Syntax-Based Statistical Machine Translation

(Or: “Can a Machine Translate Without Knowing What a Verb Is?”)

Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

MT Summit, Copenhagen, September, 2007
Topics

• **Quick review of statistical MT essentials**
  – bilingual text
  – phrase substitution models
  – language models
  – decoding

• **Syntax-based statistical MT**
  – syntax-based translation models
  – learning syntactic transformation rules from data
  – decoding
  – tree automata
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.
Statistical Machine Translation

Hmm, every time he sees “banco”, he either types “bank” or “bench” … but if he sees “banco de…”, he always types “bank”, never “bench”...

Human-translated documents
## Spanish/English corpus

Translate: Clients do not sell pharmaceuticals in Europe.

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
</table>
| 1a. Garcia and associates .  
1b. Garcia y asociados . | 7a. the clients and the associates are enemies .  
7b. los clientes y los asociados son enemigos . |
| 2a. Carlos Garcia has three associates .  
2b. Carlos Garcia tiene tres asociados . | 8a. the company has three groups .  
8b. la empresa tiene tres grupos . |
| 3a. his associates are not strong .  
3b. sus asociados no son fuertes . | 9a. its groups are in Europe .  
9b. sus grupos estan en Europa . |
| 4a. Garcia has a company also .  
4b. Garcia tambien tiene una empresa . | 10a. the modern groups sell strong pharmaceuticals .  
10b. los grupos modernos venden medicinas fuertes . |
| 5a. its clients are angry .  
5b. sus clientes estan enfadados . | 11a. the groups do not sell zenzanine .  
11b. los grupos no venden zanzanina . |
| 6a. the associates are also angry .  
6b. los asociados tambien estan enfadados . | 12a. the small groups are not modern .  
12b. los grupos pequenos no son modernos . |
<table>
<thead>
<tr>
<th>Arcturan</th>
<th>Centauri/Arcturan [Knight 97]</th>
<th>Your assignment, translate this to Arcturan:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok farok ororok lalok sprok izok enemok .</td>
<td>farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
<td></td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lalok brok anok plok nok .</td>
<td></td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat rrat dat .</td>
<td>8b. iat lat pippat rrat nnat .</td>
<td></td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
<td></td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat nnat quat oloat at-yurp .</td>
<td></td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
<td></td>
</tr>
<tr>
<td>4b. at-voon krat pippat sat lat .</td>
<td>10b. wat nnat gat mat bat hilat .</td>
<td></td>
</tr>
<tr>
<td>5a. wiwok farok izok stok .</td>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
<td></td>
</tr>
<tr>
<td>5b. totat jjat quat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
<td></td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok rarok nok izok hihok mok .</td>
<td></td>
</tr>
<tr>
<td>6b. wat dat krat quat cat .</td>
<td>12b. wat nnat forat arrat vat gat .</td>
<td></td>
</tr>
</tbody>
</table>
Ready-to-Use Online Bilingual Data

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).
Bilingual Text (200m words)

English strings

Word alignments

Chinese strings

Word-Aligned bilingual text

Phrase Pair Extraction [Och & Ney, 2004]

Vast Database of Phrase Pairs
<table>
<thead>
<tr>
<th>7 people</th>
<th>including</th>
<th>by some</th>
<th>and</th>
<th>the russian</th>
<th>the astronomers</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 people included</td>
<td>by france</td>
<td>and the russian</td>
<td>international astronomical</td>
<td>of rapporteur</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 out</td>
<td>including the</td>
<td>from</td>
<td>the french</td>
<td>and the russian</td>
<td>the fifth</td>
<td></td>
</tr>
<tr>
<td>these</td>
<td>7 among</td>
<td>including from</td>
<td>the french</td>
<td>and</td>
<td>of the russian</td>
<td>of space</td>
</tr>
<tr>
<td>these</td>
<td>7 persons</td>
<td>including from</td>
<td>the</td>
<td>of france</td>
<td>and to</td>
<td>russian</td>
</tr>
<tr>
<td>7 include</td>
<td>from the</td>
<td>of france and</td>
<td>russian</td>
<td>astronauts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>numbers include</td>
<td>from france</td>
<td>and russian</td>
<td>of astronauts who</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>populations include</td>
<td>those from france</td>
<td>and russian</td>
<td>astronauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deportees included</td>
<td>come from</td>
<td>france</td>
<td>and russia</td>
<td>in</td>
<td>astronomical personnel</td>
<td></td>
</tr>
<tr>
<td>philltrum</td>
<td>including those from</td>
<td>france and</td>
<td>russian</td>
<td>a space</td>
<td></td>
<td></td>
</tr>
<tr>
<td>including representatives from</td>
<td>france and the</td>
<td>russian</td>
<td>astronaut</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include</td>
<td>came from</td>
<td>france and russia</td>
<td>by cosmonauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include representatives from</td>
<td>french and</td>
<td>russia</td>
<td>cosmonauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include</td>
<td>came from france</td>
<td>and</td>
<td>russia’s</td>
<td>cosmonauts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes</td>
<td>coming from</td>
<td>french and</td>
<td>russia’s</td>
<td>a</td>
<td>astronaut navigation</td>
<td>member</td>
</tr>
<tr>
<td></td>
<td>french</td>
<td>and russia</td>
<td>astronauts</td>
<td>special rapporteur</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>and russia’s</td>
<td>, and</td>
<td>russia</td>
<td>rapporteur</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>, and russia</td>
<td>, and russia</td>
<td>rapporteur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>, and russia</td>
<td>or</td>
<td>russia’s</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.
这7人中包括来自法国和俄罗斯的宇航员。

Table 1: #11# the seven-member crew includes astronauts from France and Russia.

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
这7人中包括来自法国和俄罗斯的宇航员。

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
Phrase-Based Translation

Table 1: #11# the seven-member crew includes astronauts from France and Russia.

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
Components

• Training algorithms
  – Word alignment, phrase pair extraction…
    • $P(\text{chinese} \mid \text{english}) = \text{product of conditional phrase pair probabilities}$
  – English n-gram models…
    • $P(\text{english}) = \text{product of trigram probabilities}$
    • $P(w3 \mid w1 \ w2)$

• Decoding algorithm
  – $\text{argmax } e \ P(\text{english} \mid \text{chinese}) = \text{argmax } e \ P(\text{english}) \ast P(\text{chinese} \mid \text{english})$
Features and Tuning

- English n-gram language model
- Phrase pairs
  - Corpus probability of phrase pair
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length

We compute a total score for each possible translation -- a linear weighted combination of these six values. This generalizes the formula from the previous slide, if we switch to log probs.
Weight settings really affect translation quality!

plot by Emil Ettelaie

W_{TM} fixed at 1.0
(A View from the Back)

$W_{TM}$ fixed at 1.0
Hill climbing with Minimum Error-Rate Training (MERT) aka MaxBleu Training (Och, 2003)

Translation accuracy

plot by Emil Ettelaie

$W_{\text{TM}}$ fixed at 1.0
These Ideas Work!

