Machine Translation

Adam Meyers
New York University
Summary

• Human Translation
• Goals of Modern Day Machine Translation
• History of Machine Translation
• Parallel Corpora and their Role in MT
• Aligning Sentences of Parallel Corpora
• Manual Transfer Approaches and Systran
• Statistical Machine Translation
• Adding Structure to SMT
• MT using Deep Learning
• Evaluation
Translation: Human vs. Machine

• Humans do a really good job, very slowly
  – A craft, learned and perfected over centuries
  – NOT directly based on innate human abilities
  – Must understand cultural context of source & target

• Computers are faster and do a bad job
  – Many methods require much computer time to “train”
  – Best translations are literal and awkward
  – Good for tasks where error is tolerated
Human Translation from Source to Target

• Preserve meaning
  – Find idiomatic expressions with similar connotations
  – Explain/remove background knowledge required by one community, but not the other
    • Adding/subtracting whole sentences or parts of sentences
  – Change order to reflect natural order of target language
  – Dynamic (intended) rather than literal (word-for-word) meaning

• Create well-written target language text
  – Obey stylistic conventions of target language
  – Match conventions, e.g., rhyme/meter in poetry
  – Fill in “missing” information required grammatically
    • Missing gender, pronouns, politeness conventions, etc.
Examples of Translations with Glosses

• Example 1
  – Aquí se habla español [Spanish]
  – Here one speaks Spanish [English gloss]
  – Spanish is spoken here [English translation]

• Example 2
  – Todos los libros me gustan [Spanish]
  – All the books me please [English gloss]
  – I like all the books [English translation]

• Example 3
  – Quiero unas tapas [Spanish]
  – (I) want some tapas [English gloss]
  – I want some samplings of small dishes [English translation 1]
  – I want some assorted appetizers [English translation 2]
  – I want some (Spanish) Dim Sum [English translation 3]
  – I want some tapas. [English translation 4]
Computer-Aided Translation

• Translation Memory Systems
  – Professional translators of commercial text may have access to sentence/translation pairs
  – Each translation can be based on a similar instance in translation memory
  – Requires aligned parallel sentences and a similarity measure

• Using MT as a first pass, depending on quality
  – High Quality MT output can be edited by a good writer in the target language.
  – Medium Quality MT output can be edited by a professional translator.
  – Lousy MT output would take longer for a translator to fix than it would take to translate from the original text.
Goals of Modern Day MT

• Gisting
  – Provide an imperfect, but informative translation
    • Identify articles worth translating professionally
    • Multi-lingual Information Extraction or Information Retrieval

• Translating Structured Input
  – Translating forms and tables
  – Translating Controlled/Limited Languages
    • Caterpillar Manuals, Microsoft Help Text

• Literal translation

• Mostly formal language and correspondence

• Literature (esp poetry) is basically impossible
An Abbreviated History of MT

• 1947 – Warren Weaver mentions the possibility of automatic translation in a memo to Norbert Weiner
• 1954 – The Georgetown Experiment automatically translates about 60 Russian sentences to English
• 1966 – ALPAC report admits that MT is really hard and that progress has been slow: funding is cut sharply
• 1968—1976  – Commercially successful manual MT systems  
  – Systran, Logos, Meteo
• 1980s – Statistical MT (IBM) & Example-based MT (Nagao)
• 1990 – 2000 – Combining /developing statistical and example-based
• 2000 – Present – adaption of SMT to deal with syntax  
  – Phrased-based Statistical Methods (Och, Koehn, …)
  – Tree to String (Yamada, Knight, …) & sometimes more structured input
• 2013 – Present – Deep Learning MT (Kalbrenner, Blunsom, Sutskever, Cho …)
Parallel, Near Parallel and Comparable Corpora

- **A bitext** is a pair of texts such that one is a translation of the other.
- **A tritext** is a triple of texts such that they are each a translation of the others.
- **Parallel corpora** include bitexts, tritexts, and any set of N texts, such that each is a translation of the others.
- **Parallel corpora tend towards literal translations.**
- **Comparable corpora** are sets of text about the same topic.
- **A Near-parallel corpus** is a text and one or more very dynamic translations of that text.
- **Examples:** Wikipedia pages of the same topic in multiple languages vary a lot with respect to these categories.
Uses of Parallel/Comparable Corpora

