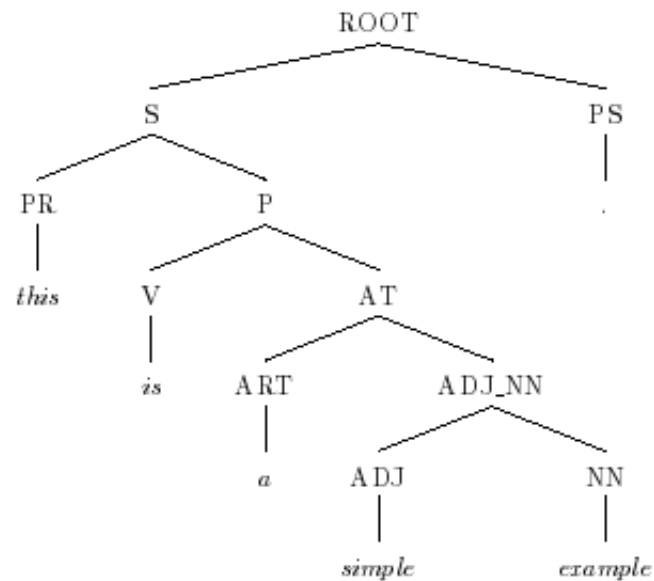
SITG Training

this is a simple example . # esto es un ejemplo simple .



SITG Training

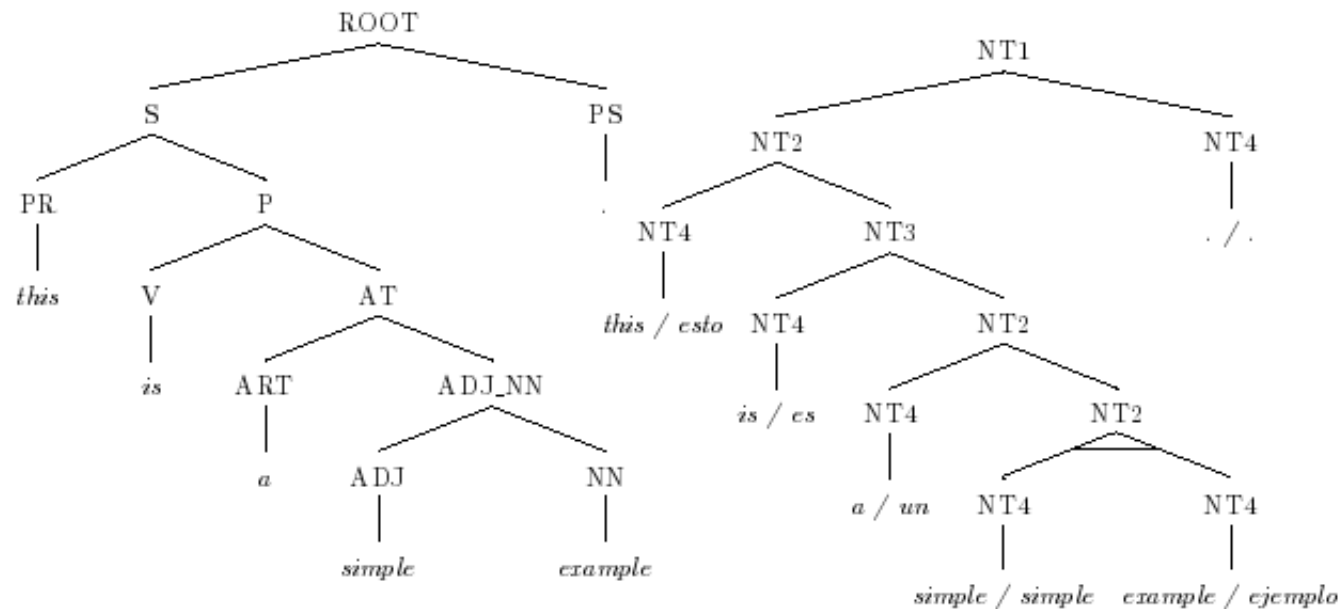
this is a simple example . # esto es un ejemplo simple .



