Machine Translation
Course 7

Diana Trandabăț

Academic year 2022-2023

What goes wrong?

- We see many errors in machine translation when we only look at the word level
  - Missing content words
    - MT: Condemns US interference in its internal affairs.
    - Human: **Ukraine** condemns US interference in its internal affairs.
  - Verb phrase
    - MT: Indonesia said that oppose the presence of foreign troops.
    - Human: Indonesia reiterated its opposition to foreign military presence.

What goes wrong?

- Wrong dependencies
  - MT: …, particularly those who cheat the audience the players.
  - Human: …, particularly those **players who cheat the audience.**
- Missing articles
  - MT: …, he is fully able to activate team.
  - Human: …, he is fully able to activate **the** team.

What goes wrong?

- Word salad:
  - the world arena on top of the u . s . sampla competitors , and since mid – July has not appeared in sports field , the wounds heal go back to the situation is very good , less than a half hours in the same score to eliminate 6:2 in light of the south African athletes to the second round .

as opposed to letter salad

How can we improve?

- Relying on language model to produce more ‘accurate’ sentences is not enough
- Many of the problems can be considered ‘syntactic’
- Perhaps MT-systems don’t know enough about what is important to people
- So, include syntax into MT
  - Build a model around syntax, or
  - Include syntax-based features in a model

Syntax-based translation

- One criticism of the phrase-based MT is that it does not model structural or syntactic aspects of the language.
- Syntax based MT uses parse trees to capture linguistic differences such as word order and case marking.
- Reordering for syntactic reasons
  - e.g., move German object to end of sentence
- Better explanation of function words
  - e.g., prepositions, determiners
- Conditioning to syntactically related words
  - translation of verb may depend on subject or object
- Use of syntactic language models

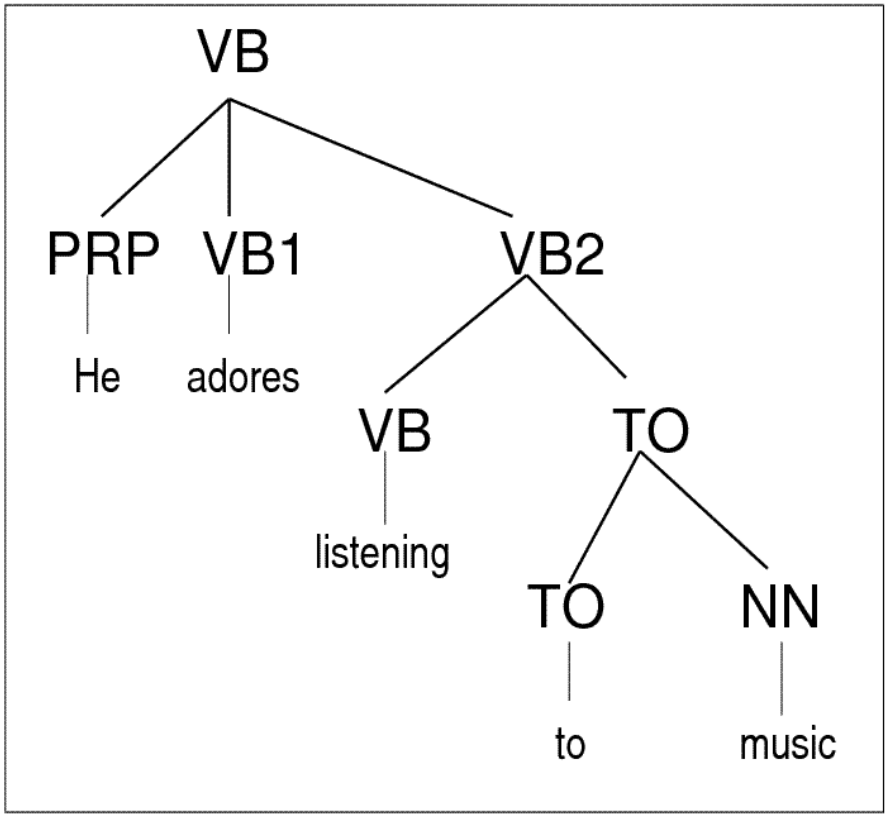
Syntax-based MT

- You have a sentence and its parse tree
- The children at each node in the tree are rearranged
- New nodes may be inserted before or after a child node
- These new nodes are assigned a translation
- Each of the leaf lexical nodes is then translated