- Acquiring bilingual dictionaries
  - All types of parallel to comparable corpora
- Creating sentence-aligned bitexts
  - Parallel corpora
- Statistical MT and Example-based MT
  - Sentence-aligned bitexts
  - Bilingual dictionaries
- Answer Keys for Automatic MT evaluation
  - Sentence-aligned bitexts
- Translation Memory for Manual Translation
  - Sentence-aligned bitexts
Aligning Sentences of Bitexts

• Problem: Given a parallel bitext, determine which sentences of the SOURCE language align with which sentences of the TARGET

• Possible mappings between source/target sentences
  – 1 to 1     X translates as X'
  – N to 1     $X_1, X_2, \ldots X_N$ in combination translate as X'
  – 1 to N     X translates as $X_1', X_2', \ldots X_M'$ combined
  – N to N     $X_1, X_2, \ldots X_N \leftrightarrow X_1', X_2', \ldots X_M'$
  – 1 to 0     Source Sentence is not translated
  – 0 to 1     Target Sentence is added information

• Scrambling: Source/Target sentences may be ordered differently
Gale and Church 1993

• “A Program for Aligning Sentences in Bilingual Corpora,” Computational Linguistics, 19:1, pp. 75-102
• Uses character lengths of sentences and dynamic programming to assign probability scores to matching sentences
• First uses this method to align paragraphs, then aligns sentences within matching paragraphs
• Uses a training corpus of manually aligned sentences
• Incorporates edit distances for differences in alignments
  – deletions, scramblings, N to 1, etc.
Quick Definitions of Standard Statistical Concepts

• Variance = average of the squares of deviations from the mean
• Standard Deviation = square root of variance
• These are used to represent values that are distributed with a normal distribution.
• Distance Measures based on Standard Deviation are on the next slide
Gale and Church 2

- Probability that two units match calculated from manually aligned sentences
  - \( c = \) average number of characters in L1 per characters in L2
  - \( s^2 = \) variance between number of characters in corresponding \([1_1, 1_2]\) sentence pairs.
  - \( \delta = \frac{l_1 - (l_2 \times c)}{\sqrt{l_1 s^2}} \)
    - Approximately the number of standard deviations from the expected length
  - \( P(\text{match} | \delta) = constant \times P(\delta | \text{match}) \times P(\text{match}) \)

- Probability of different types of matches
  - \( P(1 \text{ to } 1) = .89 \)
  - \( P(1 \text{ to } 0 \text{ or } 0 \text{ to } 1) = .0099 \)
  - \( P(2 \text{ to } 1 \text{ or } 1 \text{ to } 2) = .089 \)
  - \( P(2 \text{ to } 2) = .011 \)

- Distance is calculated to penalize deletions, mergers and scramblings
- These probabilities are combined (details omitted)
- Alignments for English/French and English/German were about 96% correct
  - Hansards Corpus (English/French Canadian Parliament proceedings)
  - Economic Reports from Union Bank of Switzerland (English/German & English/French)
Meyers, Kosaka and Grishman 1998

• “A Multilingual Procedure for Dictionary-Based Sentence Alignment”, Proceedings of AMTA'98

• Sentence Similarity score based on morphological analysis and bilingual dictionary

• Analyzes sentence alignment as a variant of the stable marriage problem. Uses a solution based on the Gale-Shapey algorithm

• Assumes that alignments occur in 10 sentence windows
  – Large gaps can throw off alignment unless some other technique (paragraph alignment) is used in addition

• Handles 1 to 1, 1 to 0, 0 to 1, N to 1 and 1 to N alignments, not N to N
  – Assumes N < 4

• Results
  – Span/Eng 1-1: 97.8/93.5/95.6 Prec/Rec/F, 1-2/2-1: 20/100/33 Prec/Rec/F
  – Jap/Eng 1-1: 90.9/72.3/80.5 Prec/Rec/F, 1-2/2-1: 13.6/42.9/20.7 Prec/Rec/F
1 to 1 version

