A Syntax-based Statistical Machine Translation Model

Alexander Friedl, Georg Teichtmeister 4.12.2006

- **Introduction**
- The model
- Experiment
- Conclusion

- Statistical Translation Model (STM):

  - mathematical model

  - statistical modelling of human-language translation

  - parameters estimated with training corpus

- First steps

  - IBM 1998: string-to-string word-based translation

  - completely independent process

- Question: Why not word-based translation?

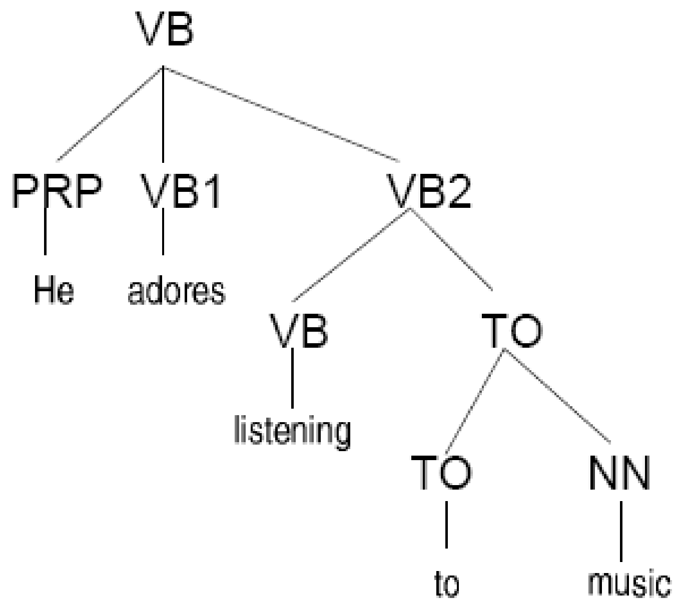
- no structural or synactic aspects

- how to handle different word order?

- Solution:

A Syntax-based Statistical Translation Model

- Input: parse tree (syntactic parser)



- Output: string

kare ha ongaku wo kiku no ga daisuki desu

- Three operations on each node of the tree:

  - reorder

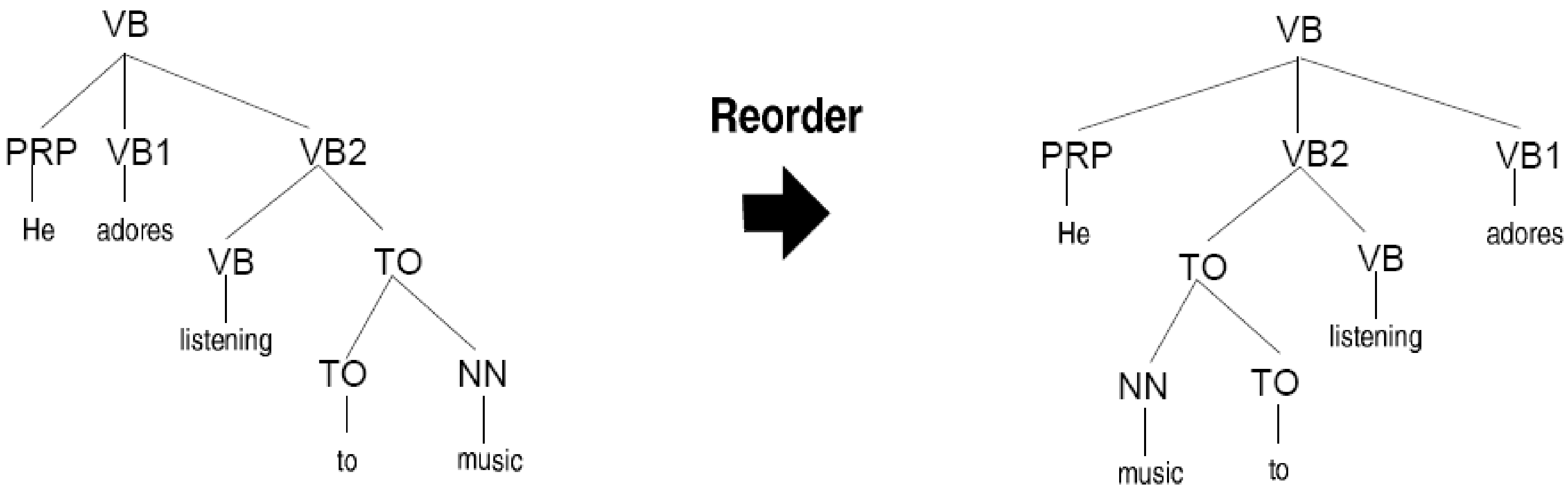
  - inserting

  - translating

- Reorder

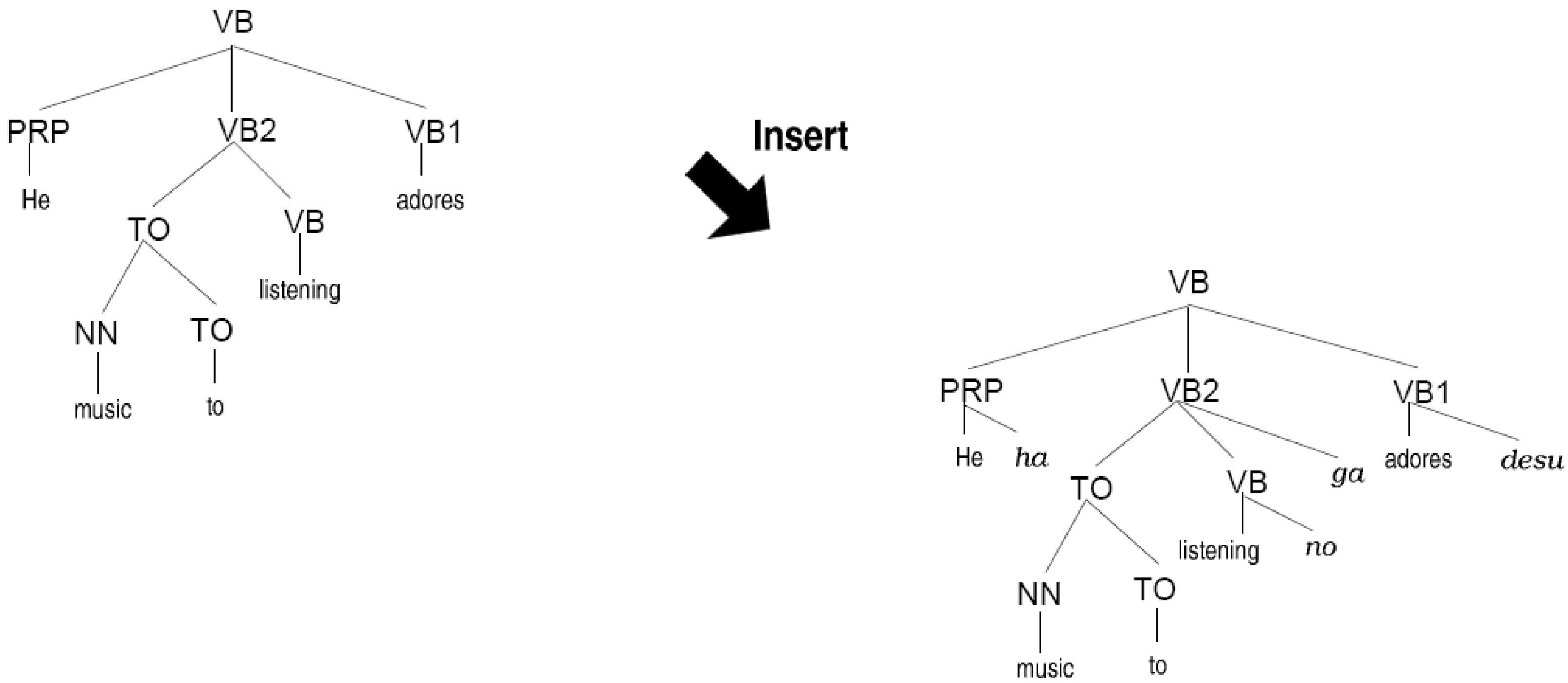
  - different word order

  - English vs. Japanese



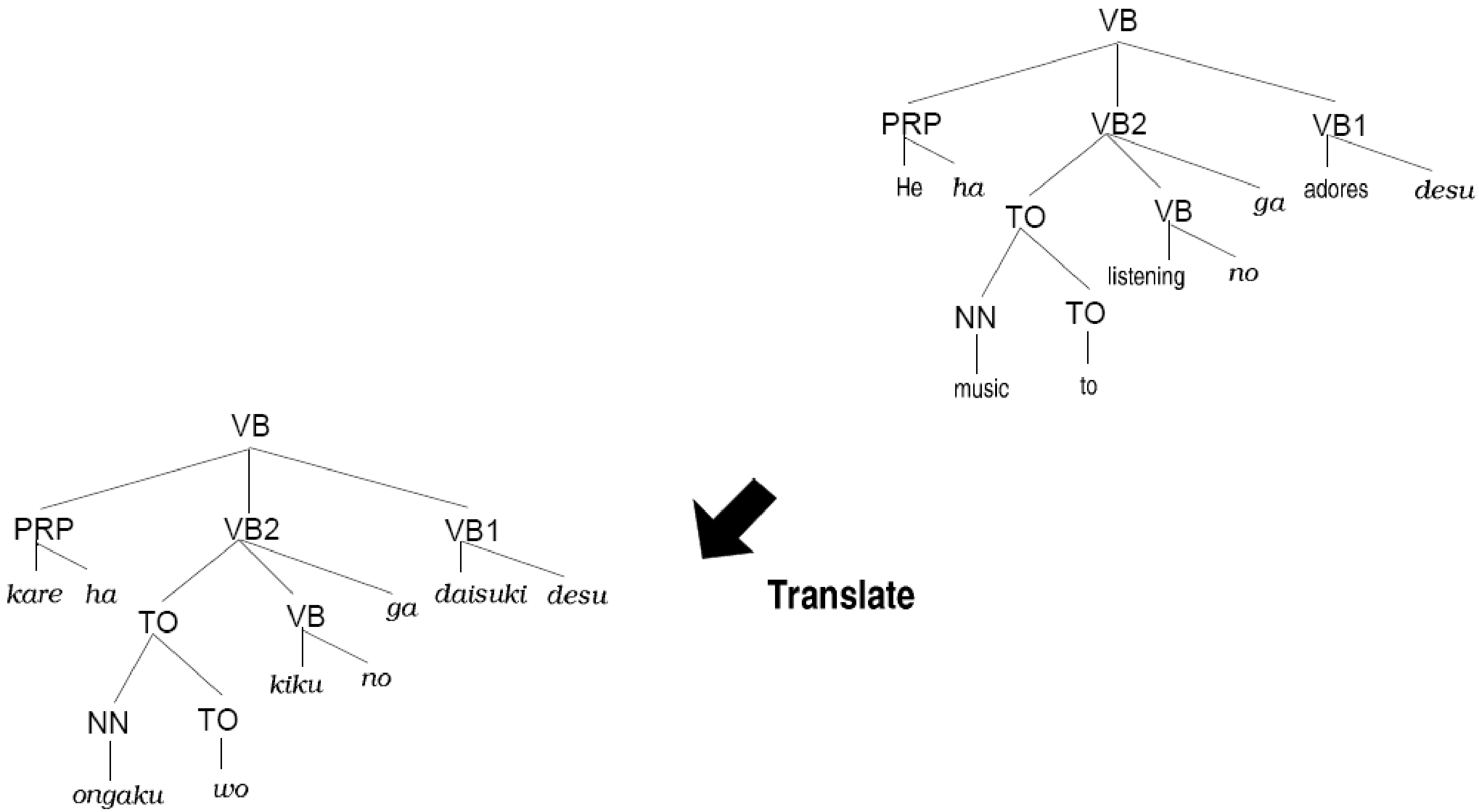
- Word-insertion

- e.g. capture linguistic differnces in specifying syntactic cases



- Translation

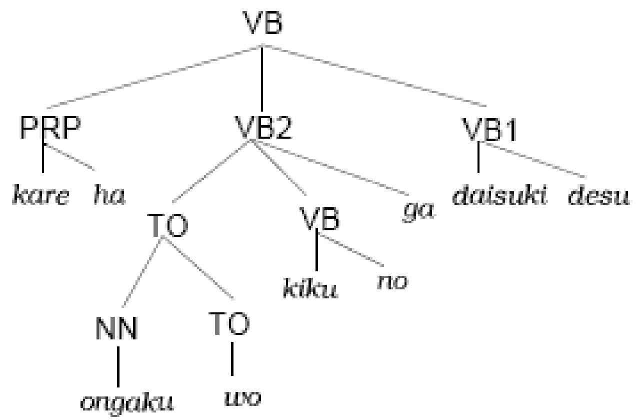
- Translate leaf words into the destination language



- Output

Reading off Leaves

kare ha ongaku wo kiku no ga daisuki desu ⬅



- Introduction
- **The model**
- Experiment
- Conclusion

- English parse tree into a noisy channel
- Output should be a Japanese sentence
- Stochastical operations an each nodes

- Reorder

  - N! possible reorderings for N children

  - probability given by the *r-table*

original order	reordering	P(reorder)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	0.723
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.251
	TO VB	0.749
TO NN	TO NN	0.107
	NN TO	0.893
⋮	⋮	⋮

r−table

- Word-insertion

- left, right oder nowhere

- probability given by the *n-table*

parent	TOP	VB	VB	VB	TO	TO	...
node	VB	VB	PRP	TO	TO	NN	...
P(None)	0.735	0.687	0.344	0.709	0.900	0.800	...
P(Left)	0.004	0.061	0.004	0.030	0.003	0.096	...
P(Right)	0.260	0.252	0.652	0.261	0.007	0.104	...

n−table

w	P(ins−w)
ha	0.219
ta	0.131
wo	0.099
no	0.094
ni	0.080
te	0.078
ga	0.062
⋮	⋮
desu	0.0007
⋮	⋮

