The Menu Bar

- Administrivia:
  - final projects -
- Agenda:
  - Combining statistics with language knowledge in MT
  - MT - the statistical approach (the “low road”)
    - Evaluation
    - Where does it go wrong? Beyond the “talking dog”
  - MT – Middleuropa ground
    - Transfer Approach: using syntax to help
    - How to combine w/ statistical information
  - Can we attain the Holy Grail?

How well does stat MT do?

- What happens if the sentence is already seen (part of training pair)?
- Then the system works just as hard
- Remembrance of translations past...?
- We get “only” 60% accuracy (but better than Systran...)
- Let’s see how to improve this by adding knowledge re syntax
- Probably even better to add knowledge re semantics... as we shall see

The game plan to get better

<table>
<thead>
<tr>
<th>MT Fluency &amp; Adequacy Bake-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct/agram</td>
</tr>
<tr>
<td>Systran (transfer) 54/74%</td>
</tr>
<tr>
<td>IBM Model 5 (Statistical) 58/67%</td>
</tr>
<tr>
<td>Stat+ Syntax transfer</td>
</tr>
<tr>
<td>Human 84/86%</td>
</tr>
<tr>
<td>Stat MT</td>
</tr>
</tbody>
</table>

MT Fluency & Adequacy Bake-off

0 100

AD Equacy
Problemos

- F in: L’atmosphère de la Terre rend un peu myopes même les meilleurs de leur télescopes
- E out: The atmosphere of the Earth returns a little myopes same the best ones of their telescopes
- (Systran): The atmosphere of the Earth makes a little short-sighted same the best of their télescopes
- (Better) The earth’s atmosphere makes even the best of their telescopes a little ‘near sighted’
- Why?

Let’s take a look at some results...

Should

<table>
<thead>
<tr>
<th></th>
<th>should</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>t(f</td>
</tr>
<tr>
<td></td>
<td>0.330</td>
</tr>
<tr>
<td>devrait</td>
<td>0.123</td>
</tr>
<tr>
<td>devraient</td>
<td>0.109</td>
</tr>
<tr>
<td>devrons</td>
<td>0.073</td>
</tr>
<tr>
<td>faut</td>
<td>0.058</td>
</tr>
<tr>
<td>doit</td>
<td>0.058</td>
</tr>
<tr>
<td>aurait</td>
<td>0.041</td>
</tr>
<tr>
<td>doivent</td>
<td>0.024</td>
</tr>
<tr>
<td>devons</td>
<td>0.017</td>
</tr>
<tr>
<td>devrais</td>
<td>0.013</td>
</tr>
</tbody>
</table>

What about...

- In French, what is worth saying is worth saying in many different ways
- He is nodding:
  - Il fait signe qui oui
  - Il fait un signe de la tête
  - Il fait un signe de tête affirmatif
  - Il hoche la tête affirmativement
Nodding hill...

<table>
<thead>
<tr>
<th>Word</th>
<th>Phi 1</th>
<th>Phi 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>signe</td>
<td>0.164</td>
<td>0.342</td>
</tr>
<tr>
<td>la</td>
<td>0.123</td>
<td>0.203</td>
</tr>
<tr>
<td>oui</td>
<td>0.086</td>
<td>0.163</td>
</tr>
<tr>
<td>fait</td>
<td>0.073</td>
<td>0.023</td>
</tr>
<tr>
<td>que</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>hoche</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>faire</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>me</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>approve</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>qui</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>un</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>faites</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

Best of 1.9 x 10^6 alignments!

Best of 8.4 x 10^9 alignments!

• Always works hard – even if the input sentence is one of the training examples
• Ignores morphology – so what happens?
• Ignores phrasal chunks – can we include this? (Do we?)...
• Can we include syntax and semantics?
• (why not?)

Other languages...

• Aligning corpus – a cottage industry
  • Instruction Manuals
  • Hong Kong Legislation - Hansards
  • Macao Legislation
  • Canadian Parliament Hansards
  • United Nations Reports
  • Official Journal of the European Communities

How can we do better?

• Systran: transfer approach
• Q: What's that?
• A: transfer rules for little bits of syntax
• Then: combine these rules with the statistical method
  • Even doing this a little will improve us to about 65%
  • Gung ho – we can get to 70%
  • Can we get to the magic number?

The golden (Bermuda?) triangle

Source (eg, Spanish)  Target (eg, English)
The Bermuda triangle revisited

Transfer station

- Transfer: Contrasts are fundamental to translation. Statements in one theory (source language) are mapped into statements in another theory (target language).
- Interlingua: Meanings are language independent and can be encoded. They are extracted from Source sentences and rendered as Target sentences.