- Fill a 10 X 10 array with similarity scores between the first 10 source and first 10 target sentences
- Select the best alignment mapping from source to target using a version of the Gale-Shapey algorithm
  - An alignment is a set of source/target pairs
- From this alignment, keep the pairs that include source sentence 1 and target sentence 2 (this can be 0, 1 or 2 pairings).
- Remove the paired sentences from consideration and advance the window, so it is 10 X 10 again.
- Repeat until all sentences are aligned
Some Details

- N to 1 algorithm for some maximal N
  - Enlarge array for N to 1 & 1 to N matches, N = 1, 2 or 3
  - Only consecutive sentences are considered
  - Thus for 10 sentences, the array is $27 \times 27 = 729$ cells
    - 10 sentences + 9 sequences of 2 + 8 sequences of 3 = 27

- Constraint: matched sentences are at most 6 apart
  - Source sentences 1 and 10 compete for target sentence 5

- Similarity based on source (S) & target (T), words
  - $Dice = \frac{2 \times |Match(S, T)|}{|S| + |T|}$

- A source and target word match if
  - Any pair of morphological forms matches bilingual dictionary
  - Dictionary can be supplemented automatically by co-occurrence of unmatched words (requires second pass)
  - Morphological forms can be generated generously by removing any possible ending (erroneous forms won't match anything)
Gale Shapey Algorithm

- Stable Marriage Problem
  - N potential husbands, each with a ranking of N potential wives
  - N potential wives, each with a ranking of N potential husbands
  - A stable matching is a set of [husband, wife] pairings such that there is no two pairs \([h_1, w_1], [h_2, w_2]\)
    such that: \(h_1\) prefers \(w_2\) to \(w_1\) and \(w_2\) prefers \(h_1\) to \(h_2\)

- Gale Shapey algorithm chooses a set of 1-1 pairs, optimizing either for husband preferences or the wife preferences
  - Applications: applicants to law schools, dating services, and obviously, sentence alignment
  - Complexity = \(O(n^2)\)

- Gale Shapey Algorithm, optimizing for source sentences:
  - Repeat the following step until there are no more unmatched source sentences:
    - Match a source sentence \(S\) with its most preferred available target sentence \(T\)
    - \(T\) is available if:
      - \(T\) is currently unmatched or
      - \(T\) is matched, but prefers \(S\) to its current match \(S'\) (Then \(S'\) becomes unmatched)

- We run once optimized for source, once for target, then keep intersection and select conflicting cases based on score

- N-to-1 matches: modified definition of match conflicts and preferring 1 to 1
Direct Transfer Manual MT

• Separate Morphological from Lexical Components
  – *John likes ice cream sandwiches →
    John like+3rd_sing ice_cream sandwich+plural
• Translate words
  – Juan gustar+3rd_sing helado sándwich+plural
• Apply transfer rules, reorder and apply morphology
  – * letter indices: translations, number indices: per/num/gen agree
  – X_i like_i Y → X'_ gustar_j Y'_j
  – noun_1 noun_2 → noun_2' de noun_1'
  – plural noun → el/la_i + plural + noun_i + plural
  – Juan gustan los sándwiches de helado
Syntactic Transfer

- Transfer Rules Based on Parse Trees
  - Idiosyncratic to parsing/semantic system assumed
  - Semi-standardization of parsing to Penn Treebank is recent and not uncontroversial
- $like \rightarrow gustar$
- More precise than direct transfer
“Deeper” Level Transfer

- Can incorporate more generalizations
- Example: morphological agreement with the subject can occur after transfer
Systran

• History
  – Oldest Commercial MT system
    • company founded 1968
    • descendant of Georgetown University system from 1950s
  – Most successful manual transfer system
  – Some current Systran systems are hybrid manual/statistical systems
  – The Engine Behind Yahoo!'s BabbleFish translation service before it was replaced by Bing translate in 2012 (current version at: http://www.systranet.com/translate/)