Translation Quality
(BLEU)

Phrase-based MT Progress

NIST Common Evaluations
(Arabic/English)
Some Lessons

• The simpler, the better
• It takes a long time just to get the bugs out!
• Every change has to be carefully checked
• Good ideas often don’t help
• Have to try lots of things
• It’s highly experimental
Statistical MT Research is Highly Experimental

Translation Accuracy (BLEU)

Actual progress at undisclosed laboratory!

Chinese/English NIST 2002 Test Set

Mar 1
Apr 1
May 1
2005
Two Ways to Improve Statistical MT Systems

- Quality of resulting translation system
- Better algorithms
- More data

Amount of bilingual training data
Can a machine translate between Chinese and English without knowing what a verb is?

- Of course
- But the output is often bad

“Frequent high-tech exports are bright spots for foreign trade growth of Guangdong has made important contributions.”

- Our phrase-based story might need some work
Syntax

Maybe we need some grammar?
Syntax will never work!
We’re better off without syntax!
Syntax has been *shown* to make things worse!
It has never worked in speech recognition!

You are crazy!

Working on syntax-based approach
to translation (nouns, verbs,
prepositional phrases…)

Syntax will never work!
You need *semantics*!
Language is about the world!
You are crazy!

ACL Language Engineers

AAAI Fellows
MT Progress

Translation Quality (BLEU)

Phrase-based MT Progress

NIST Common Evaluations

syntax didn’t work

2002 2003 2004 2005

NIST Common Evaluations
Syntax Started to Be Helpful in 2006

Translation Accuracy

Chinese/English

Phrase-based

sentences < 16 words (NIST-03/04)

all sentences (NIST-2003)
How to Add Syntax?

• Automatically parse training data
  – Many parsers are available: (Collins 97, Charniak 01, etc)

• Then many approaches are possible
  – Add **syntactic features** to phrase-based system
    • many references
  – Syntactically **re-order source sentences** into target-like word order (for training and decoding)
    • (Berger et al 94, Xia & McCord 04, Collins et al 05, etc)
  – Build **tree-to-tree** translation systems
    • (Eisner 03, Gildea 03, Melamed 04, Riezler & Maxwell 06, Cowan et al 06, etc)
  – Build **tree-to-string** translation systems
    • (Quirk et al 05, Huang et al AMTA-06, Liu et al 06, etc)
  – Build **string-to-tree** translation systems
    • (Yamada & Knight 01, Galley et al 04, Venugopal & Zollmann 06, etc)

• Let’s just look at one approach & investigate
Phrase-Based Output

Hypothesis #1

Gunman of police killed.

Decoder
Hypothesis #1
Gunman of police attack.

Decoder
Hypothesis #7
Phrase-Based Output

Gunman by police killed.

Decoder
Hypothesis #12
Phrase-Based Output

Killed gunman by police.

Decoder
Hypothesis #134
Gunman killed the police.

Decoder
Hypothesis #9,329
Phraset-Based Output

Gunman killed by police.

Problematic:
- VBD “killed” needs a direct object
- VBN “killed” needs an auxiliary verb (“was”)
- countable “gunman” needs an article (“a”, “the”)
- “passive marker” in Chinese controls re-ordering

Can’t enforce/encourage any of this!
The gunman killed by police.
Gunman by police shot.
The gunman was killed by police.
Syntax-Based Output

• Better modeling of target language structure
  – Always a verb
  – Verb is always in the right place

• Better handling of function words
  – They often don’t translate
  – But they control how the translation goes

• Better generalization in translation patterns
Syntax-Based Statistical MT

• Terminology
• Mathematical Framework
• Translation Model
• Language Model
• Decoder
These 7 people include astronauts coming from France and Russia.
These 7 people include astronauts coming from France and Russia.
Mathematical Framework

• String-based system
  \[ \text{argmax}_{e,a} P(e, a, c)^\alpha \cdot P(e)^\beta \cdot |e|^{\gamma} \cdot \ldots \]

• Tree-based system
  \[ \text{argmax}_{\text{etree},a} P(\text{etree}, a, c)^\alpha \cdot P(\text{etree})^\beta \cdot |\text{etree}|^{\gamma} \cdot \ldots \]

- translation model
- language model
- length bonus
String-to-Tree

• Mathematically, we want a weighted relation with pairs drawn from:
  – (the infinite) set of Chinese strings
  – (the infinite) set of English trees

• Good pairs should have a high weight
• Bad pairs should have a low weight

• Probabilistic generative modeling approach
  – How does a Chinese string become an English tree (or vice-versa)?
An Early Syntactic Model of Translation

[Yamada & Knight 01]

**Parse (E)**

- S
  - NP
    - he
  - VB1
    - adores
  - VB2
    - listening
    - PP
      - P
        - to
      - NN
        - music

**Reorder**

- S
  - NP
    - he
  - PP
    - P
      - to
    - NN
      - music
  - VB2
  - VB1
    - adores
    - listening

**Insert**

- S
  - NP
    - he
  - PP
    - P
      - to
    - NN
      - music
  - VB2
  - VB1
    - adores
    - listening
  - ga
    - desu

**Translate**

- S
  - NP
    - kare
  - PP
    - P
      - to
    - NN
      - music
  - VB2
  - VB1
    - ga
    - desu

- S
  - NP
    - kare
  - PP
    - P
      - to
    - NN
      - music
  - VB2
  - VB1
    - ga
    - desu

**Take Leaves**

Kare ha ongaku wo kiku no ga daisuki desu

Sentence (J)
Phrase-Based

- Grab a chunk of English string
- Decide how to translate it (using phrase pair inventory)
- Recurse on remaining input
  - Can be modeled by finite-state string transducer
  - [Mealy, 1959] → [Kumar & Byrne, 2003, HLT]
Syntax-Based

- Grab a chunk of English input tree
- Decide how to translate it
- Recurse of remaining subtrees
  - Can be modeled by tree transducer
  - [Rounds, 1970] → [Graehl & Knight, 2004, HLT]
Top-Down Tree Transducer

Original input: he enjoys listening to music

Transformation:
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input:
he enjoys listening to music,

Transformation:
he enjoys wa ga listening to music,
Top-Down Tree Transducer

Original input:  
Transformation:

he enjoys listening to music

kare wa ga enjoys

listening to music
Top-Down Tree Transducer

Original input:

Final output:

he enjoys listening to music

kare, wa, ongaku, o, kiku, no, ga, daisuki, desu
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input:

Transformation:

```
q S
 x0:NP VP 0.2
 s x0, wa, r x2, ga, q x1
 x1:VBZ x2:NP
```

```
q S
 NP VP
 PRO VBZ NP
 he enjoys SBAR
 VBG VP
 listening P NP
 to music
```

```
q S
 NP VP
 PRO VBZ NP
 he enjoys SBAR
 VBG VP
 listening P NP
 to music
```
Top-Down Tree Transducer

Original input: he enjoys listening to music

Transformation: s NP PRO ‘wa’ SBAR ‘ga’ enjoys
Top-Down Tree Transducer

Original input:

Transformation:
Top-Down Tree Transducer

Original input: he enjoys listening to music,

Transformation: kare, wa, SBAR, ga, enjoys

Transformation: r NP, q VBZ

Transformation: VBG, VP

Transformation: listening, P, NP
to music
Top-Down Tree Transducer

Original input:  

Final output:

To get total probability, multiply probabilities of the individual steps.
An Early Syntactic Model of Translation
[Yamada & Knight 01]

Parse (E)

Reorder

Insert

Translate

Take Leaves

Sentence(J)

Kare ha ongaku wo kiku no ga daisuki desu
Can be cast as a single 4-state tree transducer.