Transfer approach

- Analysis using a morphological analyser, parser and a grammar.
- Depending on approach, grammar must build syntactic and/or semantic representation.
- Transfer: mapping between S and T.
- Generation using grammar and morphological synthesizer (from analysis?)

Transfer system: 2 languages
Transfer - multiple languages

Syntactic Transfer

Syntactic transfer

<table>
<thead>
<tr>
<th>SL Tree</th>
<th>Source-side transformation</th>
<th>TL Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det the NP N1 Adj N2... Adj Nm</td>
<td>Det the NP N1 Adj N2... Adj Nm</td>
<td>Det the NP N1 Adj N2... Adj Nm</td>
</tr>
<tr>
<td>Det the NP N1 Adj N2... Adj Nm</td>
<td>Det the NP N1 Adj N2... Adj Nm</td>
<td>Det the NP N1 Adj N2... Adj Nm</td>
</tr>
</tbody>
</table>

5 transfer rules: 3 syntax, 2 lexical

Syntactic transfer

- Maps trees to trees
- No need for ‘generation’ except morphology
- Method: top-down recursive, non-deterministic match of transfer rules (where $t_v$ is a variable) against tree in source language
- Output is tree in target language (w/o word morphology)
Simple syntactic transfer example

- Rules (English-Spanish) - 3 in previous example
- 1 for NP NP; 1 for N1 N1'; one for Det Det
- Lexical correspondences
- Suppose input is as in preceding example - trace through matching

Handling other differences

- E: You like her
- S: Ella te gusta
- Lit: She you-accusative pleases
  (Grammatical object in English is subject in Spanish, and v.v.)

Syntactic transfer

Tree mapping rule for this

\[ S \rightarrow \text{tv(subj)} \rightarrow \text{vp} \rightarrow S \rightarrow \text{tv(obj)} \rightarrow \text{vp} \]
Is this systematic?

- Yes, and taxonomic too...
- Roughly 8-9 such ‘classes’ of divergence:
  1. Thematic
  2. Head switching
  3. Structural
  4. Lexical Gap
  5. Lexicalization
  6. Categorial
  7. Collocational
  8. Multi-lexeme/idiomatic
  9. Generalization/morphological

Other divergences- systematic

- E: The baby just ate
- S: El bebé acaba de comer
- Lit: The baby finish of to-eat
  Head-switching

- E: Luisa entered the house
- S: Luisa entró a la casa
- Lit: Luisa entered to the house
  Structural

Divergences diverging

- E: Camilio got up early
- S: Camilio madruó
  Lexical gap

- E: Susan swam across the channel
- S: Susan cruzó el canal nadando
  (Systran: Susan nadó a través del canal)
  Lit: Susan crossed the channel swimming
  (manner & motion combined in verb E, path in across in S, verb cruzó has motion & path, motion in gerund nadando)
  Lexicalization

Divergences, III

- E: A little bread
- S: Un poco de pan
- Lit: A bit of bread
  Categorial - difft syntactic categories

- E: John made a decision
- S: John tomó/*/hizo una decisión
- Lit: John took/*/made a decision
  Collocational - usually make goes to hacer but here a ‘support’ verb for decision
We can accommodate these...

Issues

- Q: How many rules?
- A: usually many 000s for each one-way pair
- Q: Nondeterminism - which rule to apply?
- Q: How hard is it to build a rule system?
- A: Can we learn these automatically?

Transfer picture again

Statistical MT is transfer approach
Statistical MT is a transfer approach!

- Except... Analysis & synthesis vestigial
- Transfer done statistically
- Can we do better by applying some simple analysis & synthesis?
- A: Yes, from 50s to 60+ %
- A: Yes, we can do even better if we do incorporate syntax systematically: Knight et al 2001
- We will see that there's a method behind this madness, and all the alignment methods are in effect 'learning' the transfer rules

Adding syntax to Stat MT... simply

Statistical Machine Translation Model

Source Language Model

Translation Model

Decoder

$\hat{S} = \text{argmax } P(S|T) = \text{argmax } P(S,T)$


Simple analysis and synthesis - IBM Model X

- Find word strings
- Annotate words via simple grammatical functions
- Very very very simple syntactic analysis
- Inflectional morphology
- Statistically derived word senses
Crummy but amazing improvement to stat model

- Simplest English & French syntactic regularization
- For English:
  - Undo question inversion
  - Move adverbs out of multiword verbs
  - Eg: Has the grocery store any eggs →
    The grocery store has any eggs Qinv →
    Iraq will probably not be completely balkanized →
    Iraq will be balkanized probably_m1 not_m2 completely_m3

And for French...