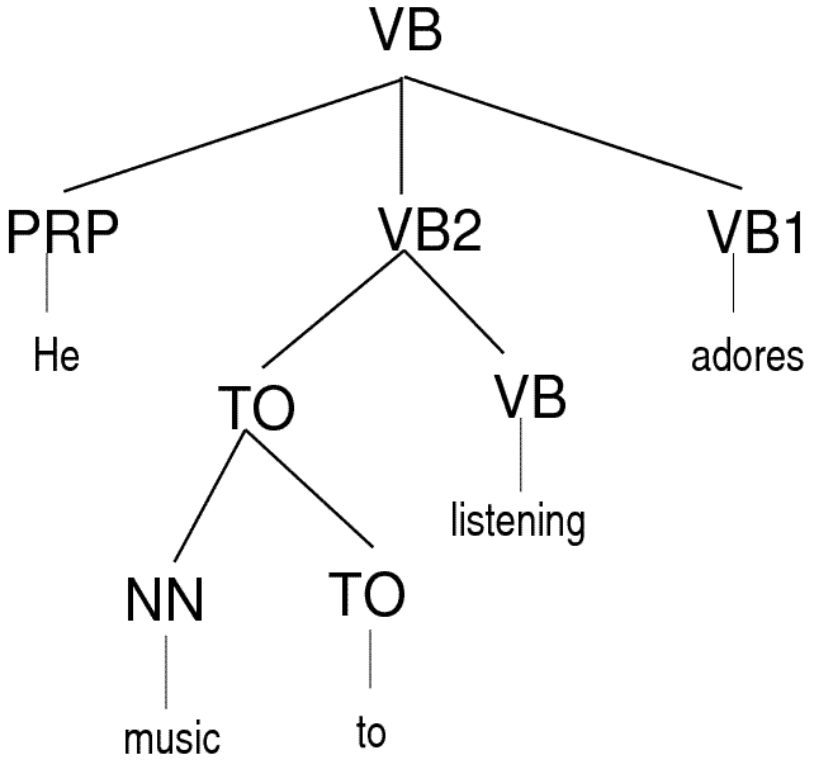
A syntax-based model

- Assume word order is based on a reordering of source syntax tree > *Reorder*
- Assume null-generated words happen at syntactical boundaries > *Insert*
- (For now) Assume a word translates into a single word > *Translate*

