
2004

Sixth Biennial Conference of the
Association for Machine Translation in the Americas

Tutorial Notes

# Introduction to Statistical Machine Translation

**Philipp Koehn**
MIT / CSAIL
**Kevin Knight**
USC / ISI & Computer Science Dept.

***September 28, 2004***
***Georgetown University***
***Washington, DC***

# Introduction to Statistical Machine Translation

Philipp Koehn
CSAIL
Massachusetts Institute of Technology

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



## Overview

- Introduction
- Data for Statistical MT
- MT Evaluation
- Word-Based Statistical MT
- Phrase-Based Statistical MT
- Advanced Training Methods
- Syntax and Semantics in Statistical MT

# Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

⇨

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 .

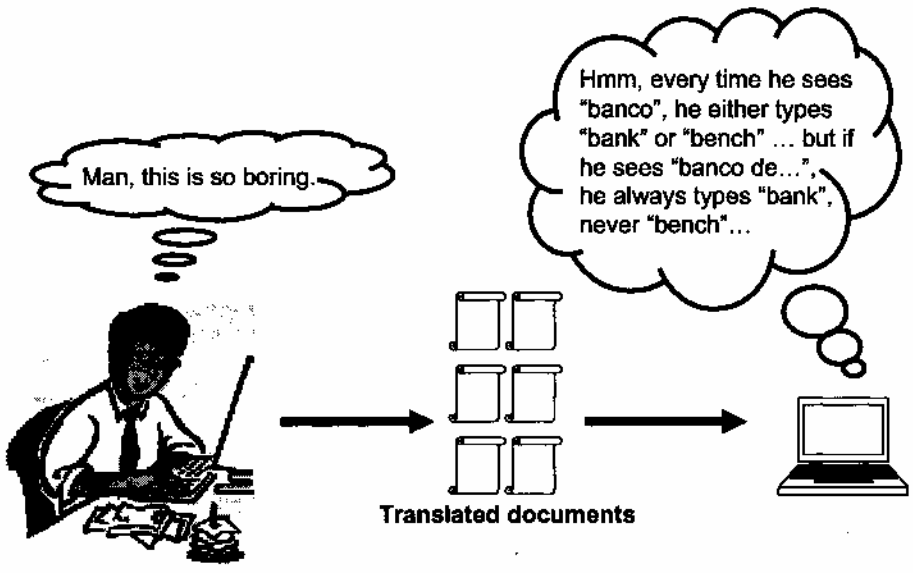
The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

About $10 billion spent annually on human translation.

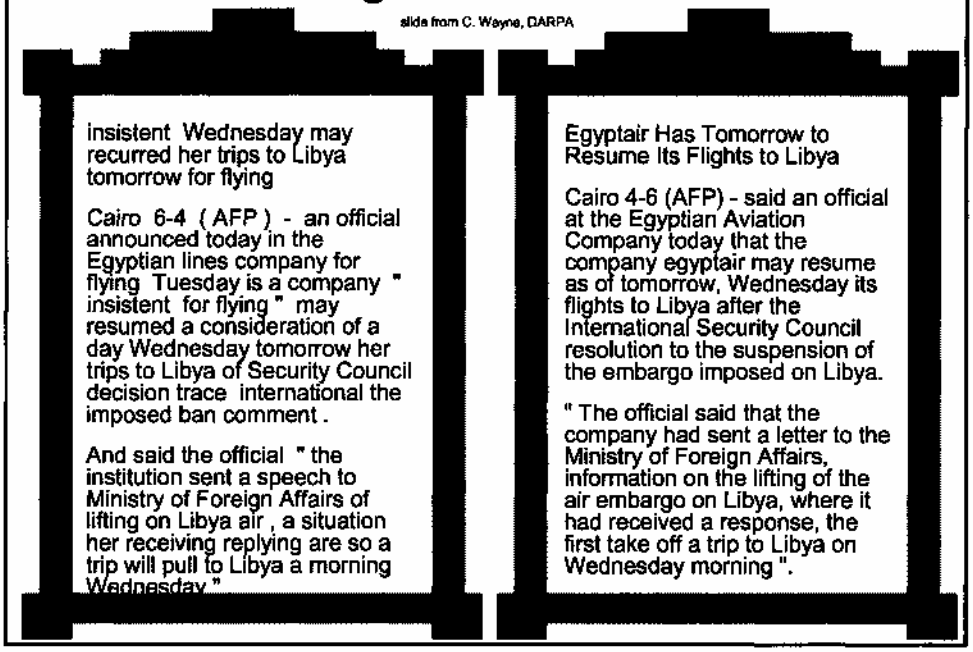
People around the world stubbornly refuse to write everything in English.