- Translation

- dependent only on the word itself

- Probability given by the *t-table*

E	adores		he		i		listening		music		to		…
J	*daisuki*	1.000	*kare*	0.952	*NULL*	0.471	*kiku*	0.333	*ongaku*	0.900	*ni*	0.216	…
			NULL	0.016	*watasi*	0.111	*kii*	0.333	*naru*	0.100	*NULL*	0.204	
			nani	0.005	*kare*	0.055	*mi*	0.333			*to*	0.133	
			da	0.003	*shi*	0.021					*no*	0.046	
			shi	0.003	*nani*	0.020					*wo*	0.038	
			⋮	⋮	⋮	⋮					⋮	⋮	

t−table

- Total probability

- product of the single operation probabilities

- Tables

- English-Japanese training corpus

- Formal Transcription

Input:

English parse tree $\mathcal{E}$ (in nodes $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$)

Output:

French sentence $\mathbf{f}$ (in words $f_1, f_2, \ldots f_m$)

- Probability getting $\mathbf{f}$ for $\mathcal{E}$

$$P(\mathbf{f}|\mathcal{E}) = \sum_{\boldsymbol{\theta}:\mathrm{Str}(\boldsymbol{\theta}(\mathcal{E}))=\mathbf{f}} P(\boldsymbol{\theta}|\mathcal{E})$$

Str(Θ($\mathcal{E}$)) is a sequence of leaf words

- P(Θ|$\mathcal{E}$): Probability of a particular set of RVs in a parse tree

$$
\begin{aligned}
\mathrm{P}(\boldsymbol{\theta}|\mathcal{E}) &= \mathrm{P}(\theta_1, \theta_2, \ldots, \theta_n | \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \\
&= \prod_{i=1}^{n} \mathrm{P}(\theta_i | \varepsilon_i).
\end{aligned}
$$

- RVs $\theta_i = <\nu_i,\rho_i,\tau_i>$ independent on each other, but dependent on features of ε_i

$$
\begin{aligned}
\mathrm{P}(\theta_i|\varepsilon_i) &= \mathrm{P}(\nu_i, \rho_i, \tau_i|\varepsilon_i) \\
&= \mathrm{P}(\nu_i|\varepsilon_i)\,\mathrm{P}(\rho_i|\varepsilon_i)\,\mathrm{P}(\tau_i|\varepsilon_i) \\
&= \mathrm{P}(\nu_i|\mathcal{N}(\varepsilon_i))\,\mathrm{P}(\rho_i|\mathcal{R}(\varepsilon_i))\,\mathrm{P}(\tau_i|\mathcal{T}(\varepsilon_i)) \\
&= n(\nu_i|\mathcal{N}(\varepsilon_i))\,r(\rho_i|\mathcal{R}(\varepsilon_i))\,t(\tau_i|\mathcal{T}(\varepsilon_i))
\end{aligned}
$$

- Probability of getting a French sentence $\mathbf{f}$ given an English parse tree $\mathcal{E}$

$$\mathrm{P}(\mathbf{f}|\mathcal{E}) = \sum_{\boldsymbol{\theta}:\mathrm{Str}(\boldsymbol{\theta}(\mathcal{E}))=\mathbf{f}} \mathrm{P}(\boldsymbol{\theta}|\mathcal{E})$$

$$= \sum_{\boldsymbol{\theta}:\mathrm{Str}(\boldsymbol{\theta}(\mathcal{E}))=\mathbf{f}} \prod_{i=1}^{n} n(\nu_i|\mathcal{N}(\varepsilon_i))\, r(\rho_i|\mathcal{R}(\varepsilon_i))\, t(\tau_i|\mathcal{T}(\varepsilon_i))$$

- Automatic Estimation of model parameters

- update parameter to maximize the likelihood of the training corpus

- Algorithm

1. Initialize probability tables
2. Reset counters
3. For each iteration the number of events are counted and weighted by the probability of events
4. Parameter re-estimated by the counts