• Languages
  – Many language pairs to/from English or French

• Multiple dictionaries for each language: idioms, morphology, compound nouns, …

• Many components are language independent, but have language specific modules

Hutchins & Somers 1992 Systran Diagram

Figure 10.1 Systran translation process
Systran: Source Language Pre-Processing

• Lookup in 3 bilingual dictionaries
  – Idioms and compound nouns – fixed multi-word dictionaries
    • with respect to, ice cream, tip top, so so, good for nothing, blow drier
  – Words – Main dictionary

• Morphological analysis
  – Nothing for English
  – For languages like Russian, stems and affixes looked up separately in Dictionaries
  – Some category info inferred from endings of OOV words
Systran 2\textsuperscript{nd} Stage: Source Language Analysis

- Homograph resolution (same spelling/different word)
  - Manual rules using adjacent POS – default: most frequent POS
- Phrase and Clause Identification:
  - A sort of shallow parsing, but looking for larger units than chunks
  - Clues: subordinate conjunctions (\textit{because}), punctuation, pronouns, …
- Identify Syntactic Relations:
  - Also like shallow parsing, but more like chunking/head identification
- Coordination and other “enumerations”
  - E.g., scope in: \textit{zinc and aluminum components}
- Identify Subjects, Predicates and semantic roles (deep cases)
  - Use special analytic dictionaries to deal with rare structures
Systran 3rd Stage: Transfer

- Translate conditional idioms (other idioms stage 1)
  - English passive *agreed* is translated as French *convenir*
  - Otherwise, *forms of agree* are translated as *être d'accord*
- Translate prepositions/postpositions
  - Previous stages needed – require syntactic/semantic info
- Lexical Routines: rules triggered by lex items
  - English *as* translates as many different French words depending on context
Systran 4rth Stage: Synthesis

• Word Translation (for words not handled by more specific rules)
• Morphological generation
  – Gender, number, tense, etc.
    • Previous rules allow agreement to be handled properly
• Syntactic Rearrangement
  – English Adj/Noun order → Spanish Noun/Adj order
• Result: Translated Sentence
How many MT Systems for N languages?

• N (N-1) transfer systems
  – English to Spanish, Spanish to English, English to German, German to English, Spanish to German, …
  – 10 languages → 180 systems (both directions)

• 2 X N Interlingua Systems
  – English to Interlingua, Interlingua to English, Spanish to Interlingua, Interlingua to Spanish, German to Interlingua, Interlingua to German, …
  – 10 languages → 20 systems
The Interlingual Approach

• Translate source language into Interlingua
  – Usually similar to automatic semantic analysis (from parse to semantics)

• Generate target language
  – Natural Language Generation

• What does an Interlingua Look Like?
  – A logical representation with standard primitives, e.g.,
    • Structure like a programming language       OR
    • Feature structure (or similar datastructure)   OR
    • Logical formulas
  – Some Pivot Language
    • English, Sanskrit, Esparanto, …

• Mostly toy systems – approach less successful than others
  – Except for resource-poor languages
Statistical Machine Translation (SMT)

- **Word Based Models**
  - based on translating individual words
  - allow for deletions, reorderings, etc.
  - Analogous to manual direct transfer systems

- **Phrase Based Models (2\textsuperscript{nd} most popular)**
  - based on translating blocks of words (may not be conventional phrases) and then words within those blocks
  - allows for deletions, reorderings, etc.

- **Models using structured text**
  - tree to string
  - synchronous grammars
  - tree to tree

- **Neural Networks (Newest and most popular)**
  - based on functions from source text to hidden layer(s) (encoding) and functions from hidden layer(s) to target text
Word Alignment

- A 1st step in training most statistical MT systems
- Map source words to target words, before various statistics are recorded (translation, distortion, etc.)
- Many systems implement other components, but use Giza++ or Berkeley word alignment programs
- Simple Example from Microsoft help text

<table>
<thead>
<tr>
<th></th>
<th>Excel</th>
<th>vuelve</th>
<th>a</th>
<th>calcular</th>
<th>valores</th>
<th>en</th>
<th>libro</th>
<th>de</th>
<th>trabajo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excel</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recalculates</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>workbook</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Word Alignment Discussion

