CMPT-825

Natural Language Processing

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

Machine Translation

- One of the earliest applications in computational linguistics
- Not a natural task: understanding language is something humans do naturally, but humans use experts trained in translation
- However, automatic real-time open vocabulary translation between languages has immense implications

Machine Translation

- It is clear that full translation of arbitrary text, such as literature and poetry for instance, is well beyond our reach

- However, given the language analysis algorithms we have seen so far we are in a position to deal with translation in the following settings:

  – tasks for which a **rough translation** is adequate – for example, on the web to find information,
  e.g. cross-language information retrieval: search in Chinese but find English pages

- tasks were a **human post-editor** can be used to improve automatic translation output

- tasks limited to **sublanguage** domains: e.g. weather forecasts (Montreal), computer manuals, etc.

- the simplest translation task is a phonetic transfer from one language to another: **transliteration**

Machine Translation

- Languages differ in many aspects: morphology, lexical variations, sentence structure, implicit information (usually missing in one language but not the other)

- However, many differences might not be relevant to the problem of machine translation
  unless the **Sapir-Whorf hypothesis** is true (language constrains or shapes thought) – all evidence points to this hypothesis as being incorrect
  (cf. *Languages of Pao* by Jack Vance)

Language Differences

- Lexical variation: **morphological differences**

E:	the	man 's	house
H:	az	ember	ház a
	the	man	house his

- Prepositions vs. Postpositions

  E: *to Kiki* vs. J: *Yuriko ni*

- Missing items: determiners, verbal information (tense, aspect, mood)

Language Differences

- Lexical variation: **lexical gaps**

E: brother	J: *otooto* (younger) J: *onisaan* (older)
E: wall	G: *Wand* (inside) G: *Mauer* (outside)
E: know	F: *connaître* (be acquainted with) F: *savoir* (know a proposition)
E: they	F: *ils* (masc) F: *elles* (fem)
G: *berg*	E: hill E: mountain
M: *tā*	E: he, she or it

Language Differences

- Differences in the sentence structure:

E:	the	bottle	floated	out
S:	la	botella	salió	flotando
	the	bottle	exited	floatingly

- Language differences specifically for MT: **structural divergences** (Dorr, 1994)

- E: I like Mary $\Leftrightarrow$ S: María me gusta a mi 'Mary pleases me'
- E: John usually goes home $\Leftrightarrow$ S: Juan suele ir a casa 'John tends to go home'
- E: I like eating $\Leftrightarrow$ G: Ich esse gem 'I eat likingly'
- E: John entered the house $\Leftrightarrow$ S: Juan entró en la casa 'John entered in the house'
- E: I stabbed John $\Leftrightarrow$ S: Yo le di puñaladas a Juan 'I gave knife-wounds to John'
- E: I am hungry $\Leftrightarrow$ G: Ich habe Hunger 'I have hunger'
- E: John broke into the room $\Leftrightarrow$ S: Juan forzó la entrada al cuarto 'John forced (the) entry to the room'

Machine Translation

- Three main models for machine translation:
  - **Interlingua** Model: (for n languages, map each language to and from a common interlingua (a language independent canonical form representing meaning for MT)
  - **Transfer** Model: (for each language pair, describe an **analysis, transfer, generation** transfer from one language to another) sometimes the transfer is reversible so that for n languages, there are n^2 transfer pairs
  - **Direct Translation** Model: approach for a single language pair, different models for different pairs, still using the language analysis algorithms for source/target but with ad-hoc transfer

Interlingua Model

- Translation as a process of extracting the meaning of the input and then expressing that meaning in the target language

- We need to map from the source language into a general meaning representation called **interlingua**, and then map from this interlingua into the target language

- Interlingua tries to represent all sentences that mean the “same” thing in the same way, regardless of the language

Interlingua Model

- One key aspect of interlingua is the identification of the **predicate-argument** structure in the source language and then map this into the equivalent predicate argument structure in the target language

- Feature structures are the common representation for predicate argument structure
  Some explicitly use first order logic

Interlingua Model

- Sounds attractive, but requires overwhelming work in deciding exactly what should go into the design of the interlingua (see, for example, the UNL interlingua)

