Using Parsed Corpora for Estimating Stochastic Inversion Transduction Grammars

UNIVERSIDAD POLITECNICA DE VALENCIA

Universidad Politécnica de Valencia

Instituto Tecnológico de Informática

Germán Sanchis, Joan Andreu Sánchez

{gsanchis,jandreu}@dsic.upv.es

May 2008

Index

Statistical Machine Translation

- SMT: efficient framework for building state-of-the-art MT systems.
- Problem originally defined as

$$\begin{aligned}\hat{\mathbf{y}} &= \operatorname*{argmax}_{\mathbf{y}} Pr(\mathbf{y}|\mathbf{x}) \\ &= \operatorname*{argmax}_{\mathbf{y}} Pr(\mathbf{x}|\mathbf{y}) \cdot Pr(\mathbf{y})\end{aligned}$$

- In practice, $Pr(\mathbf{y}|\mathbf{x})$ is modelled using log-linear models:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y})$$

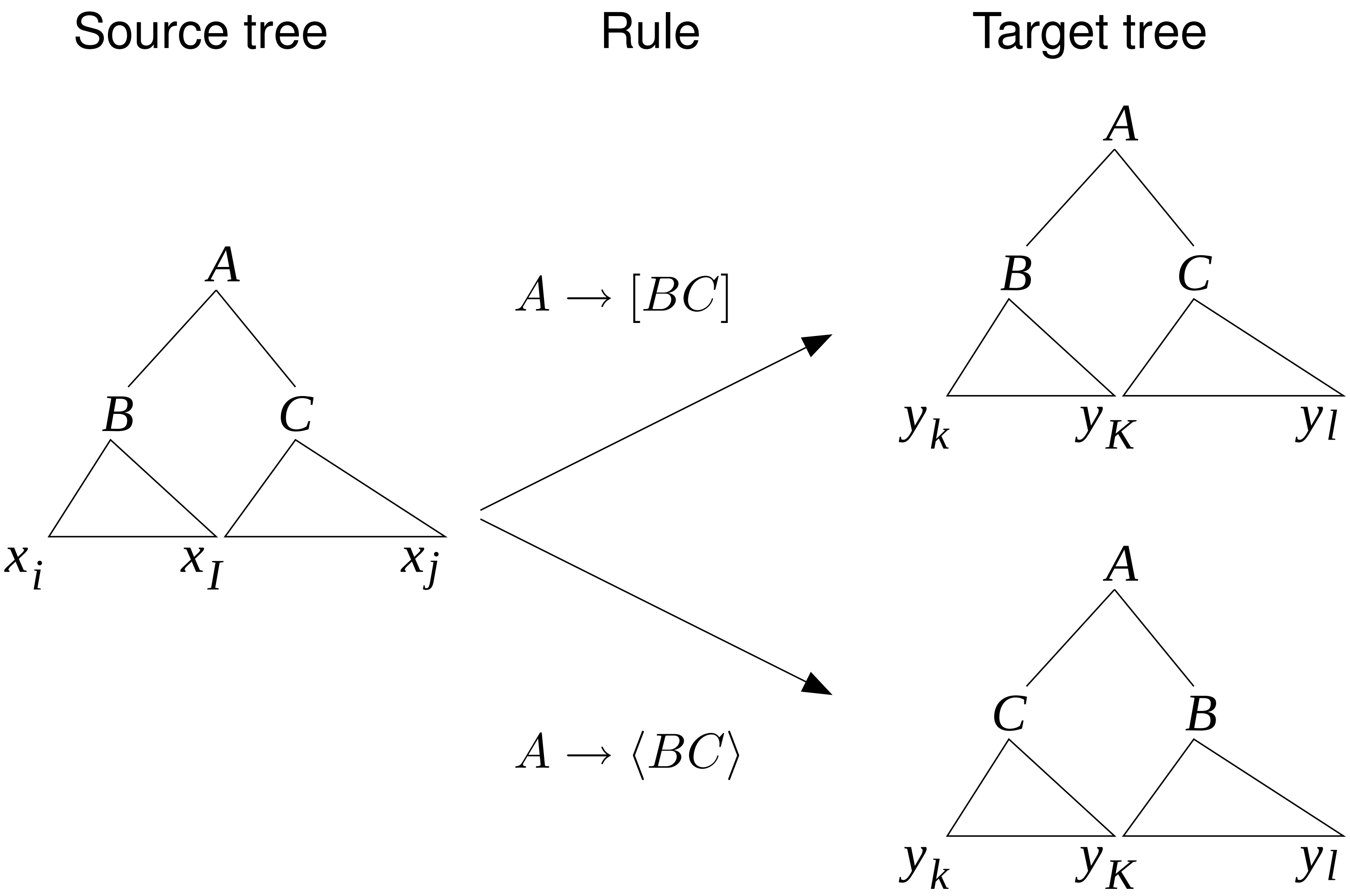
Phrase-Based models

- Systems implementing PB models are dominant in the state of the art.
- Basic translation units are bilingual phrases (segments), not single words.
- In training time, bilingual segments must be extracted: lots of techniques.
- Most common approach:
  - Heuristical extraction of phrases using word alignments.
  - Let be $(\mathbf{s}, \mathbf{t}) = x_{i+1}^{I}, y_{k+1}^{K}$
  - 5 models: $p_c(\mathbf{s}|\mathbf{t})$, $p_c(\mathbf{t}|\mathbf{s})$, $lex(\mathbf{s}|\mathbf{t})$, $lex(\mathbf{t}|\mathbf{s})$, C.