# Data-Driven Machine Translation



# Recent Progress in Statistical MT

slide from C. Wayne, DARPA




# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok **farok** ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok **farok** izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok **farok** ororok lalok sprok izok enemok .<br>7b. wat **jjat** bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok **farok** izok stok .<br>5b. totat **jjat** quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok **farok** ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok **farok** izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok **crrrok** hihok yorok zanzanok .<br>???<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok **hihok** ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok **hihok** yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok **hihok** mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok **hihok** ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok **yorok** ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok **hihok** **yorok** zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok **hihok** mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok **clok** .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok **clok** . ???<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . process of elimination |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . cognate? |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

# Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

| | |
|---|---|
| 1a. ok-voon ororok sprok .<br>1b. at-voon bichat dat . | 7a. lalok farok ororok lalok sprok izok enemok .<br>7b. wat jjat bichat wat dat vat eneat . |
| 2a. ok-drubel ok-voon anok plok sprok .<br>2b. at-drubel at-voon pippat rrat dat . | 8a. lalok brok anok plok nok .<br>8b. iat lat pippat rrat nnat . |
| 3a. erok sprok izok hihok ghirok .<br>3b. totat dat arrat vat hilat . | 9a. wiwok nok izok kantok ok-yurp .<br>9b. totat nnat quat oloat at-yurp . |
| 4a. ok-voon anok drok brok jok .<br>4b. at-voon krat pippat sat lat . | 10a. lalok mok nok yorok ghirok clok .<br>10b. wat nnat gat mat bat hilat . |
| 5a. wiwok farok izok stok .<br>5b. totat jjat quat cat . | 11a. lalok nok crrrok hihok yorok zanzanok .<br>11b. wat nnat arrat mat zanzanat . zero fertility |
| 6a. lalok sprok izok jok stok .<br>6b. wat dat krat quat cat . | 12a. lalok rarok nok izok hihok mok .<br>12b. wat nnat forat arrat vat gat . |

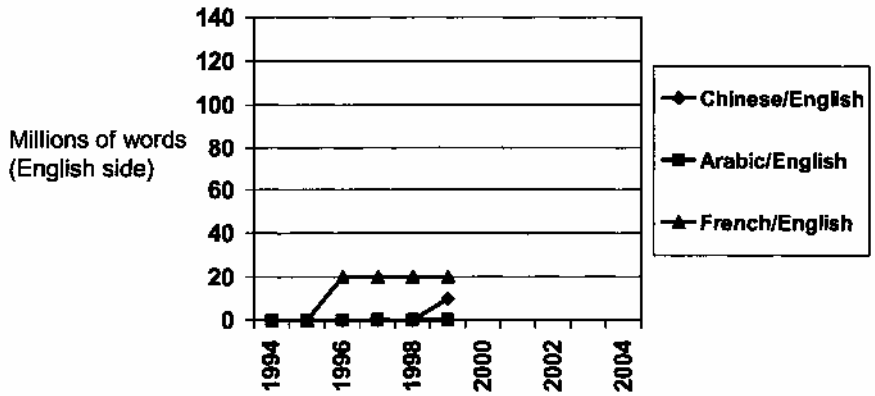
# It's Really Spanish/English

**Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa**

| | |
|---|---|
| 1a. Garcia and associates .<br>1b. Garcia y asociados . | 7a. the clients and the associates are enemies .<br>7b. los clients y los asociados son enemigos . |
| 2a. Carlos Garcia has three associates .<br>2b. Carlos Garcia tiene tres asociados . | 8a. the company has three groups .<br>8b. la empresa tiene tres grupos . |
| 3a. his associates are not strong .<br>3b. sus asociados no son fuertes . | 9a. its groups are in Europe .<br>9b. sus grupos estan en Europa . |
| 4a. Garcia has a company also .<br>4b. Garcia tambien tiene una empresa . | 10a. the modern groups sell strong pharmaceuticals .<br>10b. los grupos modernos venden medicinas fuertes . |
| 5a. its clients are angry .<br>5b. sus clientes estan enfadados . | 11a. the groups do not sell zenzanine .<br>11b. los grupos no venden zanzanina . |
| 6a. the associates are also angry .<br>6b. los asociados tambien estan enfadados . | 12a. the small groups are not modern .<br>12b. los grupos pequenos no son modernos . |

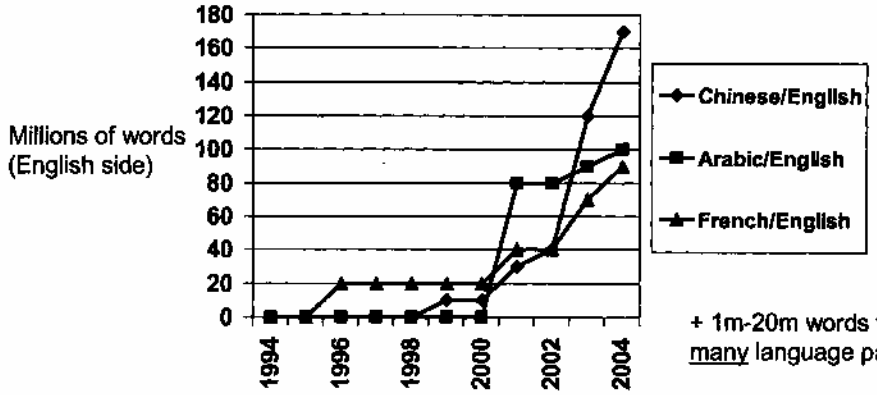
- Introduction
- **Data for Statistical MT and Data Preparation**
- MT Evaluation
- Word-Based Statistical MT
- Phrase-Based Statistical MT
- Advanced Training Methods
- Syntax and Semantics in Statistical MT