Reorder

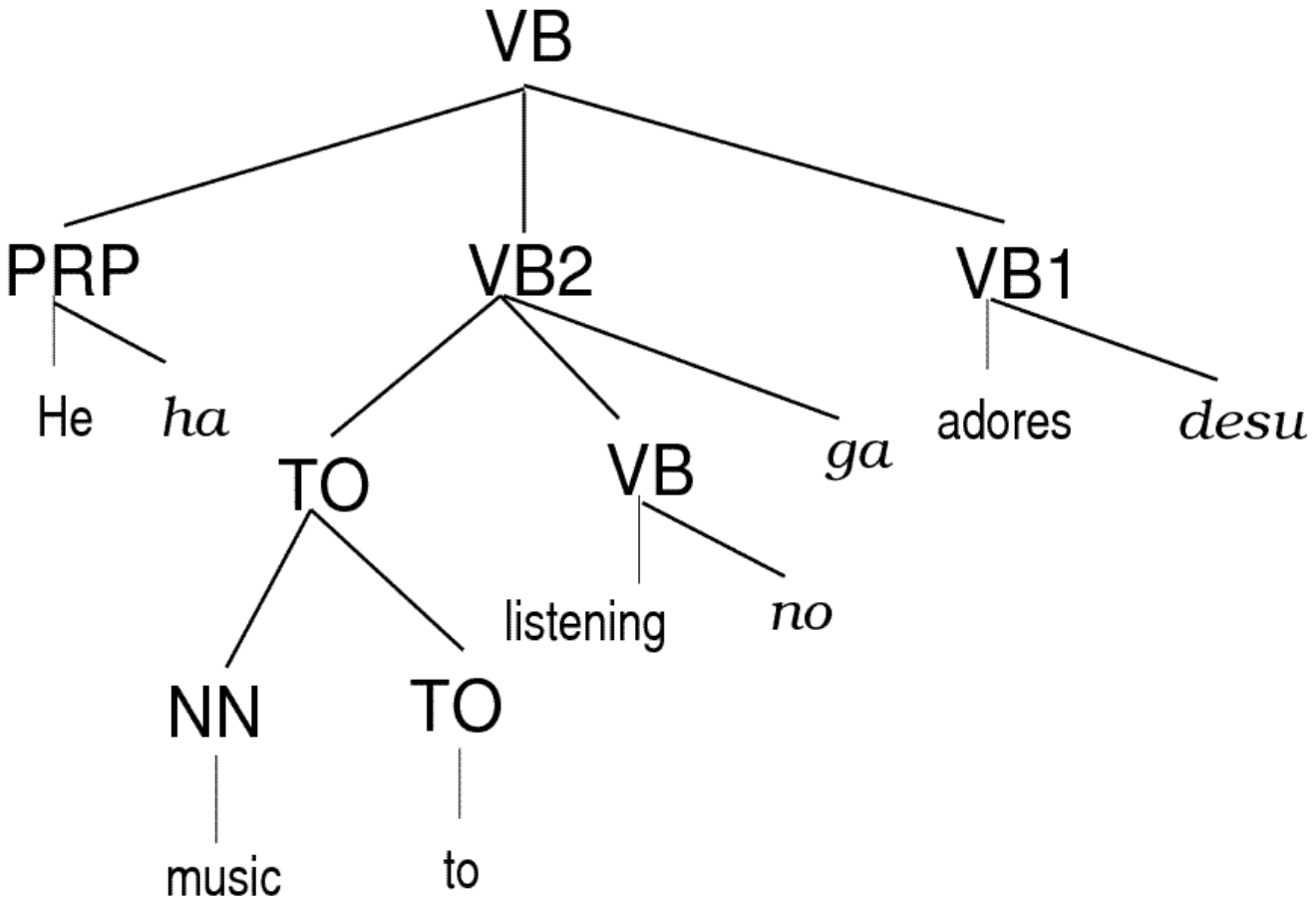


1. Original Parse Tree



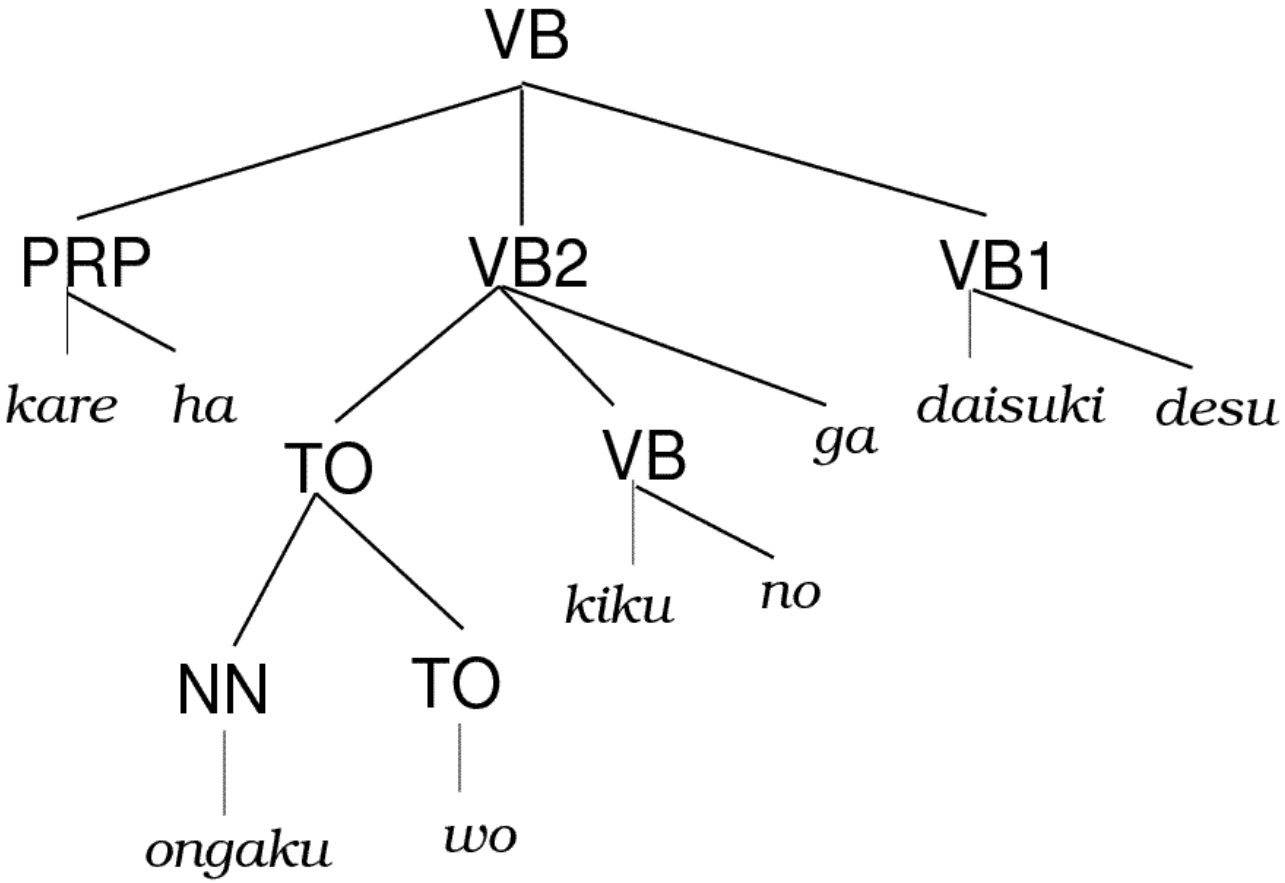
2. After Reorder

Insert



3. After Insert