Stochastic Inversion Transduction Grammars

- Originally proposed by Dekai Wu.
- Closely related to context-free grammars.
- $\tau = (N, S, W_1, W_2, R, p)$, with:
  - N: set of non-terminal symbols.
  - $S \in N$: the axiom.
  - W_1: finite set of terminal symbols of language 1.
  - W_2: finite set of terminal symbols of language 2.
  - R: finite set of rules of type:
    - lexical rules: $A \rightarrow x/\epsilon$, $A \rightarrow \epsilon/y$, $A \rightarrow x/y$.
    - direct syntactic rules $A \rightarrow [BC]$
    - inverse syntactic rules $A \rightarrow \langle BC \rangle$
  - p: a function that determines the probability of each rule.
- Analyse two strings simultaneously.

SITG example



SITGs for phrase extraction

- Analyse two strings simultaneously.

  $\Rightarrow$ Can be used to extract segments.
  $\Rightarrow$ Take into account syntax-motivated restrictions.

- Original algorithm for parsing a sentence by Wu similar to CYK, $\mathcal{O}(|\mathbf{x}|^3|\mathbf{y}|^3|\mathbf{R}|)$

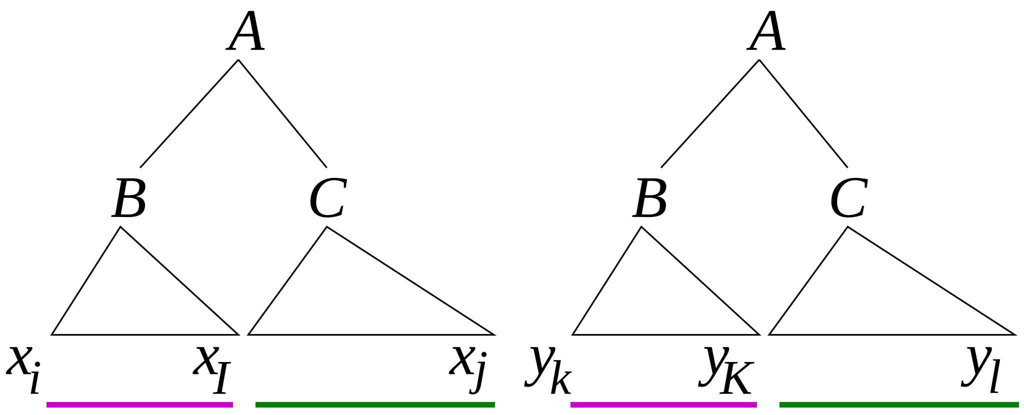
- Sánchez and Benedí, 2006: $\mathcal{O}(|\mathbf{x}||\mathbf{y}||\mathbf{R}|)$ when $\mathbf{x}$ and $\mathbf{y}$ are fully bracketed.

- Algorithm for phrase extraction:
  - Initial SITG built heuristically from word alignments.
  - Reestimation of probabilities with bracketed corpus to obtain improved SITG.
  - Training corpus parsed with SITG in order to obtain bilingual segments.
  - Inverse and direct translation probabilities:

$$p_c(\mathbf{s}|\mathbf{t}) = \frac{N(\mathbf{s},\mathbf{t})}{N(\mathbf{t})} \quad , \quad p_c(\mathbf{t}|\mathbf{s}) = \frac{N(\mathbf{s},\mathbf{t})}{N(\mathbf{s})}$$