## Ready-to-Use Online Bilingual Data




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

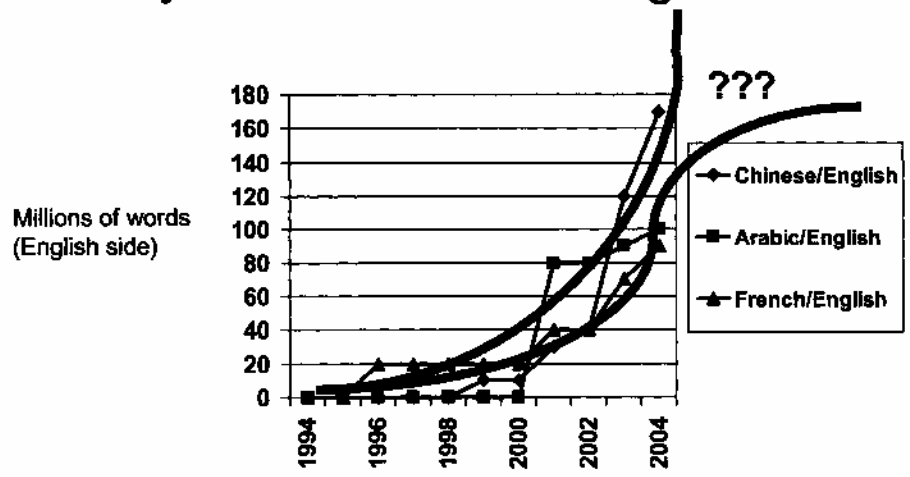
# Ready-to-Use Online Bilingual Data




+ 1m-20m words for many language pairs

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

# Ready-to-Use Online Bilingual Data




→ One Billion?

# Chinese/English News

3.5m words of human-translated text
(training data available from LDC)

| 5655 | 904 | 69258 | 7039 | 855 | 14852 | 1174 | 63 | 200 | 46 | 92428 |
|---|---|---|---|---|---|---|---|---|---|---|
| c1 | c2 | c3 | c4 | c5 | c6 | c7 | c8 | c9 | c10 | c11 |

Previously unseen 11-word
Chinese sentence
from a news article
(test data available from NIST)

Frequency of word c8
in human-translated text

# Chinese/English News

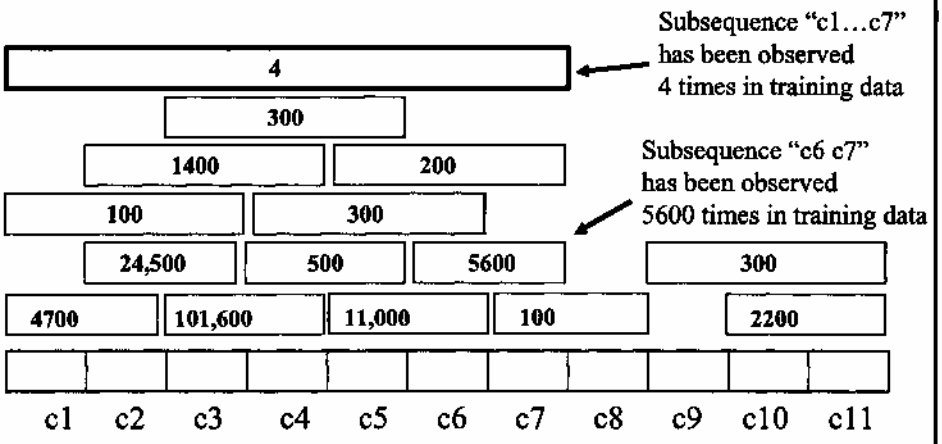
3.5m words of human-translated text
(training data available from LDC)

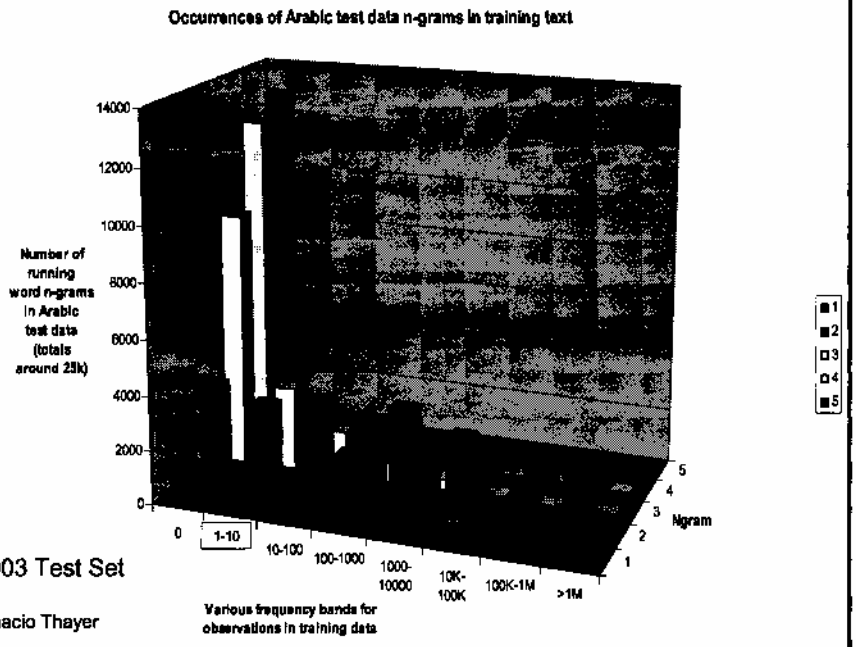
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 3 | | | | | |
| | | | | 14 | | 2 | | | | |
| | | | 1 | | 3 | | | | | |
| | | | 245 | | 5 | 56 | | | 3 | |
| | | 47 | | 1016 | | 110 | 1 | | | 22 |
| 5655 | 904 | 69258 | 7039 | 855 | 14852 | 1174 | 63 | 200 | 46 | 92428 |
| c1 | c2 | c3 | c4 | c5 | c6 | c7 | c8 | c9 | c10 | c11 |