((this (is (a (simple example)))) .) # esto es un ejemplo simple

SITG Training

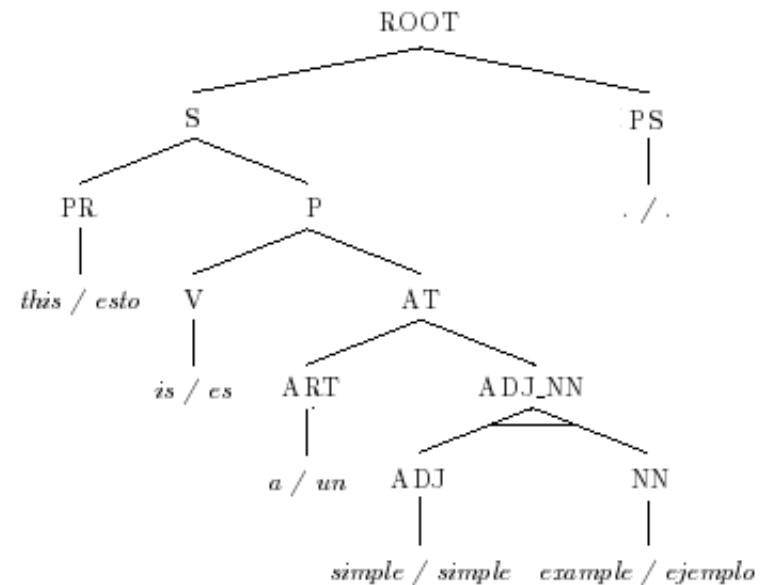
this is a simple example . # esto es un ejemplo simple .



((this (is (a (simple example)))) .) # esto es un ejemplo simple

SITG Training

this is a simple example . # esto es un ejemplo simple .

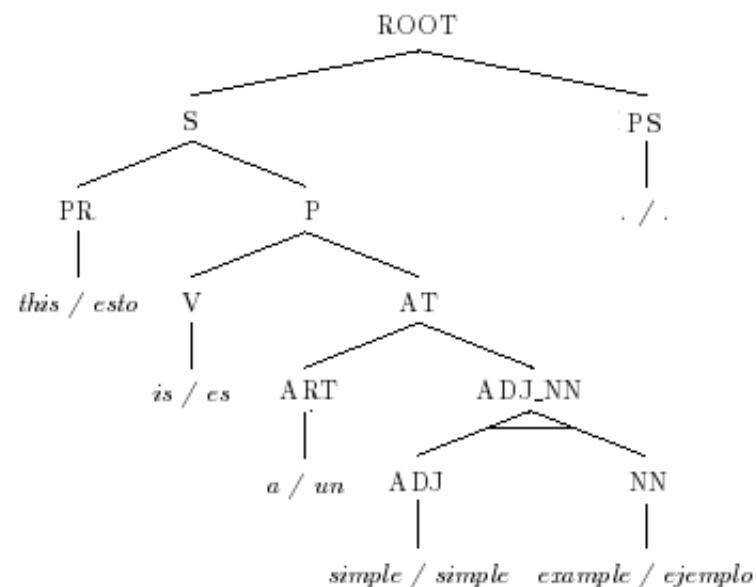


((this (is (a (simple example)))) .) # esto es un ejemplo simple

SITG Training

this is a simple example . # esto es un ejemplo simple .

ROOT --> [S PS]
S --> [PR P]
P --> [V AT]
AT --> [ART ADJ_NN]
ADJ_NN --> < ADJ NN >
PS --> . / .
PR --> this / esto
V --> is / es
ART --> a / un
ADJ --> simple / simple
NN --> example / ejemplo



((this (is (a (simple example)))) .) # esto es un ejemplo simple

Index

- 1 Introduction ▷ 1
- 2 Decoding ▷ 3
- 3 SITG Training ▷ 7
- 4 *Experiments* ▷ 16
- 5 Conclusions ▷ 19

Experimental Results

- ▶ Translation experiments over IWSLT08 Chinese-English corpus:

		Chinese	English
Training	Sentences	19.972	
	Words	171,591	188,960
	Vocabulary Size	8,428	7,182
DevSet	Sentences	489	
	Words	3,169	3,861
	OOV Words	111	115
Test	Sentences	507	
	Words	3,357	-
	OOV Words	97	-

- ▶ Baseline System: PBT system with the same phrase table.