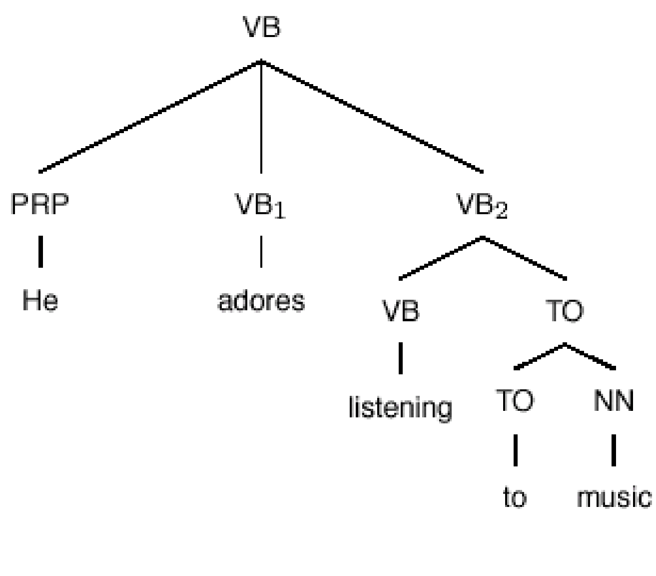
- The main problem with interlingua is the requirement to **fully disambiguate** at all times
  some aspects of a particular language pair might not be relevant for another language pair, but interlingua has to deal with both cases at once

- Other approaches to MT typically do not try to fully disambiguate, rather they try to **preserve ambiguity** where possible