Subsequence "c6 c7"
has been observed
56 times in training data

# Future Data?

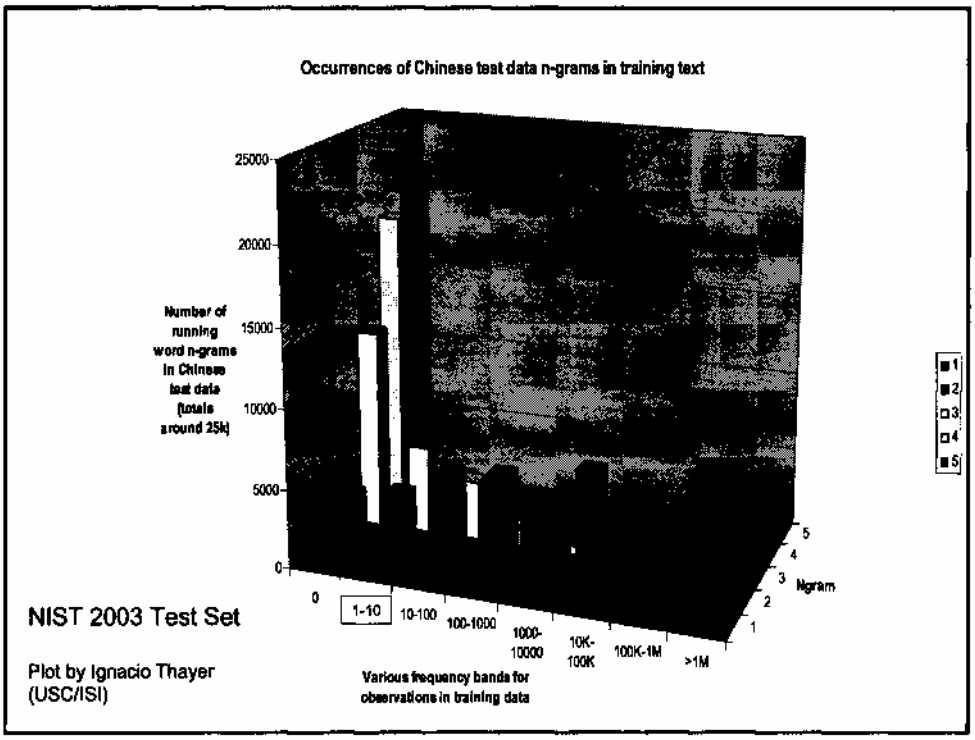
One billion words of human-translated training text?



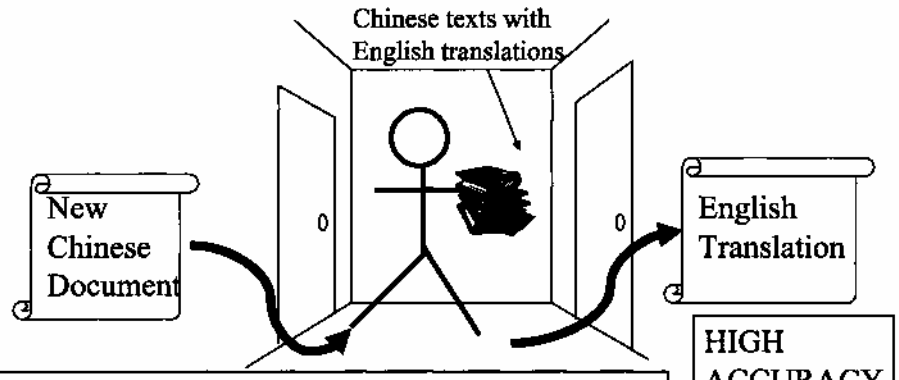




NIST 2003 Test Set

Plot by Ignacio Thayer (USC/ISI)




# Welcome to the Chinese Room



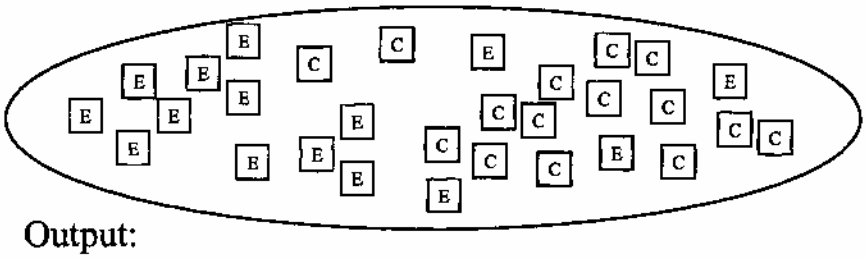

You can teach yourself to translate Chinese using *only* bilingual data (without grammar books, dictionaries, any people to answer your questions…)

# From No Data to Sentence Pairs

- Easy way: Linguistic Data Consortium (LDC)
- Really hard way: pay $$$
  - Suppose one billion words of parallel data were sufficient
  - At 20 cents/word, that's $200 million
- Pretty hard way: Find it, and then earn it!
  - De-formatting
  - Remove strange characters
  - Character code conversion
  - **Document alignment**
  - **Sentence alignment**
  - **Tokenization (also called Segmentation)**

# Document Alignment

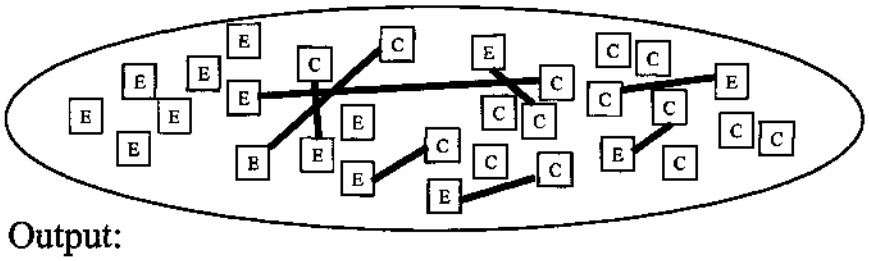
- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.