- Introduction
- The model
- **Experiment**
- Conclusion

- Experiment with small English-Japanese corpus

  - 2121 translation sentence pairs

  - taggers build the English parse trees

- Comparison with IBM 5

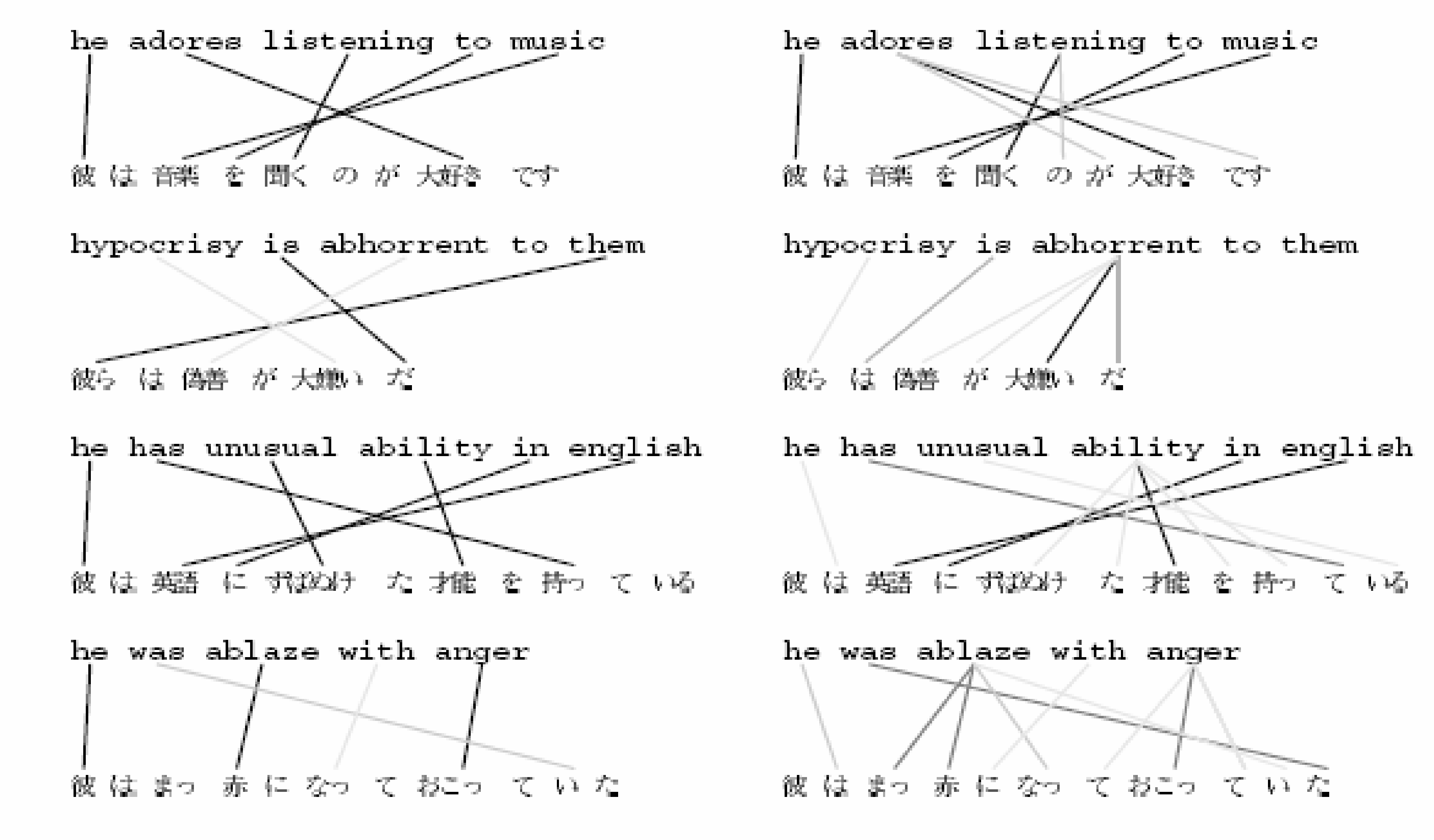
- Evaluation

  - generate the most probable alignment of the training corpus (Viterbi)

  - average score of the first 50 sentences

Alignment okay	1 point
Not sure	0,5 point
Alignment wrong	0 points

	Average score	Perfect sentences
Our Model	0,582	10
IBM Model 5	0,431	0


he adores listening to music
彼 は 音楽 を 聞く の が 大好き です
hypocrisy is abhorrent to them
彼ら は 偽善 が 大嫌い だ
he has unusual ability in english
彼 は 英語 に ずばぬけ た 才能 を 持っ て いる
he was ablaze with anger
彼 は まっ 赤 に なっ て おこっ て い た
he adores listening to music
彼 は 音楽 を 聞く の が 大好き です
hypocrisy is abhorrent to them
彼ら は 偽善 が 大嫌い だ
he has unusual ability in english
彼 は 英語 に ずばぬけ た 才能 を 持っ て いる
he was ablaze with anger
彼 は まっ 赤 に なっ て おこっ て い た

- Perplexity

- Our Model: 15,79
- IBM Model 5: 9,84

- Introduction
- The model
- Experiment
- **Conclusion**

- Syntax-based translation model
- Statistical modelling the translation process
- Syntactic information for languages with different word order
- Better alignment results in an experiment