[Gräehl & Knight 04; Gräehl, Knight & May 08]
Tree Transducers are Expressive

Phrasal Translation

Non-constituent Phrases

Non-contiguous Phrases

Context-Sensitive Word Insertion

Multilevel Re-Ordering

Lexicalized Re-Ordering

also QA, compression, paraphrasing, etc
most probabilistic tree-based models proposed 2000-2005 can be so cast
Limitations of the Top-Down Transducer Model

*Who does John think Mary believes I saw?* → *John thinks Mary believes I saw who?*

[Diagram showing syntactic structures]
Limitations of the
Top-Down Transducer Model

Who does John think Mary believes I saw?  \[\rightarrow\]  John thinks Mary believes I saw who?

[Diagram showing the transformation from one sentence structure to another.]
LIMITATIONS OF THE
TOP-DOWN TRANSDUCER MODEL

*Whose blue dog* does John think Mary believes I saw? ⇒ John thinks Mary believes I saw *whose blue dog?*

Can’t do this
Computer-Friendly Format for Tree Transducer Rules

**Phrasal Translation**

\[ VP \rightarrow \text{está, cantando} \]

\[ VBZ \quad VBG \]

\[ \text{is} \quad \text{singing} \]

\[ VP(VBZ(is), VBG(singing)) \rightarrow \text{está, cantando} \]

**Non-constituent Phrases**

\[ S \rightarrow \text{hay, x0} \]

\[ PRO \quad VP \]

\[ \text{there} \quad \text{VB} \quad x0:NP \]

\[ S(PRO(there), VP(VB(are), x0:NP)) \rightarrow \text{hay, x0} \]

**Non-contiguous Phrases**

\[ VP \rightarrow \text{poner, x0} \]

\[ VB \quad x0:NP \quad PRT \]

\[ \text{put} \quad \text{on} \]

\[ VP(VB(put), x0:NP, PRT(on)) \rightarrow \text{poner, x0} \]
These 7 people include astronauts coming from France and Russia.

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员．
Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → 
10. NP(x0:DT, CD(7), NNS(people)) → x0，7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自，x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0，x1，x2
14. VP(x0:VBP, x1:NP) → x0，x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0，x1，x2
16. NP(x0:NP, x1:VP) → x1，的，x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0，x1

Contiguous phrase pair substitution rules (alignment templates)

Higher-level rules
Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP.includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people)) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 , x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1 , x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0 , x1

Both VBP(“include”) and VBP(“includes”) will translate to “中包括” in Chinese.

In decoding Chinese, “中包括” is ambiguous and can translate back as either VBP(“include”) or VBP(“includes”).

Higher-level rules
Phrase pairs learned by alignment-templates that are relevant to this particular Chinese input sentence.

<table>
<thead>
<tr>
<th>Phrase pairs learned by alignment-templates that are relevant to this particular Chinese input sentence.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 people including by some the ruse and the ruse the astronauts</td>
</tr>
<tr>
<td>7 people included by france and the ruse international astronomical of rapporteur</td>
</tr>
<tr>
<td>7 out including the from the french and the ruse the fifth</td>
</tr>
<tr>
<td>7 among including from the french and of the ruse of space members</td>
</tr>
<tr>
<td>7 persons including from the of france and to ruse of the aerospace members</td>
</tr>
<tr>
<td>7 include from the of france and ruse astronauts the</td>
</tr>
<tr>
<td>7 numbers include from france and ruse of astronauts who</td>
</tr>
<tr>
<td>7 populations include those from france and ruse astronauts</td>
</tr>
<tr>
<td>7 deportees included come from france and ruse in astronomical personnel</td>
</tr>
<tr>
<td>7 philtume including those from france and ruse a space member</td>
</tr>
<tr>
<td>including representatives from france and the ruse astronaut</td>
</tr>
<tr>
<td>includes came from france and ruse by cosmonauts</td>
</tr>
<tr>
<td>include representatives from french and ruse cosmonauts</td>
</tr>
<tr>
<td>include came from france and ruse's cosmonauts</td>
</tr>
<tr>
<td>includes coming from french and ruse's's astronautavigation member</td>
</tr>
<tr>
<td>french and ruse's astronauts special rapporteur</td>
</tr>
<tr>
<td>and ruse's, and ruse cosmonaut</td>
</tr>
<tr>
<td>and ruse's cosmonaut</td>
</tr>
<tr>
<td>and russia's rapporteur</td>
</tr>
</tbody>
</table>

Table 1: #11# the seven - member crew includes astronauts from france and russia.
Phrase pairs learned by alignment-templates that are relevant to this particular Chinese input sentence.

<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>这 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员 ．</td>
<td>The 7 people including by some French and Russian astronauts．</td>
</tr>
</tbody>
</table>

Table 1: The seven-member crew includes astronauts from France and Russia.

Only top 5 translations-per-Chinese-phrase are shown here – there are many more.
Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people)) → x0, 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自, x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0, x1, x2
14. VP(x0:VBP, x1:NP) → x0, x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0, x1, x2
16. NP(x0:NP, x1:VP) → x1, 的, x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0, x1

The phrase “coming from” translates to “来自” only if followed by an NP (whose translation is then placed to the right of “来自”).

Higher-level rules
Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP/includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 , 7
11. VP(VBG(coming), PP(IN(from), x0:NP) → x0 , x1
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

Translate an English NP (“astronauts”) modified by a gerund VP (“coming from France and Russia”) as follows:
(1) translate the gerund VP,
(2) type the Chinese word “的”，
(3) translate the NP.

In decoding Chinese, if we analyze
(1) some Chinese into an English NP &
(2) some other Chinese into an English VP
and these two bits are separated by “的”，
then create an English NP(NP, VP) structure.

Higher-level rules
To translate “the JJ NN”, just translate the JJ and then translate the NN (drop “the”).

When we are decoding Chinese, if we create an English JJ and an adjacent English NN, we can hook these together into an NP, and also insert the word “the.”