- Output:
  - List of pairs of files that are actually translations.

# Document Alignment

- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.




- Output:
  - List of pairs of files that are actually translations.

# Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

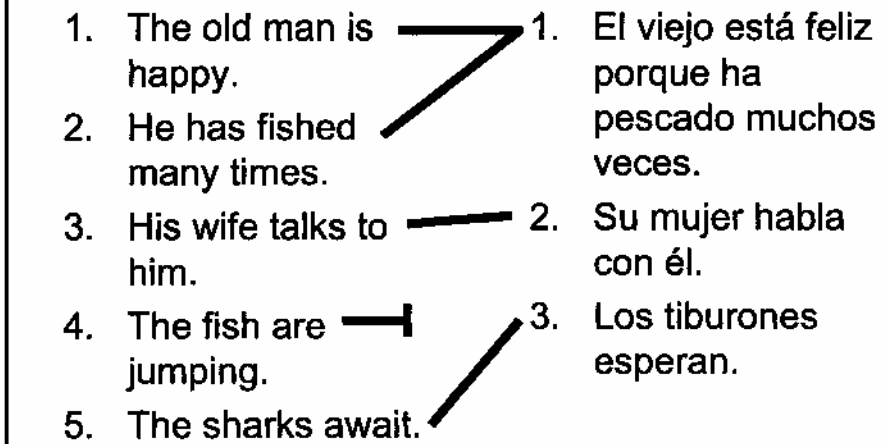
El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

# Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. *The sharks await.*

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

# Sentence Alignment



# Sentence Alignment

| | | |
|---|---|---|
| 1. The old man is happy. He has fished many times. | —— | 1. El viejo está feliz porque ha pescado muchos veces. |
| 2. His wife talks to him. | —— | 2. Su mujer habla con él. |
| 3. The sharks await. | —— | 3. Los tiburones esperan. |

Note that unaligned sentences are thrown out, and sentences are merged in n-to-m alignments (n, m > 0).

# Tokenization (or Segmentation)

- English
  - Input (some byte stream):

    `"There," said Bob.`
  - Output (7 "tokens" or "words"):

    `" There , " said Bob .`
- Chinese
  - Input (byte stream): 美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
  - Output: 美国 关岛国 际机 场 及其 办公室均接获 一名 自称 沙地 阿拉 伯富 商拉登 等发 出 的 电子邮件。

# Lower-Casing

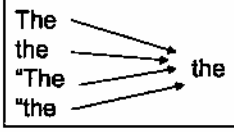
- English
  - Input (7 words):

```
" There , " said Bob .
```

  - Output (7 words):

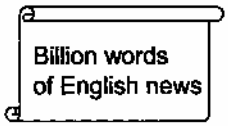
```
" there , " said bob .
```

**Idea of tokenizing and lower-casing:**

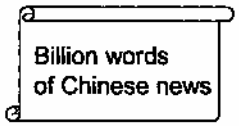



Smaller vocabulary size.
More robust counting and learning.

# What About Comparable Corpora?




A




B

**Which sentences from A are translations of which sentences from B?**

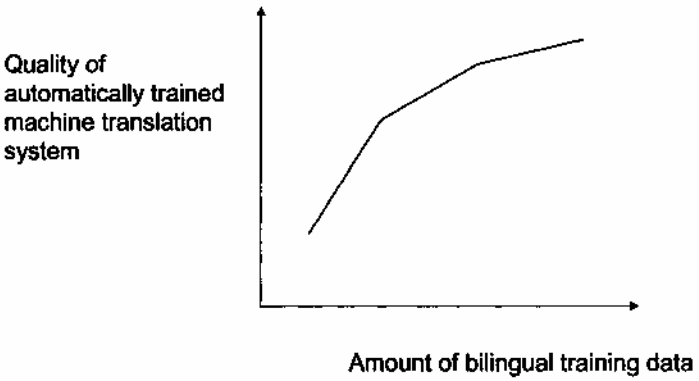
**Which words from A are translations of which words from B?**

> **"If x and y are translations, the translations of words that hang around x in corpus A may be found hanging around y in corpus B..."**

**See:**

**[Munteanu, Fraser & Marcu, 2004; Fung & Cheung, 2004; Fung & McKeown, 1995; Rapp, 1999; Koehn & Knight, 2000; Zhou & Vogel, 2002; Yamada & Hideki, 2002; Masao & Hitoshi, 2002]**

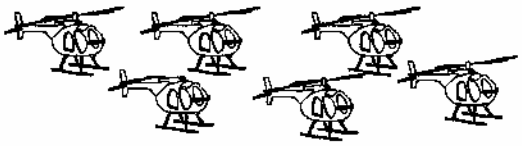
# It Is Possible to Draw Learning Curves: How Much Data Do We Need?