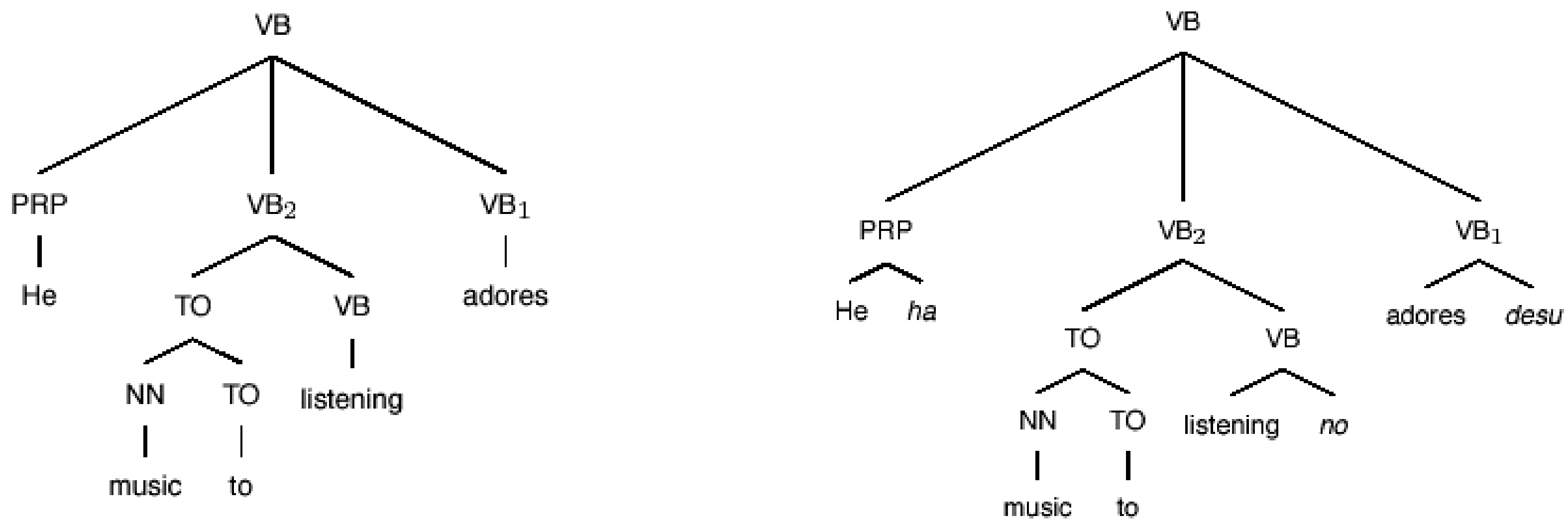
Transfer Model

- Write particular rules to transfer words, phrases, and sentence structures between a pair of languages
- Lexical transfer model $\Rightarrow$ Finite-state transducers
- More general kinds of transducers: **Synchonous Context-free Grammars**
- Sometimes, the transfer model can be used to *over-generate* multiple candidate output translations; from this output, one is selected using constraints in the target language (sometimes called **shake-and-bake** translation)

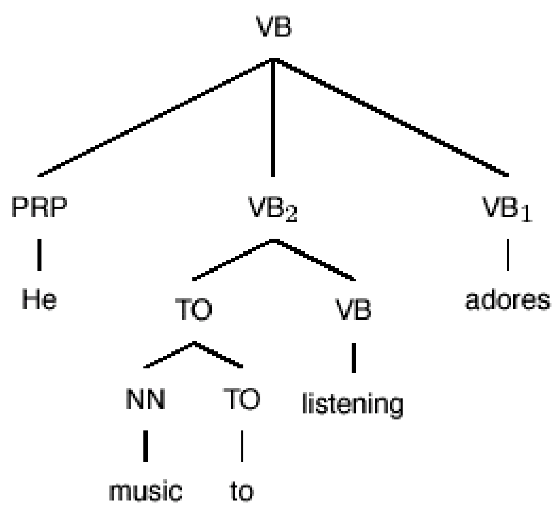
Transfer Model

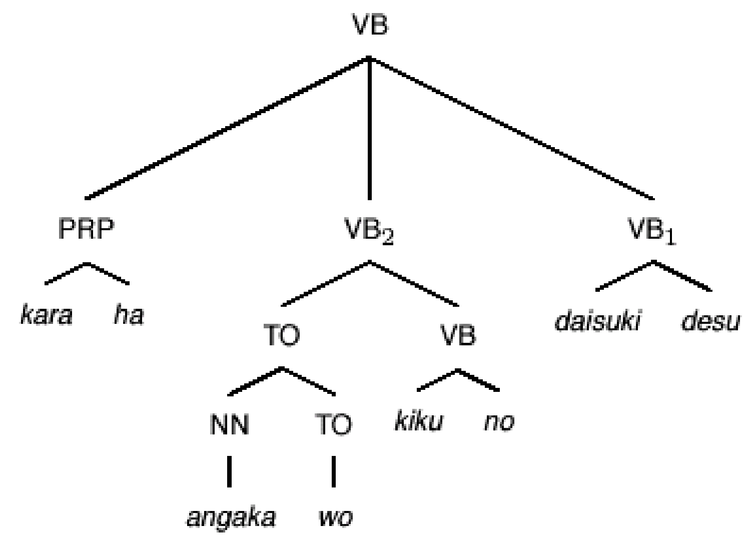


Transfer Model



Transfer Model





Direct Translation Model

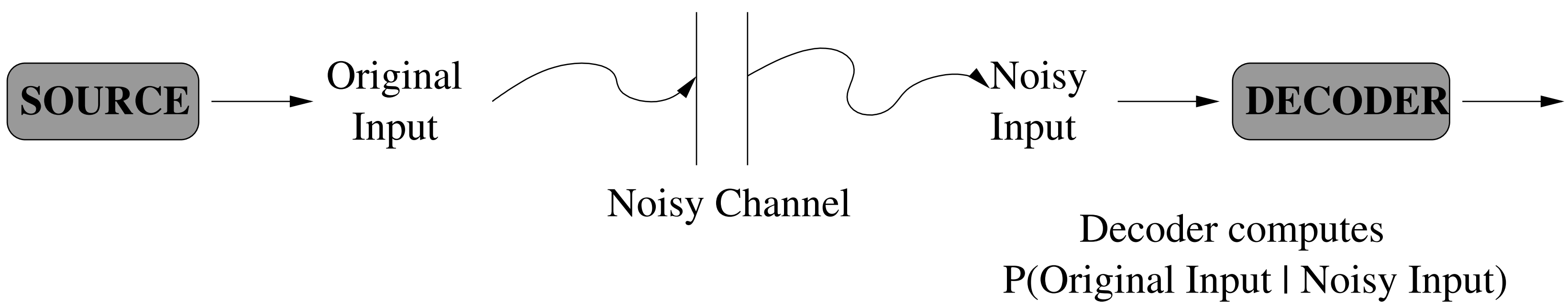
- Six stages for a Direct translation MT system