- Use some of Birch and Koehn slides
- Slides 1 to 19: Introduces the IBM Model 1 and how to use with HMM (Model 1 assumes only 1 to 1 matches)
- **Pigeon Hole Principle (Dirchlet):** If items in \( A \) are matched to items in \( B \), such that \( A \) has \( N \) items \( B \) has \( N+1 \) items, at least 1 item of \( A \) matches 2 items in \( B \).
  - B & K interpret this to favor aligning unaligned items first.
- Go back to these slides for a detailed EM walk through
- We will go back and forth for a bit.
Simplified Example of EM model

• Given
  – 4 French words: *la, maison, bleu, and fleur*
  – 4 English words: *the, house, blue and flower*
  – We only allow 1 to 1 alignments

• Starting assumption
  – Each French word has a .25 chance of being translated as a given English word
Initial Alignment Probs for 3 E/F pairs

• *la maison* → *the house* \([la/the (.25), maison/the (.25), la/house (.25), maison/house (.25)]\)
  
  – *la/the X maison/house* = \(0.25^2 = 0.0625\)
  – *maison/the X la/house* = \(0.25^2 = 0.0625\)

• *la maison bleu* → *the blue house*
  
  – *la/the X maison/house X bleu/blue* = \(0.25^3 = 0.015625\)
  – *la/the X maison/blue X bleu/house* = \(0.25^3 = 0.015625\)
  – *la/house X maison/the X bleu/blue* = \(0.25^3 = 0.015625\)
  – *la/house X maison/blue X bleu/the* = \(0.25^3 = 0.015625\)
  – *la/blue X maison/house X bleu/the* = \(0.25^3 = 0.015625\)
  – *la/blue X maison/the X bleu/house* = \(0.25^3 = 0.015625\)

• *La fleur* → *the flower*
  
  – *la/the X fleur/flower* = \(0.25^2 = 0.0625\)
  – *fleur/the X la/flower* = \(0.25^2 = 0.0625\)
Maximum Likelihood Estimates (MLE)

- For each e/f pair and for each sentence, add up the probabilities of alignments that contain that pair and regularize to 1 (initially: all prob=.25)
- Sum these scores and divide by the number of instances of f.
- Translations from X to the
  - la/the: .5 of the first set of alignments, .33 of the second set and .5 of the 3rd
    - (.5 + .33 + .5) / 3 = .44
  - maisson/the: .5 of the 1st + .33 of the 2nd, 0 in the 3rd
    - (.5 + .33)/3 = .28
  - bleu/the: 0 in the 1st + .33 of the 2nd + 0 in the 3rd
    - .33/3 = .11
  - fleur/the: 0 in the 1st and 2nd, .5 in the 3rd
    - .5/3 = .17
- house: la/house=.42, maisson/house=.42, bleu/house=.17, fleur/house=0
- blue: la/blue=.33, maisson/blue=.33, bleu/blue=.33, fleur/blue=0
- flower: la/flower=.5 maisson/flower=0, blue/flower=0, fleur/flower= .5
Expectation: Rescore Alignments

- *la maison* → *the house*
  - *la/the* (.44), *maison/the* (.28), *la/house* (.42), *maison/house* (.28)
  - *la/the* X *maison/house* = .1848
  - *maison/the* X *la/house* = .1176

- *la maison bleu* → *the blue house* (all possible alignments)
  - *la/the* X *maison/house* X *bleu/blue* = .06098
  - *la/the* X *maison/blue* X *bleu/house* = .02468
  - *la/house* X *maison/the* X *bleu/blue* = .03881
  - *la/house* X *maison/blue* X *bleu/the* = .01525
  - *la/blue* X *maison/house* X *bleu/the* = .01525
  - *la/blue* X *maison/the* X *bleu/house* = .01571

- *La fleur* → *the flower*
  - *la/the* X *fleur/flower* = .22000
  - *fleur/the* X *la/flower* = .08500
Iteration of EM