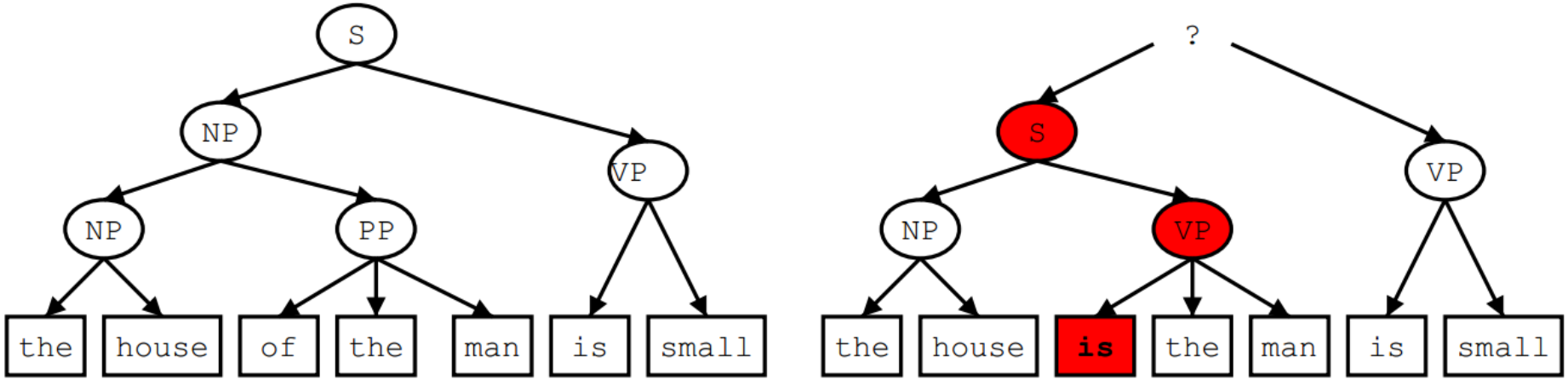
Translate



4. After Translate

Syntactic language models

- Good syntax tree => good target language
- Allows for long distance constraints