  1. morphological analysis
  2. lexical transfer of content words (including transliteration)
  3. various work relating to prepositions
  4. subject-verb-object order has to be rearranged
  5. add determiners, fix order of prepositions
  6. morphological generation

- For example, given a sentence in Japanese:

  1. watashihatsukuenouenopenwojonniageta
  2. watashi ha tsukue no ue no pen wo jon ni ageru PAST
  3. I ha desk no ue no pen wo John ni give PAST
  4. I give PAST pen on desk John to
  5. I give PAST the pen on the desk to John
  6. I gave the pen on the desk to John

Noisy Channel Model for MT



Statistical MT Model

- Use the noisy channel model as we did many times before. We use Bayesian inference to find the best translation $\hat{T}$, given the sentence in the source language S and multiple candidate translations in the target language T:

$$\hat{T} = \underset{T}{\arg\max}\ P(T)P(S \mid T)$$

- Notice that the model for $P(T)$ is a model that gives us the best perplexity in the target language

Statistical MT Model

$$\hat{T} = \underset{T}{\text{arg max}}\; P(T)P(S \mid T)$$

- In order to compute $P(S \mid T)$ we need a parallel corpus with source and target language sentences. Furthermore, we need these sentences to be in **alignment** with each other.

- Once we have an alignment we can train the model $P(S \mid T)$

Alignment for Statistical MT

- However, before we can align words within sentences (the *word alignment* problem), we have to align the sentences

- Text alignments can be computed from *parallel texts* (also called *bitexts*)

Languages	Corpus
English, French	Canadian Hansards
English, French, German	Union Bank of Switzerland reports
English, Cantonese	Hong Kong Hansards

Text Alignment

- The task is to find alignment A given two parallel text S and T:

$$\underset{A}{\text{arg max}}\; P(A \mid S, T) = \underset{A}{\text{arg max}}\; P(A, S, T)$$

- The alignment can be represented as a mapping of 1:1, 1:2, 2:2, . . . where each such mapping is called a *bead*, B_k:

$$P(A, S, T) \approx \prod_{k=1}^{K} P(B_k)$$

- In practice, the beads are restricted to $\{$ 1:1, 1:0, 0:1, 2:1, 1:2, 2:2 $\}$

Text Alignment

- Text alignment between sentences $s_1, s_2, \ldots, s_i$ and $t_1, \ldots, t_j$ can be represented as an edit distance computation:

$$D(i,j) = min \begin{cases} D(i, j-1) & + & \text{cost}(0{:}1 \text{ align } \emptyset, t_j) \\ D(i-1, j) & + & \text{cost}(1{:}0 \text{ align } s_i, \emptyset) \\ D(i-1, j-1) & + & \text{cost}(1{:}1 \text{ align } s_i, t_j) \\ D(i-1, j-2) & + & \text{cost}(1{:}2 \text{ align } s_i, t_{j-1}, t_j) \\ D(i-2, j-1) & + & \text{cost}(2{:}1 \text{ align } s_{i-1}, s_i, t_j) \\ D(i-2, j-2) & + & \text{cost}(2{:}2 \text{ align } s_{i-1}, s_i, t_{j-1}, t_j) \end{cases}$$