Most frequent deficiency of lattices is the lack of critical English function words!
### Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people)) → x0, 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自, x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0, x1, x2
14. VP(x0:VBP, x1:NP) → x0, x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0, x1, x2
16. NP(x0:NP, x1:VP) → x1, 的, x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0, x1

---

**Note that this rule goes ahead and makes “astronauts” a full NP. Might be better to have two rules:**

NNS(astronauts) → 宇航员
NP(x0:NNS) → x0

---

Higher-level rules
Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people)) → x0, 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自, x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0, x1, x2
14. VP(x0:VBP, x1:NP) → x0, x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0, x1, x2
16. NP(x0:NP, x1:VP) → x1, 的, x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0, x1

Okay, these rules look interesting.

It would be cool if we could acquire rules like these from data!!
Phrase-Based and Syntax-Based Pattern Extraction

string alignment

 ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment
Phrase-Based and Syntax-Based Pattern Extraction

**ATS** [Och & Ney, 2004]

phrase pairs consistent with word alignment

**GHKM** [Galley et al 2004, 2006]

syntax transformation rules consistent with word alignment
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

 RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
...
S(x0:NP-C x1:VP) → x0 x1
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

- VBD(felt) → 有
- VBN(obliged) → 责任
- VB(do) → 尽
- NN(part) → 一份
- NN(part) → 一份 力
- VP-C(x0:VBN x1:SG-C) → x0 x1
- VP(TO(to) x0:VP-C) → x0
- ...
- S(x0:NP-C x1:VP) → x0 x1

i felt obliged to do my part

我 有 责任 尽 一份 力
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

- VBD(felt) \rightarrow 有
- VBN(obliged) \rightarrow 责任
- VB(do) \rightarrow 尽
- NN(part) \rightarrow 一份
- NN(part) \rightarrow 一份力
- VP-C(x0:VBN x1:SG-C) \rightarrow x0 x1
- VP(TO(to) x0:VP-C) \rightarrow x0
- ... 
- S(x0:NP-C x1:VP) \rightarrow x0 x1
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
...
S(x0:NP-C x1:VP) → x0 x1

There is a unique tiling that identifies minimal translation units.
Sample “said that” rules

```
0.57 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说，x0
0.09 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 x0
0.02 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 他说，x0
0.02 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 指出，x0
0.02 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> x0
0.01 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 表示 x0
0.01 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说，x0 的
```
Sample “NP-from-NP” rules

NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 的 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 从 x1 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 的 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x0 从 x1
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 自 x1 x0
NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0,
Sample SVO rules

\[
S \quad \xrightarrow{?} \quad x_0:NP-C \quad VP \quad x_3:.
\]

\[
x_1:VBD \quad x_2:NP-C
\]

**CHINESE / ENGLISH**

- 0.82 \( S(x_0:NP-C \ VP(x_1:VBD \ x_2:NP-C) \ x_3:.\) -> x_0 \ x_1 \ x_2 \ x_3 
- 0.02 \( S(x_0:NP-C \ VP(x_1:VBD \ x_2:NP-C) \ x_3:.\) -> x_0 \ x_1 , x_2 \ x_3 
- 0.01 \( S(x_0:NP-C \ VP(x_1:VBD \ x_2:NP-C) \ x_3:.\) -> x_0 , x_1 \ x_2 \ x_3 

**ARABIC / ENGLISH**

- 0.54 \( S(x_0:NP-C \ VP(x_1:VBD \ x_2:NP-C) \ x_3:.\) -> x_0 \ x_1 \ x_2 \ x_3 
- 0.44 \( S(x_0:NP-C \ VP(x_1:VBD \ x_2:NP-C) \ x_3:.\) -> x_1 \ x_0 \ x_2 \ x_3 

Extensions to Rule Extraction from Data  [Galley et al 06]

Enumerate all ways of dealing with unaligned Chinese words.

Generate rule counts which can be normalized into probabilities.
Language Models

• Syntax-based Language Model
  – Assigns $P(\text{tree})$
    • [Collins 97; Charniak 01]
  – NOTE: Unlike parser, must be trained on domain data

• Ngram Language Model
  – Standard trigram model
  – Only judges a tree by its leaves
BREAK

• When We Come Back:
  – Review of syntax-based translation models
  – Syntax-based decoding
  – Is syntax harmful?
    • yes
    • what can be done
  – Sample outputs
  – Open problems
  – Connections to automata
  – Conclusions
  – Discussion
Phrase-Based and Syntax-Based Pattern Extraction

**ATS** [Och & Ney, 2004]

phrase pairs consistent with word alignment

**GHKM** [Galley et al. 2004, 2006]

syntax transformation rules consistent with word alignment
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
...
S(x0:NP-C x1:VP) → x0 x1
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

- VBD(felt) → 有
- VBN(oblige) → 责任
- VB(do) → 尽
- NN(part) → 一份
- NN(part) → 一份力
- VP-C(x0:VBN x1:SG-C) → x0 x1
- VP(TO(to) x0:VP-C) → x0
- S(x0:NP-C x1:VP) → x0 x1
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

- VBD(felt) → 有
- VBN(obliged) → 责任
- VB(do) → 尽
- NN(part) → 一份
- NN(part) → 一份力
- VP-C(x0:VBN x1:SG-C) → x0 x1
- VP(TO(to) x0:VP-C) → x0
- ...
- S(x0:NP-C x1:VP) → x0 x1
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

i felt obliged to do my part

RULES ACQUIRED:

VBD(felt)  →  有
VBN(obliged)  →  责任
VB(do)  →  尽
NN(part)  →  一份
NN(part)  →  一份 力
VP-C(x0:VBN x1:SG-C)  →  x0 x1
VP(TO(to) x0:VP-C)  →  x0
...
S(x0:NP-C x1:VP)  →  x0 x1
Sample “said that” rules

0.57 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 说, x0}$
0.09 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 说 x0}$
0.02 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 他说, x0}$
0.02 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 指出, x0}$
0.02 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow x0}$
0.01 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 表示 x0}$
0.01 $\text{VP(VBD("said") SBAR-C(IN("that") x0:S-C)) \rightarrow 说, x0 的}$
Sample “NP-from-NP” rules

\[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow x_1 x_0, \]

0.27 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow x_1 x_0 \]

0.15 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow \text{来自} x_1 x_0 \]

0.06 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow x_1 \text{的} x_0 \]

0.06 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow \text{从} x_1 x_0 \]

0.06 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow \text{来自} x_1 \text{的} x_0 \]

0.02 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow x_0 \text{从} x_1 \]

0.01 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow \text{自} x_1 x_0 \]

0.01 \[ \text{NP-C}(x_0:NPB, \text{PP}(\text{IN}("from") x_1:NP-C)) \rightarrow x_1 x_0 , \]
Sample SVO rules

```
S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x0 x1 x2 x3
0.82  S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x0 x1 x2 x3
0.02  S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x0 , x1 x2 x3
0.01  S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x0 , x1 x2 x3

CHINESE / ENGLISH

ARABIC / ENGLISH
```

```
0.54  S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x0 x1 x2 x3
0.44  S(x0:NP-C VP(x1:VBD x2:NP-C) x3:. ) -> x1 x0 x2 x3
```
Hiero (Chiang 05, 07)

- Phrase pairs with variables
  - e.g., “of X ↔ X de”
- Hierarchical decoding
  - the X itself could be created via other phrase pairs with variables
- Only one syntactic symbol in rules
  - X
- Translation patterns can be extracted without syntactically parses of the training data
Hiero Grammar Extraction

Australia is one of the few countries that have diplomatic relations with North Korea.