- Undo question inversion
- Combine ne...pas, rien into single words
- Move pronouns that function as direct, indirect, objects to position following verb & mark grammatical function
- Move adjs s.t. precede nouns they modify & adverbs to position following verbs they modify

French examples

- Où habite-il → Où il habite Qinv
- Il n'y en a plus → Il y en a ne_plus
- Je vous le donnerai → Je donnerai le_Pro vous_iPro (“I gave it to you”)

How well does this work?

- Pretty darn well
- Improves performance about 10%
  - 50-odd % to 60+

Now – let’s see if we can reach the next step by doing this a bit more thoroughly:
- Add linguistic features to a statistical translation model by using parse trees
Add different ‘channels’ of noise

• Instead of one noisy channel, break it out into syntactic possibilities
• Reorder – model S V O vs. S O V order (Chinese, English vs. Japanese, Turkish)
• Insertion – model case particles
• Translating – as before
Channeling - input

1. Channel Input

Reordering (r-table)

<table>
<thead>
<tr>
<th>input order</th>
<th>reordered</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP VB1 VB2</td>
<td>PRP VB1 VB2</td>
<td>0.074</td>
</tr>
<tr>
<td>VB1 PRP VB2</td>
<td>VB1 PRP VB2</td>
<td>0.073</td>
</tr>
<tr>
<td>VB VB1 PRP VB2</td>
<td>VB VB1 PRP VB2</td>
<td>0.081</td>
</tr>
<tr>
<td>VB VB VB PRP VB1</td>
<td>VB VB VB PRP VB1</td>
<td>0.083</td>
</tr>
<tr>
<td>VB1 TO VB</td>
<td>TO VB VB1</td>
<td>0.749</td>
</tr>
<tr>
<td>TO VB TO VB1</td>
<td>TO VB TO VB1</td>
<td>0.749</td>
</tr>
<tr>
<td>TO NN TO</td>
<td>TO NN TO</td>
<td>0.893</td>
</tr>
</tbody>
</table>

Channeling

- child nodes on each internal node are reordered, via R-table
- Eg: PRP-VB1-VB2 to PRP-VB2-VB1 has pr 0.723, so we pick that one
- Also reorder VB-TO → TO-VB; TO-NN → NN-TO
- Prob of the 2nd tree is therefore 0.723 x 0.749 x 0.893 = 0.484

Reordered

1. Channel Input

2. Reordered
Insertion

- Captures regularity of inserting case markers *ga*, *wa*, etc.
- No conditioning – case marker just as likely anywhere

Insertion - n-table

- Left, right, or nowhere (diff from IBM)
- 2-way table index, by (node, parent)
- EG, PRP node has parent VB

<table>
<thead>
<tr>
<th>parent node</th>
<th>TOP VB</th>
<th>VB VB</th>
<th>VB TO</th>
<th>TO TO</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB VB</td>
<td>0.735</td>
<td>0.687</td>
<td>0.344</td>
<td>0.709</td>
<td>0.900</td>
</tr>
<tr>
<td>VB PRP</td>
<td>0.014</td>
<td>0.061</td>
<td>0.034</td>
<td>0.030</td>
<td>0.003</td>
</tr>
<tr>
<td>VB Right</td>
<td>0.260</td>
<td>0.252</td>
<td>0.652</td>
<td>0.261</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Insertion – which words to insert table

<table>
<thead>
<tr>
<th>w</th>
<th>P{ins-\text{w}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ha</td>
<td>0.219</td>
</tr>
<tr>
<td>to</td>
<td>0.131</td>
</tr>
<tr>
<td>wo</td>
<td>0.099</td>
</tr>
<tr>
<td>no</td>
<td>0.094</td>
</tr>
<tr>
<td>ni</td>
<td>0.080</td>
</tr>
<tr>
<td>to</td>
<td>0.078</td>
</tr>
<tr>
<td>ga</td>
<td>0.062</td>
</tr>
<tr>
<td>desu</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

Insertion

Insert four words (ha, no, ga and desu) to create the third tree. The top VB node, two TO nodes, and the NN node inserted nothing. So, probability of obtaining the third tree given the second tree is: 4 particles x no inserts = (0.652 x 0.219)(0.252 x 0.094)(0.252 x 0.062)(0.252 x 0.007)x 0.735 x 0.709 x 0.900 x 0.800 = 3.498e-9

Translate – final channeling

• Apply the translate operation to each leaf
• Dependent only on the word itself and that no context
• Translations for the tree shown...