Parameters

- Reorder (R) – child node reordering
  - Can take any possible child node reordering
  - Defines word order in translation sentence
  - Conditioned on original child node order
  - Only applies to non-leaf nodes

Parameters cont.

- Insertion (N) – placement and translation
  - *Left*, *right*, or *none*
  - Defines word to be inserted
  - Place conditioned on current and parent labels
  - Word choice is unconditioned

Parameters cont.

- Translation (T) – 1 to 1
  - Conditioned only on source word
  - Can take on null
- Translation (T) – N to N
  - Consider word fertility (for 1-to-N mapping)
  - Consider phrase translation at each node
  - Limit size of possible phrases
  - Mix phrasal w/ word-to-word translation

Do we need the entire model to be based on syntax?

- Good performance increase
- Large computational cost
  - Many permutations to CFG rules
- How about trying something else?
  - Add syntax-based features that look for more specific things

Syntax-based Features

- Shallow
  - POS and Chunk Tag counts
  - Projected POS language model
- Deep
  - Tree-to-string
  - Tree-to-tree
  - Verb arguments

Shallow Syntax-Based Features

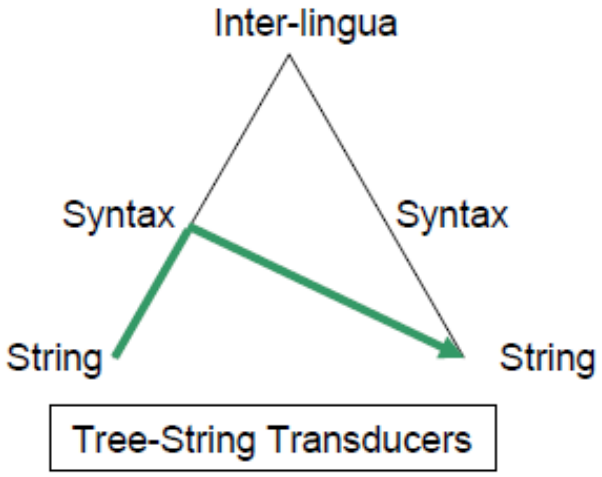
- POS and chunk tag count
  - Low-level syntactic problems with baseline system. Too many articles, commas and singular nouns. Too few pronouns, past tense verbs, and plural nouns.
  - Reranker can learn balanced distributions of tags from various features
  - Examples
    - Number of NPs in English
    - Difference in number of NPs between English and Chinese
    - Number of Chinese N tags translated to only non-N tags in English.

Shallow Syntax-Based Features

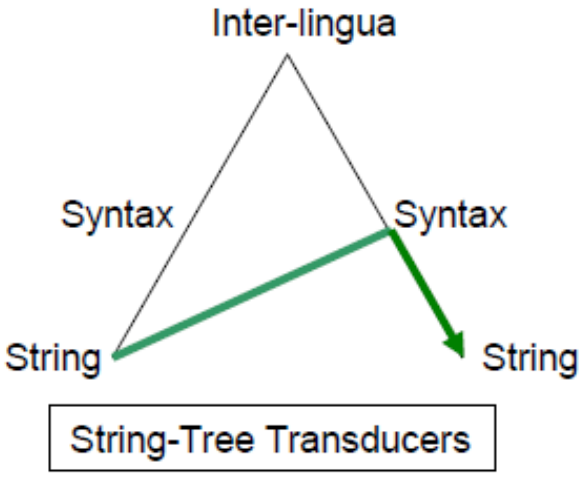
- Projected POS language model
  - Use word-level alignments to project Chinese POS tags onto the English words
    - Possibly keeping relative position within Chinese phrase
    - Possibly keeping NULLs in POS sequence
    - Possibly using lexicalized NULLs from English word
  - Use the POS tags to train a language model based on **POS N-grams**