In Hiero literature, this rule is written in synchronous grammar format:

\[
(X \rightarrow \text{have X2 with X1})
\]
Sample Hiero rules
(in tree transducer format)

X(‘s) \rightarrow  
X(the x0:X of x1:X) \rightarrow x1 的 x0
X(the x0:X that x1:X) \rightarrow x1 的 x0

X(in) \rightarrow 在
X(under x0:X) \rightarrow 在 x0 下
X(before x0:X) \rightarrow 在 x0 前

X(x0:X this year) \rightarrow 今年 x0
X(one of x0:X) \rightarrow x0 之一
X(president x0:X) \rightarrow x0 总统
### Table 1: #11# the seven-member crew includes astronauts from France and Russia.

<table>
<thead>
<tr>
<th>These 7 people including by some of the Russian the astronauts,</th>
<th>Try to use phrase pairs that have been frequently observed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>and international.</td>
<td>Try to output a sentence with frequent English word sequences.</td>
</tr>
</tbody>
</table>
Syntax-Based Decoding

• Bottom-up CKY parser
• Builds English constituents on top of Chinese spans
• Record of rule applications (the derivation) provides information to construct English tree
• Returns k-best trees
• Same decoder can handle syntax translation rules and Hiero rules
这7人中包括来自法国和俄罗斯的宇航员。
这 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员。
This 7 people include the astronauts coming from France and Russia.
这7人中包括来自法国和俄罗斯的宇航员。
“include astronauts coming from France and Russia”

“include”

“France”

“&”

“Russia”

“astronauts”

“.”

“these”

“include”

“France”

“&”

“Russia”

“astronauts”

“.”

这 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员 ．
These 7 people include astronauts coming from France and Russia.
These 7 people include astronauts coming from France and Russia.
These 7 people include astronauts coming from France and Russia.
Aside if you are familiar with monolingual parsing.
Binarization for Decoding

- For CKY decoding, all rules must be *binarized*.
- Rule with $|\text{RHS}| > 2$ must be split into rules with $|\text{RHS}| = 2$
  - $S(x_0:\text{NP} \ VP(x_1:\text{VBD} x_2:\text{NP})) \rightarrow x_1 x_0 x_2$
  - $Z(x_0:\text{NP} x_1:\text{VBD}) \rightarrow x_1 x_0$
  - $S(x_0:Z x_1:\text{NP}) \rightarrow x_0 x_1$
- Similar to putting a CFG into Chomsky normal form.
- A rule can be binarized in different ways: must pick best!
- Some translation rules cannot be binarized at all…
  - $A(x_0:B x_1:C x_2:D x_3:E) \rightarrow x_1 x_3 x_0 x_2$ [Wu 96]
- We just delete these.

- Binarization details: [Zhang, Huang, Knight, Gildea, 2006]
he said that the process is practical, which would take six months.
Why Might Syntax Help?

• Phrase-based MT output is “n-grammatical”, not grammatical
  – Every sentence needs a subject and a verb

• Re-ordering is poorly explained as “distortion” -- better explained as syntactic transformation
  – Arabic to English, VSO $\rightarrow$ SVO

• Function words have syntactic effects even if they are not themselves translated
Why Might Syntax Hurt?

• Less freedom to glue pieces of output together -- search space has fewer output strings
• Search space is more difficult to navigate
• Rule extraction from bilingual text has limitations

available phrase-based translations

this section
Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings
- Search space is more difficult to navigate
- Rule extraction from bilingual text has limitations

this section
Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings
- Search space is more difficult to navigate
- Rule extraction from bilingual text has limitations
Comparing Phrase-Based Extraction with Syntax-Based Extraction

• Quantitatively compare
  – A typical phrase-based bilingual extraction algorithm (ATS, Och & Ney 2004)
  – A typical syntax-based bilingual extraction algorithm (GHKM, Galley et al 2004)
    – These algorithms picked from two good-scoring NIST-06 systems

• Identify areas of improvement for syntax-based rule coverage
Phrase-Based and Syntax-Based Pattern Extraction

**ATS [Och & Ney, 2004]**

phrase pairs consistent with word alignment

**GHKM [Galley et al 2004]**

syntax transformation rules consistent with word alignment
ATS (Och & Ney, 2004)

PHRASE PAIRS ACQUIRED:

felt
felt obliged
felt obliged to do  →  有 责 任 尽
obliged
obliged to do  →  责 任 尽
do
part
part

i felt obliged to do my part

我 有 责 任 尽 一份 力
ATS (Och & Ney, 2004)

PHRASE PAIRS ACQUIRED:

felt \rightarrow 有
felt obliged \rightarrow 有 责任
felt obliged to do \rightarrow 有 责任 尽
obliged \rightarrow 责任
obliged to do \rightarrow 责任 尽
do \rightarrow 尽
part \rightarrow 一份
part \rightarrow 一份 力
ATS (Och & Ney, 2004)

PHRASE PAIRS ACQUIRED:

- felt → 有
- felt obliged → 有 责任
- felt obliged to do → 有 责任 尽
- obliged → 责任
- obliged to do → 责任 尽
- do → 尽
- part → 一份
- part → 一份 力

i felt obliged to do my part
我 有 责任 尽 一份 力
ATS (Och & Ney, 2004)

**PHRASE PAIRS ACQUIRED:**

- felt → 有
- felt obliged → 有责任
- felt obliged to do → 有责任尽
- obliged → 责任
- obliged to do → 责任尽
- do → 尽
- part → 一份
- part → 一份力

- i felt obliged to do my part
- 我有责任尽一份力
GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
  VP-C(x1:VBN
    x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1
I felt obliged to do my part

我有责任尽一份力

RULES ACQUIRED:

- VBD(felt) → 有
- VBN(obliged) → 责任
- VP(x0:VBD
  VP-C(x1:VBN
  x2:SG-C) → x0 x1 x2
- S(x0:NP-C x1:VP) → x0 x1
i felt obliged to do my part

我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VP(x0:VBD VP-C(x1:VBN x2:SG-C)) → x0 x1 x2
S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)
i felt obliged to do my part

我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt)  →  有
VBN(oblighed)  →  责任

VP(x0:VBD
  VP-C(x1:VBN
    x2:SG-C)  →  x0 x1 x2

S(x0:NP-C x1:VP)  →  x0 x1
I felt obliged to do my part.