• The Expectation and Maximization steps alternate until there is convergence (the probabilities do not change noticeably from iteration N to iteration N+1)
• Some of the details of scoring, e.g., presence of NULL, are omitted from example
• In the 1\textsuperscript{st} EM step, alignments are weighted equally
• For subsequent steps, the probabilities of previous alignments are used as weights, e.g., pairs in \textit{la maison} $\rightarrow$ \textit{the house} have weights of $0.1848/(0.1848+0.1176) = 0.61$ and $0.1176/(0.1848+0.1176) = 0.39$
IBM Models 1 to 5 for calculating translation probabilities for each sentence

- From Candide Project in 1980s and 1990s
- IBM model 1: Based on translation probability of each source word to each target word
- IBM model 2: Adds in distortion, probability of alignment given positions of source/target words and lengths of sentences
- IBM model 3: Adds fertility model, probability that each source word will correspond to N target words
- IBM model 4: Adds relative alignment model (modifies 2 to account for the fact that chunks move together)
- IBM model 5: Accounts for inaccuracies in 3 and 4 by only considering “vacant positions” when assigning probabilities
Phrase-Based Models

• Purportedly the highest performing systems
• In training, N to N words are aligned, not just single words
• These chunks of N words are often called “phrases”
  – But they need not be linguistic phrases
• Example alignment
  – *natuerlich hat john* [of course] *has fun with the game*
  – [spass am] *spiel*
• Phrase table acquired from alignments is used for translation
• Deletions and insertions become unnecessary
Phrase-Based Alignment

• Record all possible N to N mappings that:
  – are compatible with word alignment
  – N to N mappings are desirable (if frequent)
• It is therefore OK to have reliable mappings in which not all the words are aligned
• One popular technique:
  – Intersection of source-target & target-source word alignments
• Birch and Koehn slides 34 and 35
• It is OK to add unaligned blocks to adjacent aligned blocks
• The more probable phrase translations will be identified by an iterative process and highly ranked in the phrase table
• To limit computation, max phrase length (e.g., 6) often assumed
Decoding

• Find the most probable translation \( \hat{E} \), given:
  – Probability of translating \( F \) to a given \( E \) (a candidate \( \hat{E} \))
  – The probability of a particular \( E \) (the language model).

\[
\hat{E} = \arg \max_{E \in \text{English}} P(F | E) \times P(E)
\]

• \( P(F|E) \) is derived from probabilities trained
  – IBM Models: e.g., from previous slide
    • \((\text{la/the}) \cdot 0.44 \times (\text{maison/house}) \cdot 0.28 \times (\text{bleu/blue}) \cdot 0.11 = 0.012 \)
  – Phrase Model: probabilities from phrase table

• \( P(E) \) is based on language model
  – e.g., multiplying unigram, bigram, etc.
Translating sample sentence

• Input: *La maissan bleu*

• Translation probabilities (hypothetical):

<table>
<thead>
<tr>
<th>French</th>
<th>English</th>
<th>Unigram probabilities (count in WSJ ÷ 1 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>la</em></td>
<td>the</td>
<td>.035, <em>blue</em> = 1.3 X 10^{-4}, <em>house</em> = 6.7 X 10^{-4}, <em>flower</em> = 6 X 10^{-6}</td>
</tr>
<tr>
<td><em>maisson</em></td>
<td>blue</td>
<td>.13</td>
</tr>
<tr>
<td><em>bleu</em></td>
<td>house</td>
<td>.22</td>
</tr>
<tr>
<td><em>fleur</em></td>
<td>flower</td>
<td>.12</td>
</tr>
</tbody>
</table>

• The most probable translation would be:

  – *the house blue* = translation-prob X language prob = 4.37 X 10^{-10}
  
  • translation-prob = .70 X .41 X .50 = .1435
  
  • Lang-prob = .035 X 6.7 X 10^{-4} X 1.3 X 10^{-4} = 3.05 X 10^{-9}
More Details About Decoding

• The translation on the previous slide is the most probable, in part, because we only allow 1 to 1
  – more words → lower probabilities for all translations
  – N words implies N words in the translation

• Other models use additional components:
  – translation to/from NULL, distortion, fertility, ...