Experiment	% BLEU
Baseline	41.1
Initial SITG	41.23
Reestimated SITG	41.79
SAITG	42.85

Experimental Results

PBT	this one and what 's the difference between ?
SAITG	what 's the difference between this with that ?
Ref	how is this one different from that one ?
PBT	call mr. is three four one four five six seven .
SAITG	call mr. is three six four five seven four one .
Ref	the number for s nicholas is three six four five seven four one .
PBT	can i go to the front row ?
SAITG	is it okay to the front row ?
Ref	can i go up to the front ?

- Some rules obtained:
- $\Pr(\text{QP} \rightarrow [\text{CD CD}]) = 0.147$
 - $\Pr(\text{QP} \rightarrow \langle \text{CD CD} \rangle) = 0.046$
 - $\Pr(\text{QP} \rightarrow [\text{CD QP}]) = 0.284$
 - $\Pr(\text{QP} \rightarrow \langle \text{QP CD} \rangle) = 0.061$

Index

- 1 Introduction ▷ 1
- 2 Decoding ▷ 3
- 3 SITG Training ▷ 7
- 4 Experiments ▷ 16
- 5 *Conclusions* ▷ 19

Conclusions

- ▶ SITG based decoder.
- ▶ Analyzed heuristics to train the SITG.
- ▶ Phrase table from a PBT system.
- ▶ When no syntactic information is used, it is almost equivalent to a PBT.
- ▶ The use of linguistic information improve the results.

End

The UPV Translation System for IWSLT 2009

Authors: *Guillem Gascó i Mora*
Joan Andreu Sánchez Peiró



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

CKY-like Algorithm

$$\delta_{ij}(A) = \max_t \Pr(A \xrightarrow{*} s_i^j / t) \quad \text{For all } A \in N \text{ and } \begin{cases} 0 \leq i < j \leq |s|, \\ i, j \text{ such that } j - i \geq 1, \end{cases}$$

$$\delta_{ij}(A) = \max(\delta_{ij}^{[]} (A), \delta_{ij}^{\langle \rangle} (A), \max_t \Pr(A \rightarrow s_i^j / t)) \quad (1)$$

where

$$\delta_{ij}^{[]} (A) = \begin{cases} \max_{\substack{B, C \in N \\ i < I \leq j}} \Pr(A \rightarrow [BC]) \delta_{iI}(B) \delta_{Ij}(C) & \text{if } j - i > 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$\delta_{ij}^{\langle \rangle} (A) = \begin{cases} \max_{\substack{B, C \in N \\ i < I \leq j}} \Pr(A \rightarrow \langle BC \rangle) \delta_{iI}(B) \delta_{Ij}(C) & \text{if } j - i > 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

CKY-like Algorithm

$$\tau_{ij}(A) = \operatorname{argmax}_t (\Pr(A \Rightarrow^* s_i^j/t))$$

$$\tau_{ij}(A) = \begin{cases} t & \text{if } \Pr(A \rightarrow s_i^j/t) \text{ is the maximum in (1)} \\ \tau_{iI}(B)\tau_{Ij}(C) & \text{if } \Pr(A \rightarrow [BC])\delta_{iI}(B)\delta_{Ij}(C) \\ & \text{is the maximum in (1)} \\ \tau_{Ij}(C)\tau_{iI}(B) & \text{if } \Pr(A \rightarrow \langle BC \rangle)\delta_{iI}(B)\delta_{Ij}(C) \\ & \text{is the maximum in (1)} \end{cases} \quad (4)$$

[Back](#)