我有责任尽一份力

There is a unique tiling that identifies minimal translation units.
GHKM Syntax Rules

**Phrasal Translation**

\[
VP \rightarrow \text{está, cantando}
\]

\[
\begin{align*}
& \text{VBZ} \quad \text{VBG} \\
& \text{is} \quad \text{singing}
\end{align*}
\]

**Non-constituent Phrases**

\[
S \rightarrow \text{hay, NP}
\]

\[
\begin{align*}
& \text{PRO} \\
& \text{there} \quad \text{VP} \\
& \quad \text{VB} \quad \text{NP} \\
& \quad \text{are}
\end{align*}
\]

**Non-contiguous Phrases**

\[
VP \rightarrow \text{poner, NP}
\]

\[
\begin{align*}
& \text{vb} \\
& \text{np} \\
& \text{pRT} \\
& \text{put} \\
& \text{on}
\end{align*}
\]

**Context-Sensitive Word Insertion**

\[
\text{NPB} \rightarrow \text{NNS}
\]

\[
\begin{align*}
& \text{DT} \\
& \text{the} \\
& \text{NNS}
\end{align*}
\]

**Multilevel Re-Ordering**

\[
S \rightarrow \text{VB, NP1, NP2}
\]

\[
\begin{align*}
& \text{NP1} \\
& \text{VP} \\
& \text{VB} \\
& \text{NP2}
\end{align*}
\]

**Lexicalized Re-Ordering**

\[
\text{NP} \rightarrow \text{NP1, of, NP2}
\]

\[
\begin{align*}
& \text{NP2} \\
& \text{PP} \\
& \text{P} \\
& \text{NP1}
\end{align*}
\]
GHKM Syntax Rules

Phrasal Translation

\[ VP \rightarrow \text{está, cantando} \]

\[ \text{VBZ} \quad \text{VBG} \]

\[ \text{is} \quad \text{singing} \]

Non-constituent Phrases

\[ S \rightarrow \text{hay, NP} \]

\[ \text{PRO} \quad \text{VP} \]

\[ \text{there} \quad \text{VB} \quad \text{NP} \]

\[ \text{are} \]

Non-contiguous Phrases

\[ VP \rightarrow \text{poner, NP} \]

\[ \text{VB} \quad \text{NP} \quad \text{PRT} \]

\[ \text{put} \quad \text{on} \]

Context-Sensitive Word Insertion

\[ \text{NPB} \rightarrow \text{NNS} \]

\[ \text{DT} \quad \text{NNS} \]

\[ \text{the} \]

Multilevel Re-Ordering

\[ S \rightarrow \text{VB, NP1, NP2} \]

\[ \text{NP1} \quad \text{VP} \]

\[ \text{VB} \quad \text{NP2} \]

Lexicalized Re-Ordering

\[ \text{NP} \rightarrow \text{NP1, PP, NP2} \]

\[ \text{NP2} \quad \text{PP} \]

\[ \text{of} \]
ATS and GHKM Methods Do Not Coincide

GHKM has no built-in phrase size limit -- ATS does.

GHKM pulls unaligned English words into phrases.

GHKM only gets *minimal* rules to explain each segment pair.

GHKM forced to incorporate unaligned English words into phrases.

GHKM forced to incorporate some unaligned foreign words into phrases.

GHKM misses phrases due to parse failures.

GHKM phrases come with applicability conditions.

ATS Phrase Pairs Relevant to NIST-02

43k

161k

134k

GHKM Phrase Pairs Relevant to NIST-02
ATS and GHKM Methods Overlap

ATS Phrase Pairs actually used in 1-best decodings of NIST-02 (1,994 = 2 per sentence).

GHKM Phrase Pairs Relevant to NIST-02

1,994

GHKM phrases come with applicability conditions.

GHKM only gets minimal rules to explain each segment pair.

GHKM forced to incorporate some unaligned foreign words into phrases.

GHKM forced to incorporate unaligned English words into phrases.

GHKM misses phrases due to parse failures.

CAN WE REDUCE THIS NUMBER?
Some Methods for Improving Syntax-Based Rule Extraction

- Acquire larger rules
  Composed rules (Galley et al, 06)
  Phrasal rules (Marcu et al, 06)
- Acquire more general rules
  Re-structure English trees (Wang et al, 07)
  Re-align tree/string pairs (May & Knight, 07)
- Expand syntactic category set
  Slash categories (Zollmann & Venugopal 06)
Larger, Composed Rules

Minimal GHKM Rules:

B(e1 e2) $\rightarrow$ c1 c2
C(e3) $\rightarrow$ c3
A(x0:B x1:C) $\rightarrow$ x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) $\rightarrow$ c1 c2 x0
A(x0:B C(e3)) $\rightarrow$ x0 c3
A(B(e1 e2) C(e3)) $\rightarrow$ c1 c2 c3

"big phrasal rule"
Larger, Composed Rules

Minimal GHKM Rules:

\[ B(e_1 e_2) \rightarrow c_1 c_2 \]
\[ C(e_3) \rightarrow c_3 \]
\[ A(x_0:B \ x_1:C) \rightarrow x_0 \ x_1 \]

Additional Composed Rules:

\[ A(B(e_1 e_2) \ x_0:C) \rightarrow c_1 \ c_2 \ x_0 \]
\[ A(x_0:B \ C(e_3)) \rightarrow x_0 \ c_3 \]
\[ A(B(e_1 e_2) \ C(e_3)) \rightarrow c_1 \ c_2 \ c_3 \]

“big phrasal rule”
Larger, Composed Rules

Minimal GHKM Rules:

B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) → c1 c2 x0
A(x0:B C(e3)) → x0 c3
A(B(e1 e2) C(e3)) → c1 c2 c3

“big phrasal rule”
Larger, Composed Rules

Minimal GHKM Rules:

B(e₁ e₂) → c₁ c₂
C(e₃) → c₃
A(x₀:B x₁:C) → x₀ x₁

Additional Composed Rules:

A(B(e₁ e₂) x₀:C) → c₁ c₂ x₀
A(x₀:B C(e₃)) → x₀ c₃
A(B(e₁ e₂) C(e₃)) → c₁ c₂ c₃

“big phrasal rule”
GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VP(x0:VBD
  VP-C(x1:VBN
    x2:SG-C) → x0 x1 x2

VP(VBD(felt)
  VP-C(VBN(obliged)))
  x0:SG-C) → 有 责任 x0

S(x0:NP-C x1:VP) → x0 x1
GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt)       → 有
VBN(obliged)    → 责任