• Typically, generate K most likely translations
  – For different applications K can equal 1, 10, 1000, etc.
Tree-based Models

- So far the most successful Tree-based Models assume an isomorphism between source & target
- Sample Rule: $\text{NP} \rightarrow \text{Det}_1 \text{ NN}_2 \text{ JJ}_3 \mid \text{Det}_1 \text{ JJ}_3 \text{ NN}_2$
- Tree:
Problematic Tree: No VP rule
Solution: Change Grammar so VPs align
One Phrase Structure with 2 Strings

• String to Tree Machine Translation
  – Parser in one language is aligned with the tokens in the other language (biased to source or target)
  – More common method

• Synchronous parsing
  – A synchronous grammar is induced from the pair of source and target language texts
What about Tree to Tree alignment?

- Given $N$ source nodes and $N$ target nodes
  - alignment $i$ = set of pairs of source target nodes
  - $O(N!)$ 1 to 1 alignments (and more $N$ to 1, 1 to $N$, etc.)

- Reasonable constraints shrink the search space

- If synchronous grammars is too strict (1 to 1 partial mapping). What about weaker constraints?

- We did some experiments at NYU using logic dependency graphs (rooted DAGs, tree-like) using a dominance-preserving constraint
  - Motivation: There are cases (long distance dependencies) where linguistic analysis should work better than statistics (allowing displacements of $N$ tokens)
    - 2 Stage Manual Rule Parsers
    - Using GLARF as 2\textsuperscript{nd} stage
Dominance Preserving Constraint

- **Given** alignment $A$ including source nodes $S_1$ and $S_2$ and target nodes $T_1$ and $T_2$
- **If** Dominates($S_1$, $S_2$), **then** Dominates($T_1$, $T_2$)
Dominance-Preserving Alignment Algorithm

- Assume that Source and Target Roots are aligned.
- Compute the score of the source/target pair using the following recursive routine:
  - \( \text{Score}(X,Y) = \text{lexical score}(X,Y) + \text{highest scoring pairing of the children of } X \text{ and the children of } Y. \)
    - Lexical scores require a bilingual dictionary, which can be supplemented by automatic procedures to acquire missing (previously unaligned pairs).
- Also allow \( X \) to be aligned with one of the children of \( Y \) or \( Y \) to be aligned with one of the children of \( X \).
  - Without this step, the algorithm would be restricted to a least common ancestor preserving alignments, a subset of dominance-preserving alignments.
Tree to Tree Alignment

Excel vuelve a calcular valores en libro de trabajo.

Excel recalculates values in workbook.

Source Tree

$D = \text{volver}$

$A = \text{Excel}$

$B = \text{calcular}$

$F = \text{trabajo}$

Target Tree

$D' = \text{recalculate}$

$A' = \text{Excel}$

$B' = \text{values}$

$C' = \text{workbook}$

$\text{Excel recalculates values in workbook}$
Transfer Rules Derived From Alignment

1) \(A = \text{Excel}\) \(\rightarrow \) \(B' = \text{Excel}\)

2) \(F = \text{valores}\) \(\rightarrow \) \(B' = \text{values}\)

3) \(C = \text{libro}\)
   \(\rightarrow \) \(C' = \text{workbook}\)

4) \(D = \text{trabajo}\)
   \(\rightarrow \) \(D' = \text{recalculate}\)

\(\text{Subj} \rightarrow a\)

\(\text{Obj} \rightarrow \text{in}\)

\(\text{en}\)
A simple reordering based on Logic1 node alignment

I know the rules of tennis ↔

English in Chinese order: I know the (of) tennis rules
NYU Systems Using Dependency Graph Alignment

• Why: There are some cases (long distance dependencies) where linguistically motivated analysis should help MT

• 1996-2000
  – Toy systems for Spanish/English and Japanese/English
  – Using 2 stage parsers with manual rules

• 2010
  – Use GLARF on output of state of the art treebank parsers
  – Reordering English sentences to be like Chinese
  – Then run standard word alignment program (Giza++)
  – Achieved 1.5% improvement in Word Alignment
    • Most of the benefit from reordering large noun modifiers
  – Incremental step in larger goal:
    • use reordered English with state-of-the-art MT systems
Dominance-Preserving Constraint is too strong

