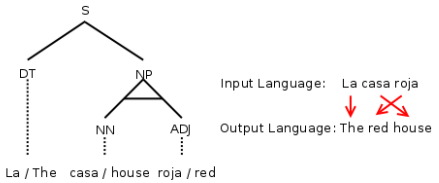
Complete Search Space Exploration for SITG Inside Probability

Guillem Gascó, Joan-Andreu Sánchez, José-Miguel Benedí
{ggasco,jandreu,jbenedi}@dsic.upv.es

Introduction

- Stochastic Inversion Transduction Grammars: $\{\Sigma, \Delta, N, S, R\}$
  - Σ, Δ input and output terminals
  - N non-terminals ans $S \in N$ initial symbol
  - R probabilistic rules of the form:
    - $A \to a/b$, $A \in N$, $a \in \Sigma \cup \{\epsilon\}$, $b \in \Delta \cup \{\epsilon\}$
    - $A \to [BC]$, $A, B, C \in N$
    - $A \to \langle BC \rangle$, $A, B, C \in N$
      ($A \to BC$ in the input language and
      $A \to CB$ in the output language)



- Used to parse two strings simultaneously.

Modified Parsing Algorithm for SITGs

1. Initialization

$$\begin{aligned}
\mathcal{E}_{i,i+1,k,k+1}[A] &= p(A \to x_{i+1}/y_{k+1}) \quad 0 \le i < |x|\ 0 \le k < |y| \\
\mathcal{E}_{i,i+1,k,k}[A] &= p(A \to x_{i+1}/\epsilon) \quad 0 \le i < |x|\ 0 \le k \le |y| \\
\mathcal{E}_{i,i,k,k+1}[A] &= p(A \to \epsilon/y_{k+1}) \quad 0 \le i \le |x|\ 0 \le k < |y|
\end{aligned}$$

2. Recursion

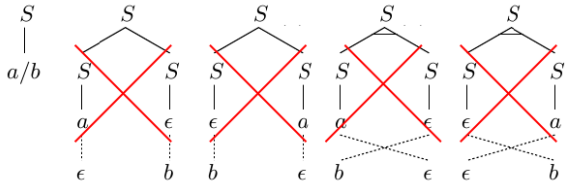
For all $A \in N$ and i, j, k, l such that $\begin{cases} 0 \le i \le |x|,\ 0 \le j \le |x| - i \\ 0 \le k \le |y|,\ 0 \le l \le |y| - k \\ j + l > 2, \end{cases} \longrightarrow j + l \ge 2$

$$\mathcal{E}_{i,i+j,k,k+l}[A] = \mathcal{E}^{[]}_{i,i+j,k,k+l}[A] + \mathcal{E}^{\langle\rangle}_{i,i+j,k,k+l}[A], \text{ where}$$

$$\mathcal{E}^{[]}_{i,i+j,k,k+l}[A] = \sum_{\substack{B,C \in N \\ 1 \le I \le j,\ 1 \le K \le l \\ I((j-I)+K(l-K)) \ne 0}} p(A \to [BC])\ \mathcal{E}_{i,i+I,k,k+K}[B]\ \mathcal{E}_{i+I,i+j,k+K,k+l}[C] \longrightarrow ((j-I)+(l-K)) \times (I+K) \ne 0$$

$$\mathcal{E}^{\langle\rangle}_{i,i+j,k,k+l}[A] = \sum_{\substack{B,C \in N \\ 1 \le I \le j,\ 1 \le K \le l \\ I((j-I)+K(l-K)) \ne 0}} p(A \to \langle BC \rangle)\ \mathcal{E}_{i,i+I,k+K,k+l}[B]\ \mathcal{E}_{i+I,i+j,k,k+K}[C] \longrightarrow ((j-I)+K) \times (I+(l-K)) \ne 0$$

- Some valid parse trees cannot be explored:



- Two other restrictions are substituted to guarantee termination.
- The modifications proposed allow the algorithm to explore the whole search space.
- Time complexity: $O(N^3|x|^3|y|^3)$
- The correctness of the modified algorithm has been proved (See the paper).

Experiments

- Bilingual corpora used:
  - IWSLT2009 Chinese-English

Set	Stat.	Ch	En
	Sents	42K	
Tr.	Words	330K	380K
	Sents	511	
Test	Words	3K	3K

  - Hansard French-English

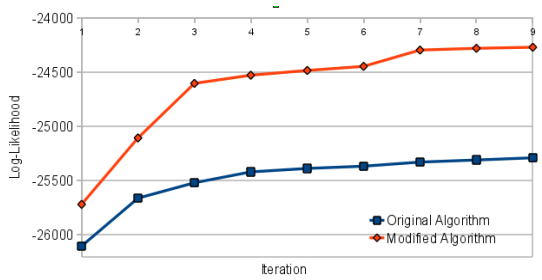
Set	Stat.	Fr	En
	Sents	997K	
Tr	Words	16.5M	14.2M

- Also using bracketing information (results between []).
- A: % of sentences for which the original algorithm cannot find the most probable parse tree.
- B: % of sentences not parsed by the original algorithm.

Experiment	A	B
Ch - En	36.25%	0.24%
[Ch] - En	37.21%	1.4%
Ch - [En]	36.97%	1.02%
[Ch] - [En]	40.93%	3.92%

Experiment	A
Fr - En	27.73%
[Fr] - En	28.06%
Fr - [En]	28.51%
[Fr] - [En]	30.56%

- Both algorithms have been tested for ITG inference purposes by using the Viterbi reestimation process.
- Loglikelihood of the IWSLT test for each algorithm:



Conclusions

- Original parsing algorithms for ITGs cannot explore the whole search space.
- The proposed modification have been proved to solve this problem.
- The non-explored trees are important in real scenarios.
- The modified algorithm performs better for inference purposes.

Acknowledgements

Supported by the EC (FEDER/FSE), the Spanish MICINN and MITyC and the Generalitat Valenciana under projects MIPRCV, iTrans2 and grant Prometeo 2009. Also supported by the scholarship BFPI/2007/117.