Phrase-Based Translation: Discontinuous Phrases and Hierarchical Reordering

Michel Galley
MSR

with slides borrowed from
Kristina Toutanova, Chris Quirk, Philipp Koehn
Types of SMT systems

- **syntax-based**
  - (lectures 6-8)
- **phrase-based**
  - (lectures 3-4)
- **word-based**
  - (lectures 2-3)

This lecture
Phrase-based SMT

- Translate more than one word at a time
  - “Phrase”: contiguous sequence of words (not necessarily linguistic)

tomorrow  I will fly  to the conference  in Canada
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)
Benefits of phrase-Based SMT

- Phrases offer: multi-word expressions, help select the correct function words (e.g., "for"), and enable local re-orderings.

How phrase-based SMT would work with massive amounts of data.
Drawbacks of phrase-Based SMT

How phrase-based SMT work in practice (actual SMT output):

- **Poor lexical substitutions:** long phrases of the *training* data often not usable at *test* time, and short phrases often yield poor translations. Phrase-based SMT is particularly vulnerable to data sparseness.

- **Poor word order:** phrase-based word reordering models lack rich analyses (e.g., syntactic) to guide the translation process.
Remedies

• **More expressive representations**
  - e.g., syntax-based SMT (lectures 6-8)
  - syntactic categories (NP, VP) less subject to data sparseness than words
  - syntactically-informed reordering

*or*

• **Extended phrased-based SMT (this lecture)**
  - discontinuous phrases
  - hierarchical phrase-based reordering
Rescue workers search for survivors in collapsed houses.

Decoding with CKY: $O(n^3)$

(→ Syntax and parsing LING 571)
Discontinuous Phrase-based SMT

• Naturally copes with long-distance word dependencies (e.g., between 在 and 里)
• “Phrases”: arbitrary subsequences of source/target sentences (again, need not be linguistically well-formed).

[Galley and Manning, 2010]
In the last part of this lecture, we will see how to use both gaps and hierarchical reordering in phrase-based SMT.
Discontinuous Phrase-based SMT

Step I: Phrase Induction
<table>
<thead>
<tr>
<th>Non</th>
<th>,</th>
<th>je</th>
<th>ne</th>
<th>veux</th>
<th>plus</th>
<th>jouer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>do</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>want</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>play</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Och and Ney, 2004]
Hierarchical Phrase Extraction

[Chiang, 2007]

No, I do not want to play anymore.
Discontinuous Phrase Extraction

Non | , | je | ne | veux | plus | jouer
---|---|---|---|---|---|---
No | ☐ | ☐ | ☐ | ☐ | ☐ | ☐
, | ☐ | ☐ | ☐ | ☐ | ☐ | ☐
I  | ☐ | ☐ | ☐ | ☐ | ☐ | ☐
do | ☐ | ☐ | ☐ | ☐ | ☐ | ☐
not| ☐ | ☐ | ☐ | ☐ | ☐ | ☐
want| ☐ | ☐ | ☐ | ☐ | ☐ | ☐
to | ☐ | ☐ | ☐ | ☐ | ☐ | ☐
play| ☐ | ☐ | ☐ | ☐ | ☐ | ☐
anymore | ☐ | ☐ | ☐ | ☐ | ☐ | ☐

Consistency wrt. word alignment is the only constraint.

Simpler than hierarchical phrase extraction.
**Problem:** enumerate all phrases (continuous or not) that appear both in the input **test** sentence and the **training** data.

**Approach:** find collocated **sub-patterns** [Rahman et al., 2006], exploiting algorithmic speed-ups [Lopez, 2007]

<table>
<thead>
<tr>
<th>Query pattern:</th>
<th>ne ... pas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-patterns:</td>
<td>ne</td>
</tr>
<tr>
<td></td>
<td>pas</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ne sais pas . tu ne sais plus . #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ne sais pas . tu ne sais plus . #</td>
</tr>
<tr>
<td></td>
<td>ne sais plus . #</td>
</tr>
<tr>
<td></td>
<td>nuit. je ne sais pas . tu ne sais plus</td>
</tr>
<tr>
<td></td>
<td>pas . tu ne sais plus . #</td>
</tr>
<tr>
<td></td>
<td>pas une seule nuit. je ne sais pas . plus . #</td>
</tr>
<tr>
<td></td>
<td>sais pas . tu ne sais plus . #</td>
</tr>
</tbody>
</table>
Sample extracted phrases

Continuous phrase-based models: [Och and Ney, 2004]

je
jouer
ne veux plus jouer
je ne veux plus jouer

I
to play
do not want to play anymore
I do not want to play anymore

Context-free models (Hiero) add: [Chiang, 2005]

ne veux plus \(X\)
je ne veux plus \(X\)

do not want \(X\) anymore
I do not want \(X\) anymore

Discontinuous models add:

veux
dO ... want
ne ... plus
not ... anymore
je ne ... plus
I ... not ... anymore
veux ... jouer
do ... want to play