CD_+0_M_+1	NN_+3	NN_-1	NN_+2_NN_+3
14 (*measure*)	open	border	cities

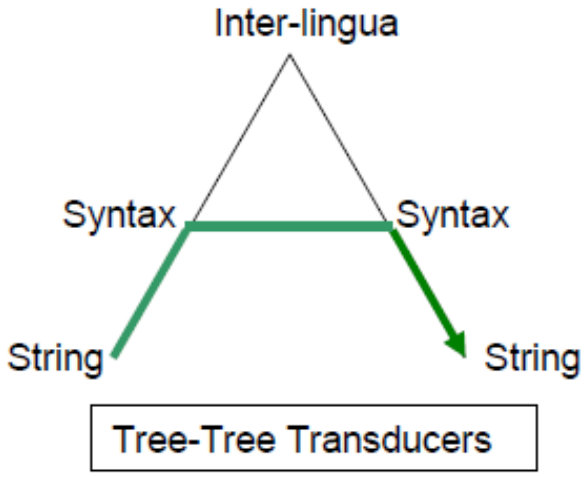
Deep Syntax-based MT



Tree-String Transducers



String-Tree Transducers



Tree-Tree Transducers

Deep Syntax-Based Features

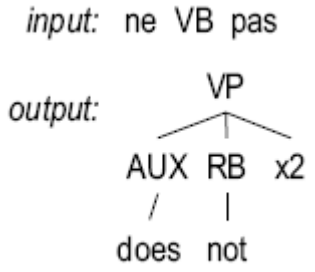
- **Tree to string**
  - Models explain how to transduce a **structural representation** of the source language input into **a string** in the target language
  - During decoding:
    - Parse the source string to derive its structure
    - Decoding explores various ways of decomposing the parse tree into a sequence of composable models, each generating a translation string on the target side
    - The best-scoring string can be selected as the translation

No.	Rule		
(1)	(IP (NP) (VP) (PU))	$X_1\ X_2\ X_3$	1:1 2:2 3:3
(2)	(NP (NN 枪手))	The gunman	1:1 1:2
(3)	(VP (SB 被) (VP (NP (NN)) (VV 击毙)))	was killed by X	1:1 2:4 3:2
(4)	(NN 警方)	police	1:1
(5)	(PU 。)	.	1:1

Deep Syntax-Based Features

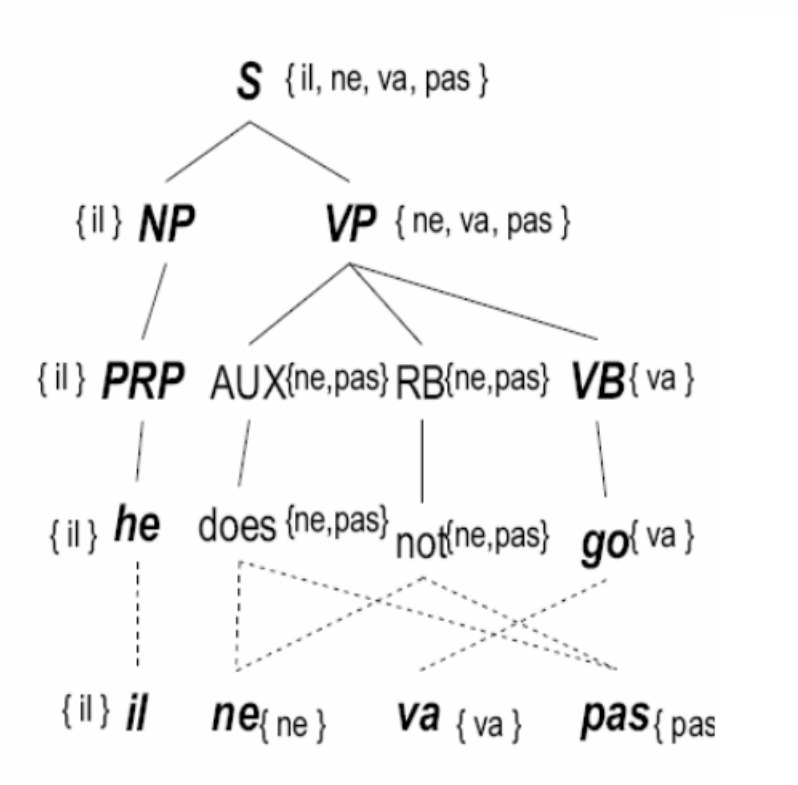
- **String-to-Tree:**
  - Models explain how to transduce **a string** in the source language into a **structural representation** in the target language
  - During decoding:
    - No separate parsing on source side
    - Decoding results in set of possible translations, each annotated with syntactic structure
    - The best-scoring *string + structure* can be selected as the translation