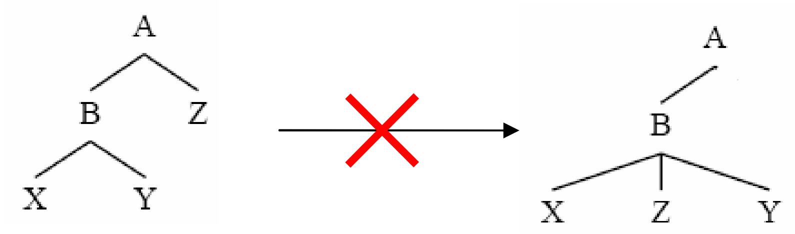
Outline – 2nd Part

- Problems of Tree to String
- Clone Operation
- Tree to Tree

- Phrasal Translation
- Translation System

Problem of Tree to String Model

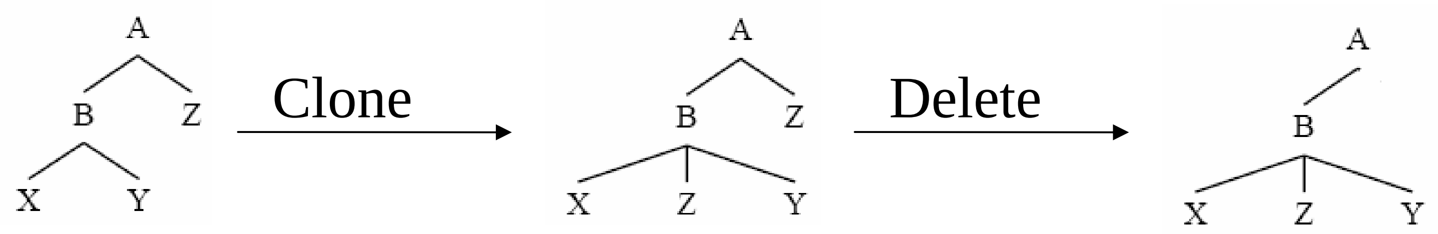
- Not all re-orderings of terminal nodes are possible



- Constrains syntactic correspondence between native and foreign language

The Clone operation

- Insert a copy of a subtree at any point in the tree
- Delete original node



When to clone?