(5 other phrase pairs)
Discontinuous Phrase-based SMT

Step II: Decoding
Phrase-Based Decoding

- **Contiguous case**: (previous lecture)

- **Discontinuous case**: same components:
  - Hypothesis stacks
  - Future cost estimation (dynamic programming)
  - Hypothesis recombination
  - Distortion limit
Discontinuous Phrase-Based Decoding

X: 他说 正在 为 这 一 访问 作出 安排。

he said are to this visit to arrangements
he said that are ... for this one access arrangements ... made
he said are for this visit made arrangement
for this visit

y: he said
c:
x: oo------oo
score:-1.3

y: arrangements
c: made
x: 000000
score:-3.2

y: are
c: made
x: 000000
score:-4.8

y: made
c: for this
x: 000000
score:-6.1

y: for this
c:
x: 000000
score:-7.2

y: visit
c:
x: 00000000
score:-8.5
he said arrangements are made for this visit.

Discontinuous Phrase-Based Decoding

X: 他说 正在 为 这一 访问 作出 安排。
Contiguous phrase-based SMT:

在这种情况 下，当 生命权 被 剥夺 时，...

under such circumstances, when the right was deprived of, can...

Discontinuous phrase-based SMT:

在这种情况 下，当 生命权 被 剥夺 时，...

under such circumstances, when he was deprived of the right to life, it...
中组部要求进一步加大对领导干部和干部选拔任用工作的监督力度。

COD has called for further increase of supervision and the intensity on leading cadres.

Discontinuous phrase-based SMT:

COD has called for further increase of supervision of leading cadres and...
Longer phrases at test time with discontinuous phrases
Accurate long-distance reordering (with phrase-based SMT)
Drawbacks of phrase-Based SMT

How phrase-based SMT work in practice (actual output):

- **Poor lexical substitutions**: long phrases of the *training* data are often not usable at *test* time, and short phrases yield poor translations. Phrase-based SMT is particularly vulnerable to data sparseness.

- **Poor word order**: phrase-based word reordering models lack rich analyses (e.g., syntactic) to guide the translation process.
Long-distance reordering in phrase-based SMT: how?

- Ideally, would use syntactic analyses to guide word re-ordering (but don’t have these analyzes in phrase-based SMT).
- Will present simple and effective substitute.
The Russian side hopes to hold consultations with Iran on this issue in the near future.

- Phrase reordering turned into a sequence classification problem.
- 3 classes: Monotone, Swap, and Other
The Russian side hopes to hold consultations with Iran on this issue in the near future.

- Phrase reordering turned into a sequence classification problem.
- 3 classes: Monotone, Swap, and Other
the russian side hopes to hold consultations with iran on this issue in the near future

• Phrase reordering turned into a sequence classification problem.

• 3 classes: Monotone, Swap, and Other
the russian side hopes to hold consultations with iran on this issue.

Source: [8]
Stack: [1-3][9-12][6-7]
Operation: reduce
Orientation: M
The Russian side hopes to hold consultations with Iran on this issue in the near future.

Source: [4-5]
Stack: [1-3][9-12][6-7][8]
the Russian side hopes to hold consultations with Iran on this issue in the near future.

Source: [4-5]
Stack: [1-3][9-12][6-8]
the russian side hopes to hold consultations with Iran on this issue in the near future. The Russian side hopes to hold consultations with Iran on this issue in the near future.

This type of deterministic shift-reduce parsing is proved to be linear time [Galley and Manning, 2008; Huang et al., 2009].
## Results

<table>
<thead>
<tr>
<th></th>
<th>Chinese-English</th>
<th>Arabic-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-based SMT</td>
<td>31.08</td>
<td>50.02</td>
</tr>
<tr>
<td>phrase-based SMT</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>with standard lexicalized reordering</em> (Moses)</td>
<td>32.16</td>
<td>50.87</td>
</tr>
<tr>
<td>discontinuous phrase-based SMT</td>
<td>32.93</td>
<td>51.47</td>
</tr>
<tr>
<td><em>with standard lexicalized reordering</em> (Moses)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>discontinuous phrase-based SMT</td>
<td>33.75 (+1.59)</td>
<td>51.85 (+0.98)</td>
</tr>
<tr>
<td><em>with long-range lexicalized reordering</em> [Galley and Manning; EMNLP 2008]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**BLEU [%]**
More information in:

• Accurate Non-Hierarchical Phrase-Based Translation
  Michel Galley and Christopher Manning
  In Proc. of NAACL-HLT, 2010
• A Simple and Effective Hierarchical Reordering Model.