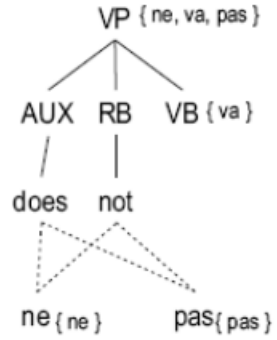
ne VB pas → (VP (AUX (does) RB (not) x2

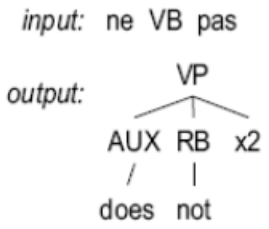


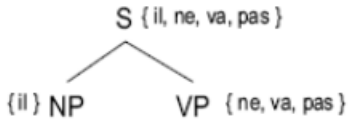
String-to-Tree

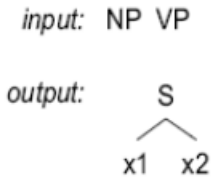
- Learn a direct translation model from word-level aligned corpus
- Extract reordering patterns











Deep Syntax-Based Features

- Tree to Tree
  - Models explain how to transduce a **structural representation** of the source language input into **a structural representation** in the target language
  - During decoding:
    - Decoder **synchronously** explores alternative ways of parsing the source-language input string and transduce it into corresponding target-language structural output.
    - The best-scoring structure + structure can be selected as the translation

Tree to Tree cont.