- Introduction
- Data for Statistical MT
- **MT Evaluation**
- Word-Based Statistical MT
- Phrase-Based Statistical MT
- Advanced Training Methods
- Syntax and Semantics in SMT Models

# MT Evaluation

- Source only
- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization
- Objective usage testing



- Automatic:
  - WER (word error rate)
  - BLEU (Bilingual Evaluation Understudy)
  - NIST
  - Named-Entity

# BLEU Evaluation Metric

(Papineni et al, ACL-2002)

**Reference (human) translation:**
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 .

**Machine translation:**
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the")
- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

# BLEU Evaluation Metric

(Papineni et al, ACL-2002)

**Reference (human) translation:**
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 .

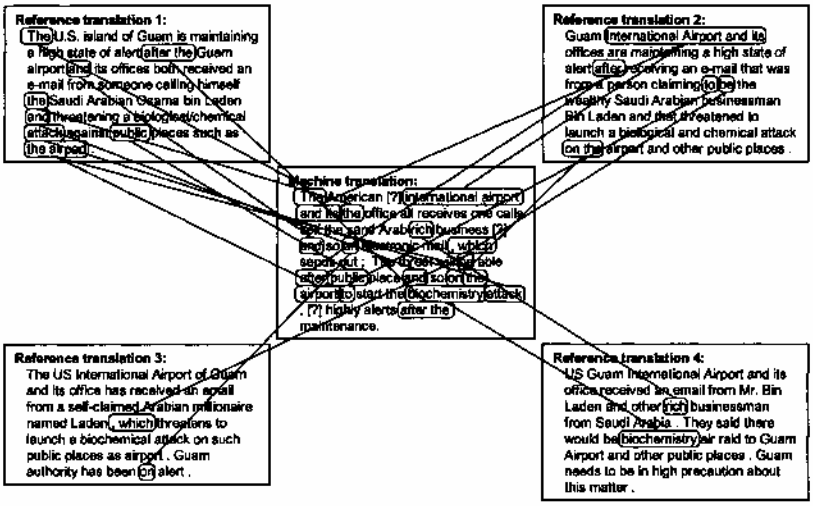
**Machine translation:**
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

# Multiple Reference Translations



# DARPA MT Evaluation Corpus: 11 Human Translations of 100 Chinese News Articles

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.

Last week 's fight took at least 12 lives.

The fighting last week killed at least 12.

The battle of last week killed at least 12 persons.

At least 12 people lost their lives in last week 's fighting.

At least 12 persons died in the fighting last week.

At least 12 died in the battle last week.

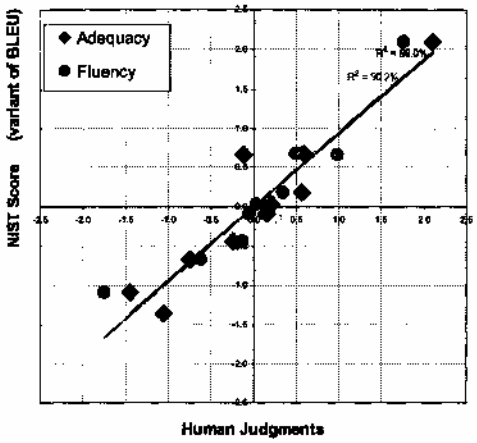
At least 12 people were killed in the fighting last week.

During last week 's fighting , at least 12 people died.

Last week at least twelve people died in the fighting.

Last week 's fighting took the lives of twelve people.

# BLEU Tends to Predict Human Judgments




slide from G. Doddington (NIST)

# BLEU in Action

| | |
|---|---|
| 枪手被警方击毙。 | (Foreign Original) |
| the gunman was shot to death by the police . | (Reference Translation) |
| the gunman was police kill . | #1 |
| wounded police jaya of | #2 |
| the gunman was shot dead by the police . | #3 |
| the gunman arrested by police kill . | #4 |
| the gunmen were killed . | #5 |
| the gunman was shot to death by the police . | #6 |
| gunmen were killed by police ?SUB>0 ?SUB>0 | #7 |
| al by the police . | #8 |
| the ringer is killed by the police . | #9 |
| police killed the gunman . | #10 |

# BLEU in Action

| | |
|---|---|
| 枪手被警方击毙。 | (Foreign Original) |
| the gunman was shot to death by the police . | (Reference Translation) |
| the gunman was police kill . | #1 |
| wounded police jaya of | #2 |
| the gunman was shot dead by the police . | #3 |
| the gunman arrested by police kill . | #4 |
| the gunmen were killed . | #5 |
| the gunman was shot to death by the police . | #6 |
| gunmen were killed by police ?SUB>0 ?SUB>0 | #7 |
| al by the police . | #8 |
| the ringer is killed by the police . | #9 |
| police killed the gunman . | #10 |

| | | |
|---|---|---|
| **green** | = 4-gram match | (good!) |
| **red** | = word not matched | (bad!) |

# BLEU in Action

| | | |
|---|---|---|
| **枪手被警方击毙。** | | **(Foreign Original)** |
| **the gunman was shot to death by the police .** | | **(Reference Translation)** |
| **the gunman was police kill .** | **#1** | Machine |
| **wounded police jaya of** | **#2** | Machine |
| **the gunman was shot dead by the police .** | **#3** | **Human** |
| **the gunman arrested by police kill .** | **#4** | Machine |
| **the gunmen were killed .** | **#5** | Machine |
| **the gunman was shot to death by the police .** | **#6** | **Human** |
| **gunmen were killed by police ?SUB>0 ?SUB>0** | **#7** | Machine |
| **al by the police .** | **#8** | Machine |
| **the ringer is killed by the police .** | **#9** | Machine |
| **police killed the gunman .** | **#10** | **Human** |

| | | |
|---|---|---|
| **green** | = 4-gram match | (good!) |
| **red** | = word not matched | (bad!) |

# BLEU: Problems?

- Not currently very sensitive to global syntactic structure.
  - though you can dispute that ...
- Doesn't care if an incorrectly translated word is a name or a preposition
  - *gave it to Albright* (reference)
  - *gave it at Albright* (translation #1)
  - *gave it to altar* (translation #2)
- What happens when a program reaches human level performance in BLEU but the translations are still bad?
  - maybe sooner than you think ...