VP(x0:VBD
   VP-C(x1:VBN
      x2:SG-C) → x0 x1 x2

VP(VBD(felt)
   VP-C(VBN(obliged))
      x0:SG-C) → 有 责任 x0

S(x0:NP-C x1:VP) → x0 x1
GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VP(x0:VBD
   VP-C(x1:VBN
      x2:SG-C) → x0 x1 x2

VP(VBD(felt)
   VP-C(VBN(obliged))
      x0:SG-C) → 有 责任 x0

S(x0:NP-C x1:VP) → x0 x1

minimal rules tile the tree/string/alignment triple. composed rules are made by combining those tiles.
## Larger, Composed Rules

<table>
<thead>
<tr>
<th>Composed limit (internal nodes in composed rule)</th>
<th># of rules acquired</th>
<th>Unacquired phrase pairs used in ATS 1-best decodings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = minimal</td>
<td>2.5m</td>
<td>1994</td>
</tr>
<tr>
<td>2</td>
<td>12.4m</td>
<td>1478</td>
</tr>
<tr>
<td>3</td>
<td>26.9m</td>
<td>1096</td>
</tr>
<tr>
<td>4</td>
<td>55.8m</td>
<td>900</td>
</tr>
</tbody>
</table>
“Phrasal” Syntax Rules

• SPMT Model 1 (Marcu et al 2006)
  – consider each foreign phrase up to length L
  – extract smallest possible syntax rule that does not violate alignments

<table>
<thead>
<tr>
<th>Method</th>
<th>Unacquired ATS Phrase Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal</td>
<td>1994</td>
</tr>
<tr>
<td>Composed 4</td>
<td>900</td>
</tr>
<tr>
<td>SPMT M1</td>
<td>676</td>
</tr>
<tr>
<td>Both</td>
<td>663</td>
</tr>
</tbody>
</table>
Restructuring English Training Trees

The Israeli Prime Minister Ariel Sharon
## Restructuring English Training Trees

<table>
<thead>
<tr>
<th>Method</th>
<th>Unacquired ATS Phrase Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal</td>
<td>1994</td>
</tr>
<tr>
<td>+ Composed 4</td>
<td>900</td>
</tr>
<tr>
<td>+ SPMT M1</td>
<td>663</td>
</tr>
<tr>
<td>+ Restructuring</td>
<td>458</td>
</tr>
</tbody>
</table>
# Effects of Coverage Improvements on Syntax-Based MT Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Chinese/English Trained on 9.8m words</th>
<th>Arabic/English Trained on 4.1m words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev-02</td>
<td>Test-03</td>
</tr>
<tr>
<td>ATS</td>
<td>36.00</td>
<td>34.31</td>
</tr>
<tr>
<td>GHKM minimal</td>
<td>39.11</td>
<td>38.85</td>
</tr>
<tr>
<td>GHKM composed 2</td>
<td>41.59</td>
<td>40.90</td>
</tr>
<tr>
<td>GHKM composed 3</td>
<td>42.28</td>
<td>41.62</td>
</tr>
<tr>
<td>GHKM composed 4</td>
<td>42.63</td>
<td>41.82</td>
</tr>
<tr>
<td>GHKM minimal + SPMT</td>
<td>41.01</td>
<td>40.34</td>
</tr>
<tr>
<td>GHKM composed 4 + SPMT</td>
<td>43.30</td>
<td>42.17</td>
</tr>
<tr>
<td>+ Left binarization of etrees</td>
<td>43.45</td>
<td>42.41</td>
</tr>
</tbody>
</table>

NIST Bleu r4n4
Improved English Binarization

Why are Penn Treebank Trees Problematic for Translation?

维克多·切尔诺梅尔金 及 其 同事

俄罗斯 首相 维克多·切尔诺梅尔金

和 他的 同事

维克多·切尔诺梅尔金 及 其 同事
Improved English Binarization

Why are Penn Treebank Trees Problematic for Translation?
Binarizing English Trees

Right binarize

Left binarize
Simple Binarizations

(1) unbinarized tree

(2) left-binarization

(3) right-/head-binarization

(4) left-binarization

(5) right-binarization

(6) left-binarization

(7) right-/head-binarization
Parallel Binarization
Parallel Binarization

维克多·切尔诺梅尔金

Relation: NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB

NPB
Forest-Based Rule Extraction

• Gets all minimal rules consistent with word alignment and some binarization
• Run EM algorithm to determine best binarization of each node in each tree
Binarization Using EM

- e-tree
- parallel binarization
- e-forest
- forest-based extraction of minimal rules
  - rules
  - derivation forests
  - EM

- f-string, alignment
- project e-tree
  - viterbi derivation for each example
- composed rule extraction (Galley et al., 2006)
  - rules for decoding
## Experimental Results
(Wang, Knight, Marcu 2007)

<table>
<thead>
<tr>
<th>Type of Binarization</th>
<th># of Rules Learned</th>
<th>Test Bleu (NIST-03)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>63.4m</td>
<td>36.94</td>
</tr>
<tr>
<td>Left</td>
<td>114.0m</td>
<td>37.47 (p=0.047)</td>
</tr>
<tr>
<td>Right</td>
<td>113.0m</td>
<td>37.49 (p=0.044)</td>
</tr>
<tr>
<td>Head</td>
<td>113.8m</td>
<td>37.54 (p=0.086)</td>
</tr>
<tr>
<td>EM</td>
<td>115.6m</td>
<td><strong>37.94 (p=0.0047)</strong></td>
</tr>
</tbody>
</table>
Tree binarized by EM training
Syntax-Based Word Alignment

• GIZA++ string-based alignments
  – are errorful
  – don’t match our syntax-based MT system

• We would like to use our tree-based translation model to align data
Syntax-Based Word Alignment

English trees
Foreign strings

GIZA++ \rightarrow initial word alignments \rightarrow \text{GHKM syntax rule extraction} \rightarrow \text{minimal rules}

\text{EM alignment ("Training Tree Transducers", Graehl & Knight'04)}

\rightarrow \text{Viterbi derivations} \rightarrow \text{Improved word alignments}

\rightarrow \text{composed rule extraction}

May & Knight, 07

Result: +0.5-1.0 Bleu

\rightarrow \text{better rules for decoding}
Remarks

• Phrase-based and syntax-based extraction algorithms have different coverage.
• Syntax-based coverage can be improved:
  – composed rules
  – phrasal rules
  – binarizing English trees with EM
  – re-aligning tree/string pairs with EM
• Improvements lead to better translation accuracy.
Some Sample Outputs
The Chinese-funded enterprises have become the largest external investor in Macao.
Kinana thanked China for substantial assistance to Tanzania in the past.
the two-day seminar is jointly sponsored by the WTO Shanghai Research Center and Shanghai Foreign Service Company Limited.

the two-day seminar is organized jointly by Shanghai Foreign Services Ltd. and Shanghai Research Center of the World Trade Organization.
Ding Hao completed his primary school at the children welfare institute, and then went to a nearby township middle school. He subsequently entered junior high school at a school near their home towns.
Input: 他确信， 加 - 中两国可以成为很好的合作伙伴。
Reference: he assured that canada and china can become very good partners.
AlTemp-e: he was convinced that 0 | the 1 | two countries 2 | can 3 | become good 4 | partners 5 | . 7
AlTemp-f: 他确信 , 0 | 加 1 | 、 中 2 | 两国 3 | 可以 4 | 成为 很 好 的 5 | 合作伙伴 6 | 、 7
[dev-little] 1-Best: he is convinced that canada, china could be a good partner.