- Weaker than synchronous grammar
- There are real cases for violations 1 and 2
- Violation 1 does not handle unclear modifier attachment
  - *Mary sent out a letter [to John]*
    - [sent out [a letter to John]]
    - [sent out [a letter] [to John]]
- Violation 2 ignores so-called head-switching phenomena
  - *Er tanzt gerne* [German]
  - *He dances with-pleasure* [English gloss]
  - *He likes to dance.* [English translation]
- Both violations are often found in parsing errors
- Common violation 2 instances for Chinese/English
  - Quantifier/transparent noun, e.g., тяжелые → *series of*
MT using Deep Learning

- Relatively new and very popular
- NYU's Prof. Kyunghyun Cho is one of the leading researchers in this area:
- Brief introduction in the next few slides
Source to Hidden to Target

• Lines represent functions from source to hidden layer, and from hidden layer and target
• Hidden Vector contains parameters for functions

Source: Mary read the book
Hidden: ●●●●●●●●●●●●●●●●
Target: María leyó el libro

• The parameters are initialized randomly and modified incrementally by training the system on parallel text
Hidden Layer is Like an Inter-Lingua

- Hidden layer of fixed number of nodes assumed between source and target words.
- Lots of connections are assumed between source and hidden layer and between hidden layer and target.
- Training the system results in:
  - Encoder translating source sentence to hidden layer
  - Decoder translating hidden layer to target sentence
- Some systems attempt to use the same hidden layer for multiple language pairs (in theory, like an inter-lingua)
  - This is interesting, but speculative
Deep Learning MT (Last Slide)

• Decoder translates incrementally using N source words (e.g., N=4) to predict the next target words in the translation
• Results comparable to phrase based MT.
• Advantages cited include:
  – Not necessary to manually design feature sets
  – Somewhat better quality
Human Evaluation of MT

• Human Evaluation: Effective & Expensive

• Method 1: Rate translations on several dimensions:
  – fluency – how intelligible is output
    • Includes clarity and naturalness
  – Fidelity – does translation contain all and only information from source
    • Includes adequacy, informativeness

• Method 2: How much editing is required to render the machine output into a good translation?
  – Track this in dollars, time or numbers of key strokes
Automatic Evaluation

• Automatic Methods: inexpensive, predominant, imperfect
  – At minimum, an evaluation metric shows improvement:
    • If a system improves, the score improves
    • If a system degrades, the score degrades
  – Output is rated on its “closeness” to the human translations
• Bleu: statistical definition of “closeness to human translation”
  – Many benchmarks are Bleu scores for particular test sets
  – Multiple human translations are provided for test set
  – Precision of n-grams in system output found in reference translations
    • N-gram is correct if in any of the references
  – Penalizes shorter output
  – Criticized for favoring statistical systems over manual (Systran) despite human evaluations to the contrary
MEANT: Automatic Evaluation Based on Semantic Role Labeling

- Steps
  - Step 1: Run SRL system for Answer Key and System Output and represent each as a graph
  - Step 2: Align graphs
  - Step 3: Measure similarity between graphs (based on F-score)
- These authors show a higher correlation with manual evaluation using this metric than other automatic metrics
- Previous papers by Wu's group describe evaluation incorporating manual input
- Subsequent papers describe improvements to the system
Summary

• The best statistical systems currently use the phrase-based approach
  – These are arguably the best systems overall
• Systran is a (proprietary) competitive system that is probably uses a combination of approaches including many manual rules and dictionaries
  – May be competitive with statistical approaches (unclear because the most commonly used score is arguably biased)
• There is research in alternatives using more linguistically motivated analysis
  – Sometimes in conjunction with statistical systems
Additional Information

• There has been some research on translating poetry, e.g., by Google:

• Interlingua and Pivot systems are sometimes used for resource-poor languages (other methods not possible for practical reasons)
Readings

• Required: J & M Chapter 25
• Various Optional Readings mentioned throughout slides