- Should a clone be inserted as child of ε_j

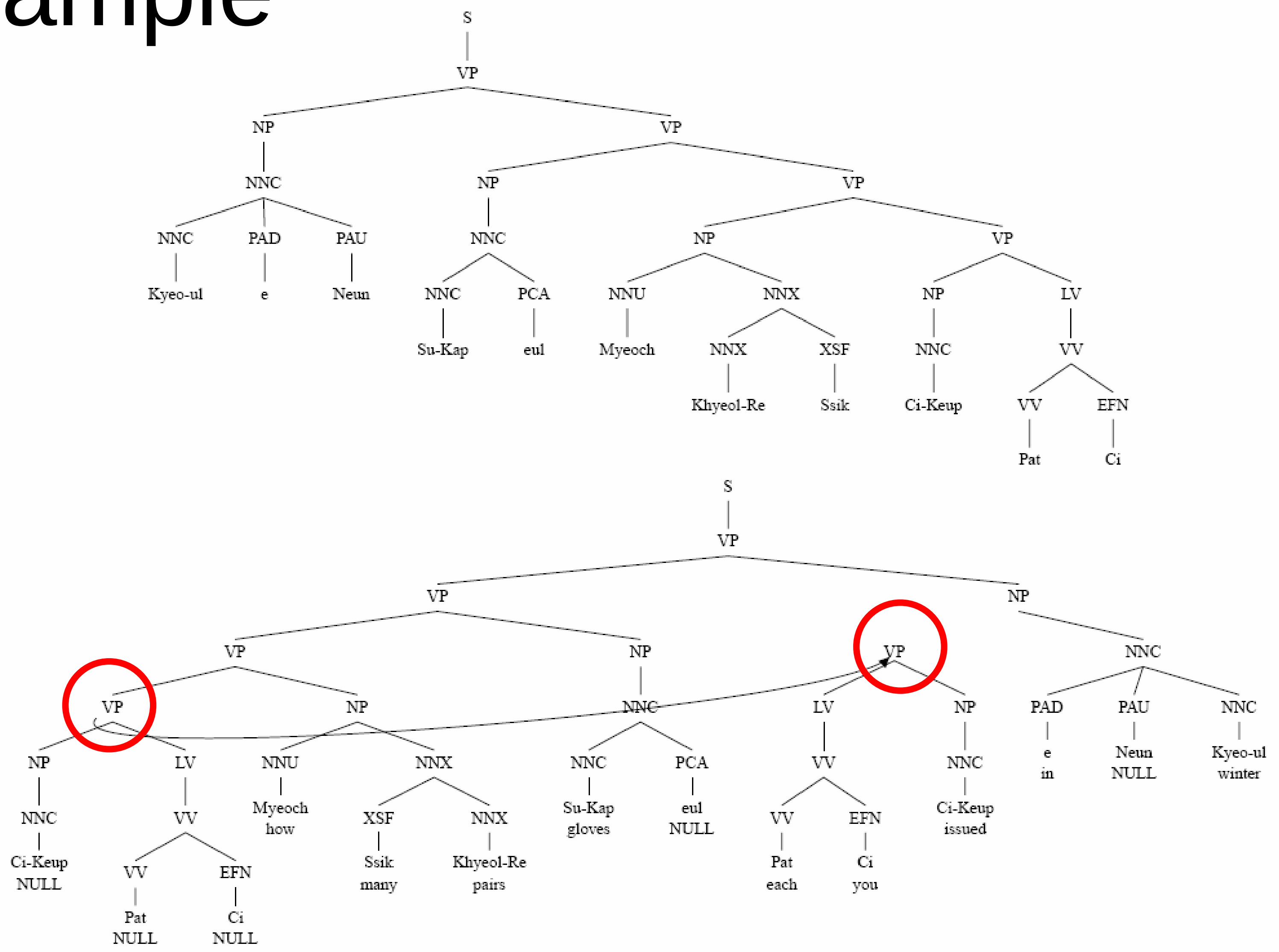
$$P_{ins}(clone \mid \varepsilon_j)$$

- Decide which node should be cloned

$$P_{clone}(\varepsilon_i \mid clone = 1) = \frac{P_{makeclone}(\varepsilon_i)}{\sum_k P_{makeclone}(\varepsilon_k)}$$

- Probability of cloning is independent of previous cloning operations

Example


S
VP
NP
VP
NNC
NP
VP
NNC
PAD
PAU
NNC
NP
VP
Kyeo-ul
e
Neun
NNC
PCA
NNU
NNX
NP
LV
Su-Kap
eul
Myeoch
NNX
XSF
NNC
VV
Khyeol-Re
Ssik
Ci-Keup
VV
EFN
Pat
Ci
S
VP
VP
NP
VP
NP
VP
NNC
VP
NP
NNC
LV
NP
PAD
PAU
NNC
NP
LV
NNU
NNX
NNC
PCA
VV
NNC
e
in
Neun
NULL
Kyeo-ul
winter
NNC
VV
Myeoch
how
XSF
NNX
Su-Kap
gloves
eul
NULL
VV
EFN
Ci-Keup
issued
Ci-Keup
NULL
VV
EFN
Ssik
many
Khyeol-Re
pairs
Pat
each
Ci
you
Pat
NULL
Ci
NULL

Tree – to - Tree

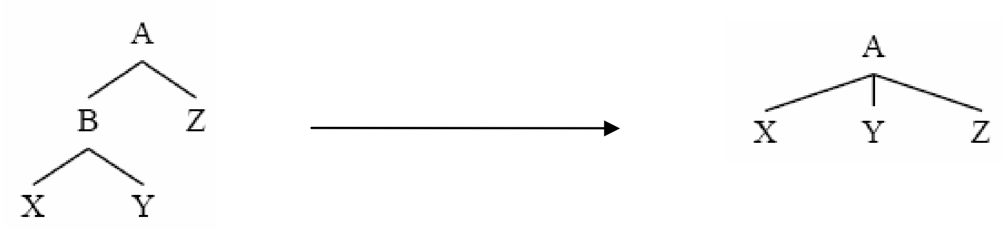
- Syntactic trees for foreign and native language
- Output tree instead of output string
- Additional tree transformations
  - Single source node → Two target nodes
  - Two source nodes → Single target node
- New model: $P(T_b \mid T_a)$

Building the output tree

- At each level at the output tree:
  - Choose a elementary tree
  - Align the children of the the elementary tree
- Translate the leaves

Elementary tree

- $P_{elem}(t_a \mid \varepsilon_a \Rightarrow children(\varepsilon_a))$
- Means that two nodes can grouped together
- *Example:*
  - *Nodes A and B are considered a elementary tree*

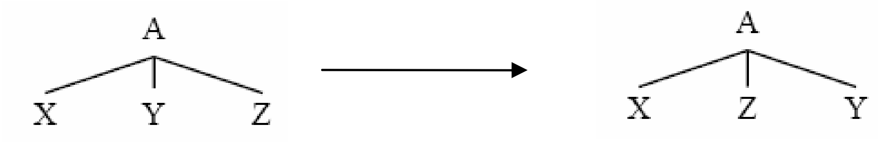


Alignment of children

- All children of an elementary tree are aligned at once according to:

$$P_{align}(\alpha \mid \varepsilon_a \Rightarrow children(t_a))$$

- Insertions and Deletions are also done in this step



Tree – to – Tree Clone

- Same reordering problems as Tree – to – String
- Cloning will now added to alignment step

$$P_{ins}(clone \in \alpha \mid \varepsilon_a \Rightarrow children(t_a))$$

$$P_{clone}(\varepsilon_i \mid clone \in \alpha) = \frac{P_{makeclone}(\varepsilon_i)}{\sum_k P_{makeclone}(\varepsilon_k)}$$