[dev-little] 1-Best Tree

```
S
  | VP
  |   | NPB
  |   | VHZ
  |   | VP-C
  |     | PRP
  |     | is
  |     | VBN
  |     | SBAR-C
  |       | that
  |       | IN
  |       | S-C
  |   | NPB
  |   | MD
  |   | VP
  |   | VP-C
  |     | NNP
  |     | NNP
  |     | VB
  |     | NP-C

Subject-Verb Agreement:
```
Input: the French foreign minister made the above statement in a meeting of the foreign affairs commission of the French national congress.
Reference: the French foreign minister made the above statement in a meeting of the French national congress.
AlTemp-e: French foreign minister in the French National Assembly yesterday, the statement delivered by the foreign affairs committee meeting yesterday.
[dev-little] 1-Best: the French foreign minister was the statement issued at the French National Assembly foreign affairs committee meeting yesterday.

[dev-little] 1-Best Tree
this year made love in california and southern areas of torrential rains were blamed on el nino.
Lots of Open Problems
Chomsky’s Program [1957]

- Algorithmically distinguish between grammatical and ungrammatical sentences:
  - John thinks Sara hit the boy
  - * The hit thinks Sara John boy
  - John thinks the boy was hit by Sara
  - Who does John think Sara hit?
  - John thinks Sara hit the boy and the girl
  - * Who does John think Sara hit the boy and?
  - John thinks Sara hit the boy with the bat
  - What does John think Sara hit the boy with?
  - Colorless green ideas sleep furiously.
  - * Green sleep furiously ideas colorless.
This Research Program has Contributed Powerful Ideas

- Context-free grammar
- Formal language hierarchy
- Syntax, Phonology...
This Research Program is Really Unfinished

Type in your English sentence here:

Is this grammatical?
Is this sensible?
Lots of Open Problems

- Modeling English fluency, using trees
  - phrase-based output – need to parse it to score it
  - syntax-based output – already in scorable tree form
  - initial work: [Charniak, Knight, and Yamada, 2003]

- Choosing syntactic categories that are appropriate for translation
  - initial work: [B. Huang and K. Knight, 2006]

- Decoder search in runtime translation
  - Search errors hurt MT accuracy
  - Faster speed is needed to support experimentation
  - Some key ideas to date:
    - cube pruning [Chiang, 2007]
    - rule binarization [Zhang, Huang, Knight, Gildea, 2006]
Lots of Open Problems

• More context for rule choice
  – compare word-based SMT
    • context-sensitive word translation probabilities [Berger et al 96]
  – compare phrase-based SMT
    • bilingual n-gram translation models [de Gispert & Mariño 02]
    • context-based phrasal TM “WSD” [Chan, Ng, Chiang 07; Carpuat & Wu 07]

• Morphology in translation rules
• More generally applicable rules
  – Adjoining transducers (tree-adjoining grammar)
• Open theory problems in the underlying automata models…
Tree Automata

Tiburon: A Tree Automata Toolkit

- Developed by Jonathan May, ISI
- First version distributed in April 2006, includes tutorial
- Inspired by string automata toolkits
- Prototype ideas, teach tree automata to yourself or others

- You cast your problem in terms of tree acceptors and transducers
  - doesn’t have to be MT
- You get implemented algorithms for free
  - e.g., Kumar/Byrne’03 (use AT&T FSM for MT)
  - e.g., Pereira/Riley’96 (use AT&T FSM for ASR)
Tiburon: A Tree Automata Toolkit

Towards simplifying system ideas:

\[ e = \text{yield(best-tree(intersect(lm.rtg,}
\[ b\text{-apply(cstring, tm.tt)))} \]

What tree automata operations are needed/supported?
<table>
<thead>
<tr>
<th>String World</th>
<th>Tree World</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-best …</td>
<td>… paths through a lattice (Viterbi, 1967; Eppstein, 1998)</td>
</tr>
<tr>
<td>EM training</td>
<td>Forward-backward EM (Baum &amp; Welch, 1971)</td>
</tr>
<tr>
<td>Determinization …</td>
<td>… of weighted string acceptors (Mohri, 1997)</td>
</tr>
<tr>
<td>Intersection</td>
<td>WFSA intersection</td>
</tr>
<tr>
<td>Applying transducers</td>
<td>string → WFST → WFSA</td>
</tr>
<tr>
<td>Transducer composition</td>
<td>WFST composition (Pereira &amp; Riley, 1996)</td>
</tr>
</tbody>
</table>
Classes of Tree Transducers

copying
non-copying
deleting
non-deleting
copying rule
deleting rule
Classes of Tree Transducers

Expressive power theorems in Maletti, Graehl, Hopkins, Knight (submitted)
Classes of Tree Transducers

copying

non-copying

deleting

non-deleting

bottom up transducers

L-MBOT M’06

MGHK’08

GHKM’04

MGHK’08

GS’84

GS’84

T3

B3

T2

B2

T

B

xT = TR

xLT

xLNT

T

LT

LB = LTR

LNT

GK’04

GS’84
Classes of Tree Transducers

copying
non-copying
deleting
non-deleting

xLNT
(xRHS, output-e)
GHKM’04

xLNT
(xRHS, output-e)
MGHK’07

xLNT
(xRHS, input-e, output-e)

LNT
(xRHS, input-e)

LNT
(xRHS, input-e)

LNT
(xRHS, output-e)
GS’84

LNT
(xRHS, e-free)

LNT
(xRHS, e-free)

T

B

T^3

B^3

xLT
LT

xT

xT_R=T_R

T^2

B^2

L-B=L_T_R

L-MBOT
M’06

GK’04

GS’84

copying
non-copying
deleting
non-deleting
Transducer hierarchy recently related to **synchronous grammars** by (Shieber 04, 06). See tutorials by Chiang/Knight ACL’06 & Shieber ESSLLI’05.
Research Synergy

Machine Translation Research

Clean representations, fast algorithms

Challenges coming from empirical data

Tree Automata Research

Clean representations, fast algorithms

User feedback

Tree Automata Software

New applications, prototypes, projects…
This Is Interdisciplinary Research

- Machine Learning
- Engineering
- Linguistics
- Data
- Efficient search algorithms
- Automata theory
- Grid computing

...