# BLEU: Prospects

- As MT advances, MTE will fail to distinguish between humans and machines
- So MTE must advance to be able to distinguish better…
- So MT must advance…
- So MTE must advance…
- A benign arms race?
  - of course, there is way more data for developing MT (bilingual text) than for developing MTE (human judgments) …

# 11 Human Translation Agencies Employed to Translate 100 Chinese News Articles

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.

Last week 's fight took at least 12 lives.

The fighting last week killed at least 12.

The battle of last week killed at least 12 persons.

At least 12 people lost their lives in last week 's fighting.

At least 12 persons died in the fighting last week.

At least 12 died in the battle last week.

At least 12 people were killed in the fighting last week.

During last week 's fighting , at least 12 people died.

Last week at least twelve people died in the fighting.

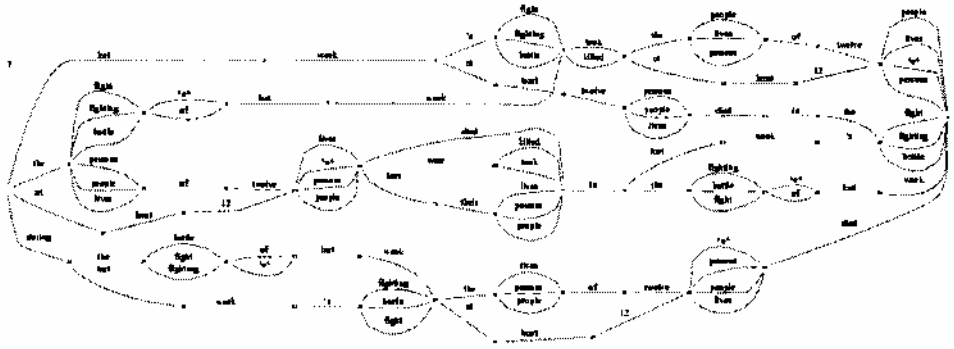
Last week 's fighting took the lives of twelve people.

# Merging Translations

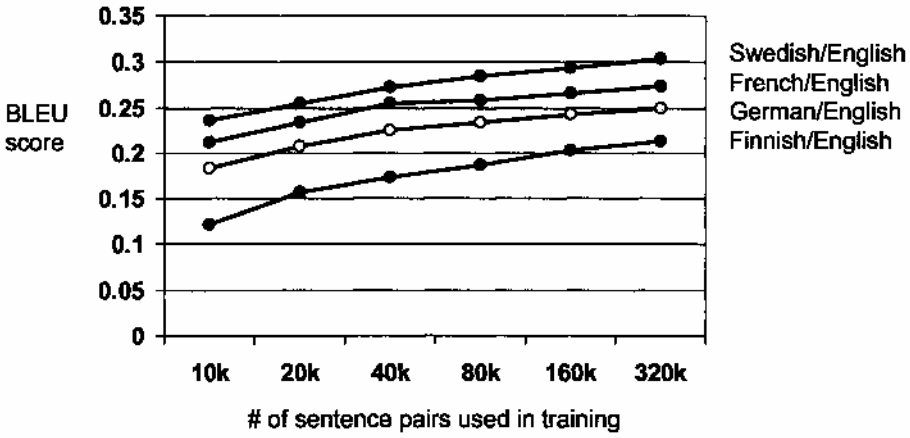
(Pang, Knight, and Marcu, NAACL-HLT 2003)

Create word graphs by merging paraphrases
=> from 10 sentences to over a thousand

11th human translation is often found in the graph!