Translation, t-table
Total probability for this (e,j) pair

- Product of these 3 ops
- But many other combinations of these 3 ops yield same Japanese sentence, so must sum these pr’s...
- Actually done with 2121 E/J sentence pairs
- Uses efficient implementation of EM (50 mins per iteration, 20 iterations)

Statistical Machine Translation Model

\[ P(S) \times P(T|S) = P(S,T) \]

\[ \hat{S} = \text{argmax} \ P(S|T) = \text{argmax} \ P(S,T) \]

Syntax-based MT

Random variables N, R, T each representing one of the channel operations for each (E)nglish node $\varepsilon$

- Insertion $N(\nu)$
- Reorder $R(\rho)$
- Translation $T(\tau)$

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>Insertion</th>
<th>Reorder</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(\nu)$</td>
<td>$R(\rho)$</td>
<td>$T(\tau)$</td>
<td></td>
</tr>
</tbody>
</table>

**Other possible transformed trees:**

$S_{q1} = <n_1, r_1, t_1>_{NP} \quad VP_{q2} = <n_2, r_2, t_2>_{NP} \quad Det_{q4} = <n_4, r_4, t_4>_{NP} \quad run_{q3} = <n_3, r_3, t_3>_{NP}$

Parameter estimation via EM

### Parameter estimation via EM

1. Initialize all probability tables: $n(\nu, N)$, $r(\rho, R)$, and $t(\tau, T)$
2. Reset all counters $c(\nu, N)$, $c(\rho, R)$, and $c(\tau, T)$
3. For each pair $\langle v, f \rangle$ in the training corpus:
   - For all $\theta$, such that $f = \text{String}(\theta(v))$,
     - Let $cnt = P(\theta(v)|\varepsilon) / \sum_{\theta \in \text{Strings}(\varepsilon)} P(\theta(v)|\varepsilon)$
   - $c(\nu, N(v)) += cnt$
   - $c(\rho, R(v)) += cnt$
   - $c(\tau, T(v)) += cnt$
4. For each $(\nu, N)$, $(\rho, R)$, and $(\tau, T)$:
   - $n(\nu, N) = c(\nu, N) / \sum_{\nu \in \text{Strings}(\varepsilon)} c(\nu, N)$
   - $r(\rho, R) = c(\rho, R) / \sum_{\rho \in \text{Strings}(\varepsilon)} c(\rho, R)$
   - $t(\tau, T) = c(\tau, T) / \sum_{\tau \in \text{Strings}(\varepsilon)} c(\tau, T)$
5. Repeat steps 2-4 for several iterations (until little change) [20 steps]
Parameter estimation via EM

$O(\sum_{\nu}\rho_{\nu}\tau_{\nu})$
for all possible combinations of parameters $(\nu, \rho, \tau)$

$O(\sum_{\nu}\rho_{\nu}|\nu|\rho_{\nu}|\rho)$

Results vs. IBM Model 5

Results for 50 sentence pairs

Perfect = all alignments OK for 3 judges
Scoring: 1.0 = OK; 0.5 = not sure; 0 = wrong

<table>
<thead>
<tr>
<th></th>
<th>Alignment ave. score</th>
<th>Perfect sents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Model</td>
<td>0.582</td>
<td>10</td>
</tr>
<tr>
<td>IBM Model</td>
<td>0.431</td>
<td>0</td>
</tr>
</tbody>
</table>

For E-F, goes up to 70%!
Can we get to the next step up – “Gold Standard” of 80%??

Problemos

- F in: L’atmosphère de la Terre rend un peu myopes mèmes les meilleurs de leur télescopes
- E out: The atmosphere of the Earth returns a little myopes same the best ones of their telescopes
- (Systran): The atmosphere of the Earth makes a little short-sighted same the best of their télescopes
- (Better) The earth’s atmosphere makes even the best of their telescopes a little ‘near sighted’
- Why?
Pourquois?

• French verb *rend* can be ‘return’ or ‘make’
• French word *même* can be ‘same’ or ‘even’ – translation systems get it dead wrong

Problem of context

• General vs. specialized use of word
• “Dear Bill,” to German:
  • Liebe Rechnung –
  • “beloved invoice”
  • (Systran) Liebe Bill
  • Solution: consult word senses?

Anaphora and beyond...

• Die Europäische Gemeinschaft und ihre Mitglieder
  • The European Community and its members
  • Die Europäische Gemeinschaft und seine Mitglieder
• The monkey ate the banana because it was hungry
  • Der Affe ass die Banane weil er Hunger hat
  • Der Affe ass die Banane, weil sie hungrig war
• The monkey ate the banana because it was ripe
  • Der Affe ass die Banane weil sie reif war
• The monkey ate the banana because it was lunch-time
  • Der Affe ass die Banane weil es Mittagessen war
• Sentence-orientation of all systems makes most anaphora problematic (unresolvable?); possibly a discourse-oriented ‘language model’ is the only chance