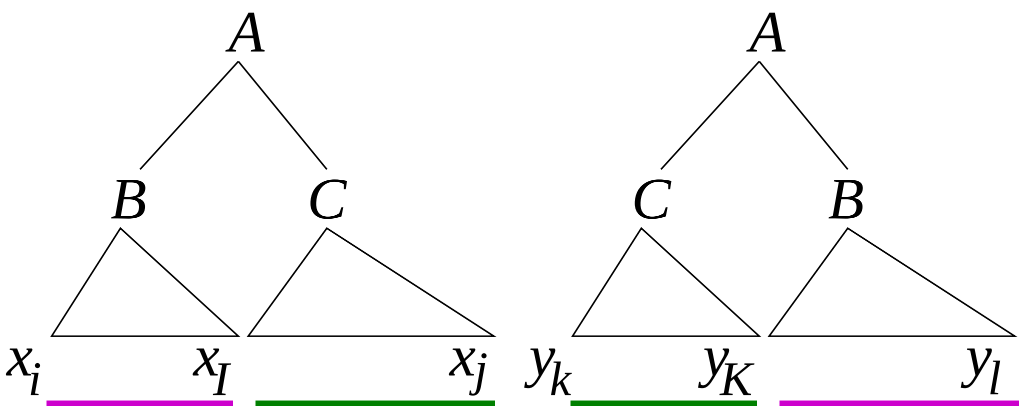
Phrase extraction example

Direct translation rule: $A \rightarrow [BC]$



$$\Rightarrow \begin{cases} \{x_{i+1} \ldots x_I, y_{k+1} \ldots y_K\} \\ \{x_{I+1} \ldots x_j, y_{K+1} \ldots y_l\} \end{cases}$$

Inverse translation rule: $A \rightarrow \langle BC \rangle$



$$\Rightarrow \begin{cases} \{x_{i+1} \ldots x_I, y_{K+1} \ldots y_l\} \\ \{x_{I+1} \ldots x_j, y_{k+1} \ldots y_K\} \end{cases}$$

Adding Syntactic Translation Probabilities

- When obtaining $\hat{T}_{\mathbf{x},\mathbf{y}}$, a subtree $\hat{T}_{\mathbf{s},\mathbf{t}}$ is obtained as well for a specific $(\mathbf{s},\mathbf{t})$
- This defines a joint probability $\hat{p}(\mathbf{s},\mathbf{t})$.
- Given that the corpus is bracketed, different $\hat{T}_{\mathbf{s},\mathbf{t}}$ may be obtained.
  $\Rightarrow$ different $\hat{p}(\mathbf{s},\mathbf{t})$ may exist.
- Let be Ω the multiset of spans obtained from a training sample.
- Let be $\Omega_{\mathbf{s},\mathbf{t}} \subseteq \Omega$ a multiset of $(\mathbf{s},\mathbf{t})$ spans.

$$\Rightarrow E_{\Omega}(\hat{p}(\mathbf{s},\mathbf{t})) = \frac{\displaystyle\sum_{\omega \in \Omega_{\mathbf{s},\mathbf{t}}} \hat{p}_{\omega}(\mathbf{s},\mathbf{t})}{|\Omega|}$$

$$\Rightarrow p_s(\mathbf{s}|\mathbf{t}) = \frac{E_{\Omega}(\hat{p}(\mathbf{s},\mathbf{t}))}{E_{\Omega}(\hat{p}(\mathbf{t}))} \quad \text{and} \quad p_s(\mathbf{t}|\mathbf{s}) = \frac{E_{\Omega}(\hat{p}(\mathbf{s},\mathbf{t}))}{E_{\Omega}(\hat{p}(\mathbf{s}))}.$$

Experimental results

- Corpus: Europarl

		Spanish	English
Training	Sentences	730K	
	Different pairs	716K	
	Vocabulary size	103K	64K
	Average length	21.5	20.8
Development	Sentences	2000	
	Average length	30.3	29.3
	Out of vocabulary	208	127
Devtest	Sentences	2000	
	Average length	30.2	29.0
	Out of vocabulary	207	125

Experimental results

- Translation results for a SITG with 1, 2, 3 and 4 non-terminal symbols.
- It. 0: Heuristically obtained SITG, only $p_c(\cdot|\cdot)$
- It. 1: One estimation iteration, $p_c(\cdot|\cdot)$
- + syntactic: adding $p_s(\cdot|\cdot)$

	It. 0		It. 1		+ syntactic	
non terms	BLEU	WER	BLEU	WER	BLEU	WER
1	26.8	62.5	26.9	62.6	27.7	61.6
4	26.6	63.2	27.9	61.5	28.9	60.0

Discussion

- Comparatively, best result reported so far with this technique was 23.0 BLEU.
- Best score obtained with Moses: 31.0 BLEU.
- with only direct and inverse models: 29.6 BLEU vs our 27.9 / 28.9.

  ⇒ Not directly comparable with Moses' best score: we have no lexical models.
  ⇒ Will add lexical models in the future.
  ⇒ Traditional PB models cannot obtain syntactic scores!
  ⇒ Moses best score uses 19M segment pairs, we use half that amount.

- Adding non-terminal symbols seems to improve.

Conclusions and ongoing/future work

- Conclusions:
  - Alternative, competitive method for phrase extraction.
  - Importance of parsed corpora for estimating SITG.
- Future work:
  - Add lexical probabilities.
  - Combine SITG's phrase table with Moses' phrase table.
  - Research ways to exploit reordering information in SITGs.

Questions? Comments? Suggestions?