Text Alignment

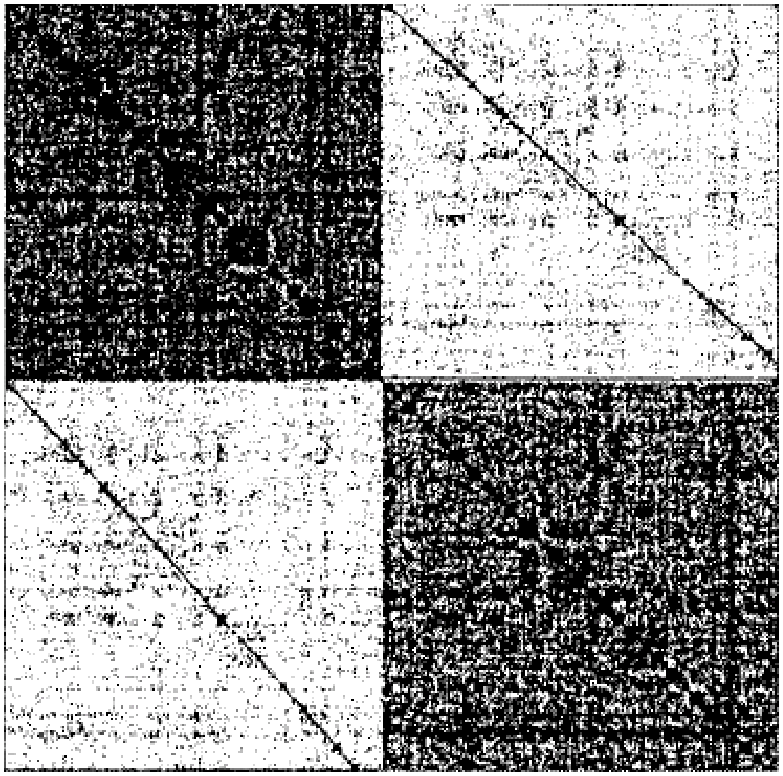
- For example, take sentences s_1, s_2, s_3, s_4 and t_1, t_2, t_3.

- Consider a 2:1 alignment combined with two 1:1 alignments:
  $L_1 =$ cost(align(s_1, s_2, t_1)) + cost(align(s_3, t_2)) + cost(align(s_4, t_3))

- Compare with:
  $L_2 =$ cost(align(s_1, t_1)) + cost(align(s_2, t_2)) + cost(align($s_3, \emptyset$)) + cost(align(s_4, t_3))

- The cost function is typically set to match the lengths of the sentences.

Dot Plot for the Canadian Hansards (Eng,Fr)



Dot Plot for the Canadian Hansards (Eng,Fr)

- Alternative method for aligning sentences and words simultaneously.
- Concatenate the source and target text and then plot a point for each matching word in position (i, j)
- Notice the large overlap when the source and target is plotted against itself (with the diagonal for self-similar words)
  More importantly, notice there is also a thin diagonal for similar words between the source and target corpus.
- Use a geometric methods to grow a diagonal line (almost a line) to match the words and sentences between source and target texts.

Statistical MT Model

$$\hat{T} = \underset{T}{\text{arg max}}\; P(T)P(S \mid T)$$

- Once we have aligned sentences we can compute the alignment between words in the source language and target language

- For example, lets consider translating a French sentence (f) of length m to an English sentence (e) of length l:

$$P(f \mid e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(f_j \mid e_{a_j})$$

- f_j is a word in f; a_j is the position in e aligned with f_j; e_{a_j} is the word in e aligned with f_j.

Statistical MT Model

- $P(w_f \mid w_e)$ is the *translation probability*

- The m sums $\sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l}$ sum over all possible alignments of French words to English words

- If $a_j = 0$ then the French word f_j is aligned to the empty string (i.e. it is deleted).

- This means that an English word can be aligned with multiple French words, but each French word is aligned with at most one English word.

Statistical MT Model

- For a particular alignment, we multiple the translation probabilities. For example, for the alignments *(Jean, John), (aime, loves), (Marie, Mary)*:

$$
\begin{aligned}
P(\textit{Jean aime Marie} \mid \textit{John loves Mary}) = \\
P(\textit{Jean} \mid \textit{John}) \times \\
P(\textit{aime} \mid \textit{loves}) \times \\
P(\textit{Marie} \mid \textit{Mary})
\end{aligned}
$$

Word Alignment for Statistical MT