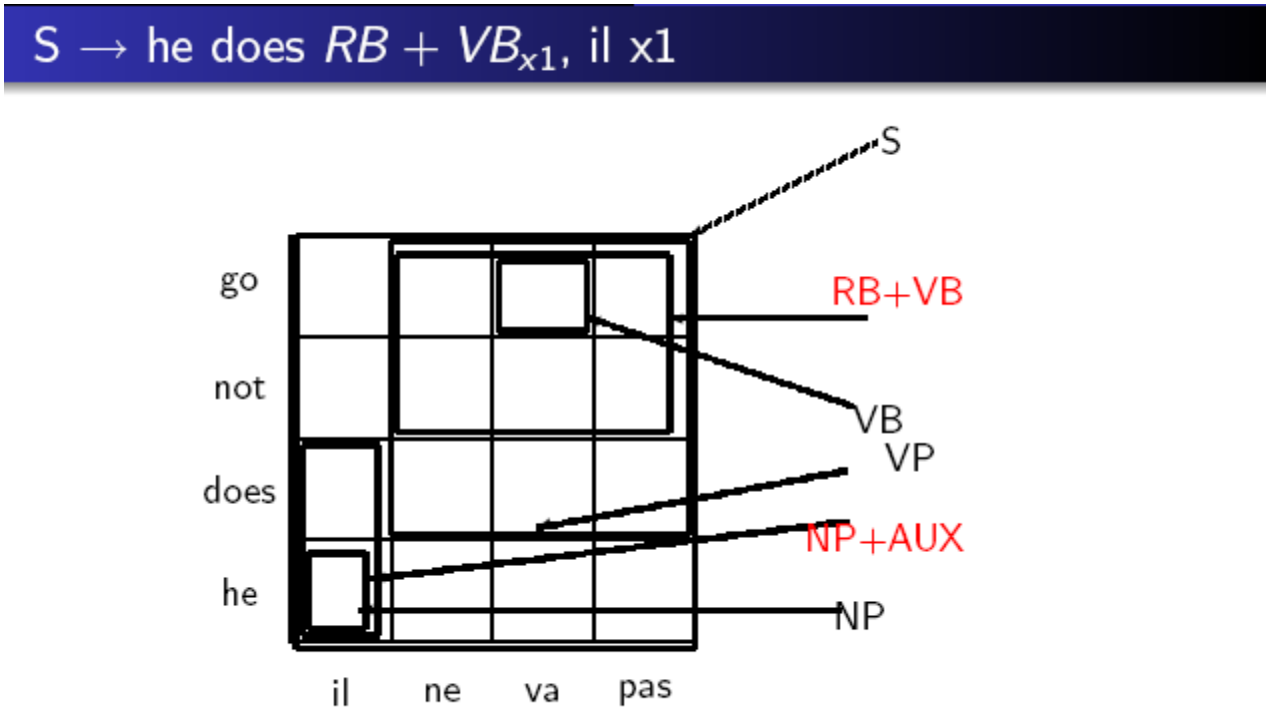
- At each level of the tree:
  1. At most one of the current node's children is grouped with the current node into a single elementary tree with its probability conditioned on the current node and its children.
  2. An alignment of the children of the current elementary tree is chosen with its probability conditioned on the current node an the children of child in the elementary tree. This is similar to the reorder operation in the tree-to-string model, but allows for node addition and removal.
- Leaf-level parameters are ignored when calculating probability of tree-to-tree.

Verb Arguments

- Idea: A feature that counts the difference in the number of arguments to the main verb between the source and target sentences
- Perform a breadth-first search traversal of the dependency trees
  - Mark the first verb encountered as the main verb
  - The number of arguments is equal to the number of its children
  - Account for differences in the number of arguments

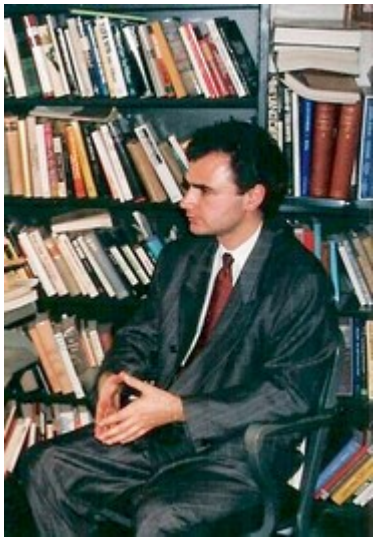
Syntax-augmented Phrase based MT

- Similar to phrase-based machine translation, but includes syntax in the creation of phrases.



“The world cannot be translated;
It can only be dreamed of and touched.”

Dejan Stojanovic, *The Creator*



As opposed to “letter salad”

- I cnduo't bvleiee taht I culod aulaclty uesdtannrd waht I was rdnaieg. Unisg the icndeblire pweor of the hmuan mnid, aocdcrnig to rseecrah at Cmabrigde Uinervtisy, it dseno't mttaer in waht oderr the lterets in a wrod are, the olny irpoamtnt tihng is taht the frsit and lsat ltteer be in the rhgit pclae. The rset can be a taotl mses and you can sitll raed it whoutit a pboerlm. Tihs is bucseae the huamn mnid deos not raed ervey ltteer by istlef, but the wrod as a wlohe. Aaznmig, huh? Yaeh and I awlyas tghhuot slelinpg was ipmorantt! See if yuor fdreins can raed tihs too.

back