Parameter comparison

	Tree-to-String	Tree-to-Tree
elementary tree grouping		$P_{elem}(t_a \mid \varepsilon_a \Rightarrow children(\varepsilon_a))$
re-order	$P_{order}(\rho \mid \varepsilon \Rightarrow children(\varepsilon))$	$P_{align}(\alpha \mid \varepsilon_a \Rightarrow children(t_a))$
insertion	$P_{ins}(\text{left, right, none} \mid \varepsilon)$	α can include "insertion" symbol
lexical translation	$P_t(f \mid e)$	$P_t(f \mid e)$
with cloning	$P_{ins}(\text{clone} \mid \varepsilon)$	α can include "clone" symbol
	$P_{makeclone}(\varepsilon)$	$P_{makeclone}(\varepsilon)$

Experiments

- Data from Korean – English corpus (Military domain)
- Korean suffixes often carry meaning
  - This suffixes became leaves in the syntax tree
  - Vocabulary was reduced from 10059 to 3279
- Average Korean: 13 words and 21 tokens
- Average English: 16 words

Results

	Alignment Error Rate
IBM Model 1	.37
IBM Model 2	.35
IBM Model 3	.43
Tree-to-String	.42
Tree-to-String, Clone	.36
Tree-to-String, Clone $P_{ins} = .5$	.32
Tree-to-Tree	.49
Tree-to-Tree, Clone	.36

$$AER = 1 - \frac{2 |A \cap G|}{|A| + |G|}$$

Word pairings by System (A) and Gold Standard(G)

Phrasal Translation

- 1 to 1 word translation not perfect
- Compound nouns
  - German vs English
- Idiomatic phrases
  - “To kick the bucket” vs “Den Löffel abgeben”

Model

- 1 to 1 Model: $t(\tau \mid \mathrm{T}) = t(f \mid e)$
- 1 to N with fertility μ:

$$t(\tau \mid \mathrm{T}) = t(f_1 f_2 .. f_l \mid e) = \mu(l \mid e) \prod_{i=1}^{l} t(f_i \mid e)$$

- N to N:

$$ph(\phi \mid \Phi) = t(f_1 f_2 .. f_l \mid e_1 e_2 .. e_m) =$$

$$\mu(l \mid e_1 e_2 .. e_m) \prod_{i=1}^{l} t(f_i \mid e_1 e_2 .. e_m)$$

Incorporation into TM

- $P(\theta_i \mid \varepsilon_i) = \lambda_{\Phi_i} ph(\phi_i \mid \Phi_i) + (1 - \lambda_{\Phi_i}) r(\rho_i \mid R_i) n(\upsilon_i \mid N_i)$

- R: feature for reordering
- N: feature for insertion
- ρ: Reorder operation
- υ: Insert operation

From Models to the Translation

- Task: Translate from **F**oreign to **E**nglish

$$P(E \mid F)$$

- Reformulation

$$P(E \mid F) = P(F \mid E)P(E)$$

Language Model

- Probability of an English Sentence P(E)
- N-gram LM in original Implementation
- No syntactic Information used
- Improvement: Immediate-head parsing for language models

Immediate-head parsing

- English Sentence
- Non-lexical PCFG → Large Parse Forest
- Pruning of the Large Parse Forest
  - Which edges have high probability of being correct?
- Evaluation of pruned Parse with a lexical PCFG

Decoder

- Decoder works in reverse direction to TM
- Find most probable syntactic tree E from a Sentence F
- Basic idea: “Translate“ Parse tree using TM to the foreign language
- Parse the foreign sentence
- “Translate” back to English
- And check LM

Conclusion

- Improvements to Syntactic Translation Model
  - Tree to String Clone
  - Tree to Tree
  - Tree to Tree Clone
  - Phrasal Translation
- Brief overview over Translation Process

References

- A Syntax based Statistical Translation Model

  Kenji Yamada and Kevin Knight, Proceedings of the Conference of the Association for Computational Linguistics 2001
- Loosely Tree-based Alignment for Machine Translation

  Daniel Gildea, In Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03), Supporo, Japan, 2003.
- Syntax baed Language Models for Statistical Machine Translation

  Eugene Charniak, Kevin Knight and Kenji Yamada, In MT Summit IX. Intl. Assoc. for Machine Translation, 2003
- A Decoder for Syntax-Based Statistical MT

  Kenji Ymada and Kevin Knight, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02) , 2002

Thank you for your attention!