Implementation of this work is part of Phrasal,
Stanford's open-source SMT system:

http://nlp.stanford.edu/software/software/phrasal
Backup slides
Running time: phrase-based running time is practically $O(n)$
Limited empirical adequacy of context-free translation models

[Søgaard and Kuhn, 2009]

<table>
<thead>
<tr>
<th>Structure</th>
<th>Hand alignments</th>
<th>Language</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>cross-serial</td>
<td>a b c a b</td>
<td>French–English</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>d e f d e f g</td>
<td>Spanish–English</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portuguese–English</td>
<td>9.5</td>
</tr>
<tr>
<td>inside-out</td>
<td>a b c d</td>
<td>French–English</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>d e f g</td>
<td>Spanish–English</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portuguese–English</td>
<td>7.0</td>
</tr>
<tr>
<td>context-free</td>
<td>a b c</td>
<td>French–English</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>d e f</td>
<td>Spanish–English</td>
<td>81.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portuguese–English</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Similar findings in [Wellington et al., 2006; Søgaard and Wu, 2009]
What's wrong with context-free translation models?

Discontinuous phrases:
- ne ... pas
- not
- sais
- do ... know

Hierarchical phrases:
- ne $X_1$ pas
- not $X_2$
- $X_2$ sais $X_2$
- do $X_1$ know

Neither SCFG nor STAG can handle this example [Søgaard, 2008]
Linear distortion (Phrasal vs. Moses)

- Phrasal linear distortion penalizes gaps:

\[ d(s_2) = 3 \quad d(s_1) = 1 \quad d(s_2) = 2 \quad d(s_3) = 3 \quad d(s_2) = 2 \quad d(s_4) = 1 \]

Lexically-sensitive reordering model

- Phrasal and Moses use lexical orientation models [Tillmann, 2004]
- Hierarchical translation models are lexically sensible by default [Chiang, 2007]

\[ p(\text{swap} | \text{世界上 in the world}) = 0.9 \]
\[ p(x \rightarrow \text{世界上} \quad x \text{ in the world}) = 0.9 \]
Experimental Setup

Joshua [Li et al., 2009]
• State-of-the-art hierarchical MT system.

Moses [Koehn et al., 2007]
• Standard phrase-based MT system.

Phrasal [Cer, Galley, Jurafsky, Manning; NAACL 2010 demo]
• Stanford's open-source phrase-based MT system.
• Supports **discontinuous phrases.**

Notes:
• *All systems use default features and settings (Moses’ == Phrasal's).*
• *With no discontinuous phrases, Phrasal produces the same output as Moses (to enable meaningful comparison).*
Training (same in all systems)

Method
- minimum error-rate training [Och, 2003]

Features

Joshua and Moses' standard features:
- language model
- phrase translation probabilities (relative freq.)
- lexical translation probabilities (relative freq.)
- structural features (word penalty, phrase penalty, etc.)
- linear distortion
Experimental Setup: Data

Chinese-English TM
- *Train:* 23.3M Chinese words, 28M English words (LDC)
- *Test:* MT03-MT08

Arabic-English TM
- *Train:* 18.8M Arabic words, 19.5M English words (LDC)
- *Test:* MT03 & MT05

English LM
- 5-gram LM with modified Kneser-Ney smoothing.
- Trained on Gigaword, AFP and Xinhua sections.
### Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Chinese-English</th>
<th>Arabic-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>hierarchical MT (baseline)</td>
<td>31.90</td>
<td>50.31</td>
</tr>
<tr>
<td>phrase-based MT</td>
<td>31.08</td>
<td>50.02</td>
</tr>
<tr>
<td>phrase-based MT with lexicalized reordering [Tillmann, 2004]</td>
<td>32.16</td>
<td>50.87</td>
</tr>
<tr>
<td>discontinuous phrase-based MT with lexicalized reordering [Tillmann, 2004]</td>
<td>32.93</td>
<td>51.47</td>
</tr>
<tr>
<td>discontinuous phrase-based MT with long-range lexicalized reordering [Galley and Manning; EMNLP 2008]</td>
<td>33.75 (+1.85)</td>
<td>51.85 (+1.54)</td>
</tr>
</tbody>
</table>

**BLEU [%]**  
(NIST Chinese-English 2003-2008  
NIST Arabic-English 2003 and 2005)
Conclusion

same experimental setting

discontinuous phrases | simple generative model features | MERT training data

two different decoders

TREE-BASED DECODER (Joshua)

STRING-BASED DECODER (this work)

contributing factors

BLEU

31.90

33.75 (+1.85)

• experimental technology (2005 – Present)
• context-freeness
• large search space (hypergraphs)

• mature technology (1990s – Present)
• no assumptions about data
• only scoring strings
Rescue workers search for survivors in collapsed houses.
How to decode with syntactic translation rules?

救援人员在倒塌的房屋里寻找生还者。

[rescue]    [staff]     [in]  [collapse]  [of]    [house] [in] [search]    [survivors]

- Convert syntax rules into a CFG and decode in $O(n^3)$ with CKY.
- Parse Chinese input as an English sentence!