# Sample Learning Curves

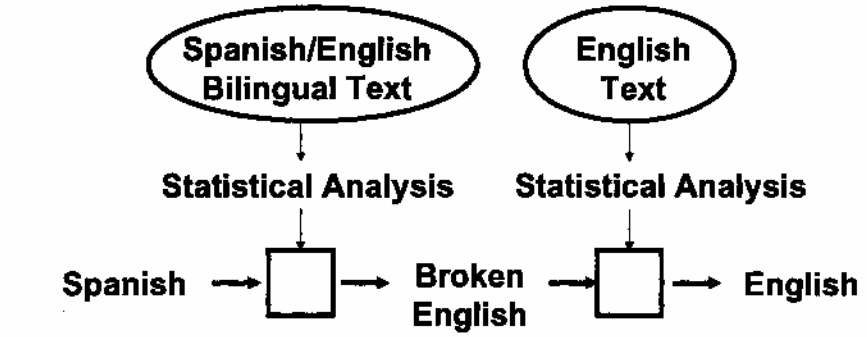



Experiments by
Philipp Koehn

# Common MT Evaluations

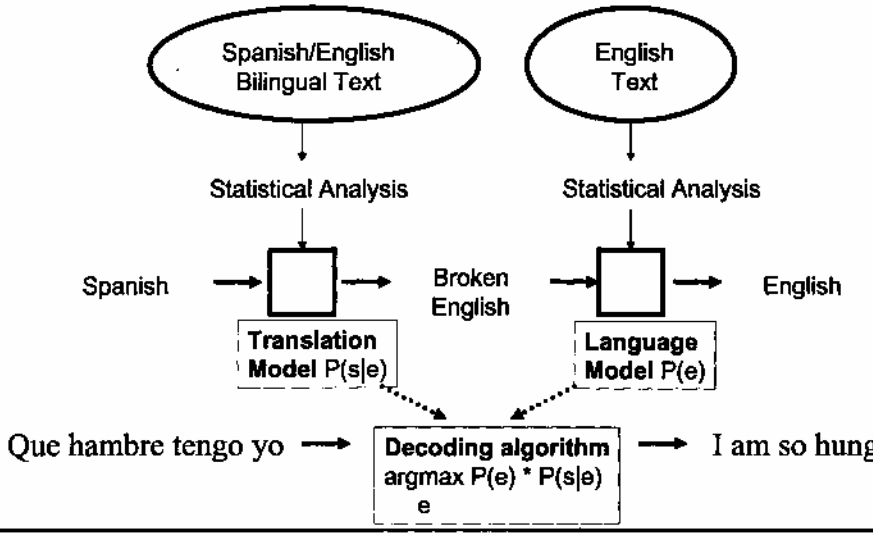
- New common evaluations administered by NIST
  - www.nist.gov/speech/tests/mt/index.htm
- First evaluation held June, 2002
  - 7 participants
  - Scores returned immediately by automatic email BLEU/NIST scoring server
- Second evaluation held May, 2003
  - 10 participants
- Third evaluation held May, 2004
  - 17 participants

# Statistical MT Systems

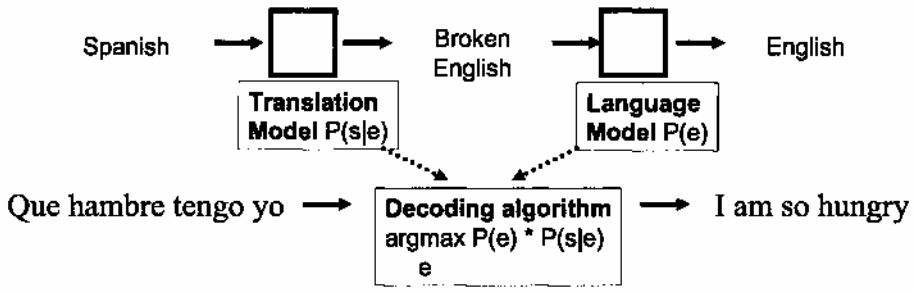



Que hambre tengo yo → What hunger have I, Hungry I am so, I am so hungry, Have I that hunger … → I am so hungry

# Statistical MT Systems



# Bayes Rule




Given a source sentence s, the decoder should consider many possible translations … and return the target string e that maximizes

**P(e | s)**

By Bayes Rule, we can also write this as:

**P(e) x P(s | e) / P(s)**

and maximize that instead. P(s) never changes while we compare different e's, so we can equivalently maximize this:

**P(e) x P(s | e)**

# Three Problems for Statistical MT

- Language model
  - Given an English string e, assigns P(e) by formula
  - good English string -> high P(e)
  - random word sequence -> low P(e)
- Translation model
  - Given a pair of strings <f,e>, assigns P(f | e) by formula
  - <f,e> look like translations -> high P(f | e)
  - <f,e> don't look like translations -> low P(f | e)
- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)

# The Classic Language Model

## Word N-Grams

Goal of the language model:

He is on the soccer field
He is in the soccer field

Is table the on cup the
The cup is on the table

Rice shrine
American shrine
Rice company
American company

# The Classic Language Model

## Word N-Grams

Generative story:

```
w1 = START
repeat until END is generated:
    produce word w2 according to a big table P(w2 | w1)
    w1 := w2
```

P(I saw water on the table) =

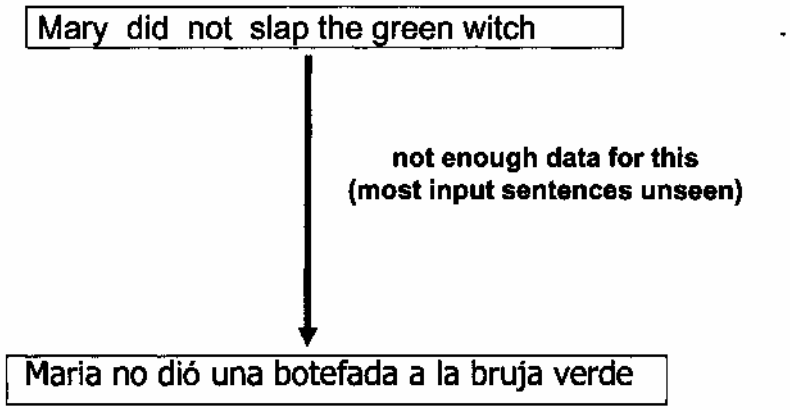
P(I | START) *
P(saw | I) *
P(water | saw) *
P(on | water) *
P(the | on) *
P(table | the) *
P(END | table)

**Probabilities can be learned from online English text.**

# Translation Model

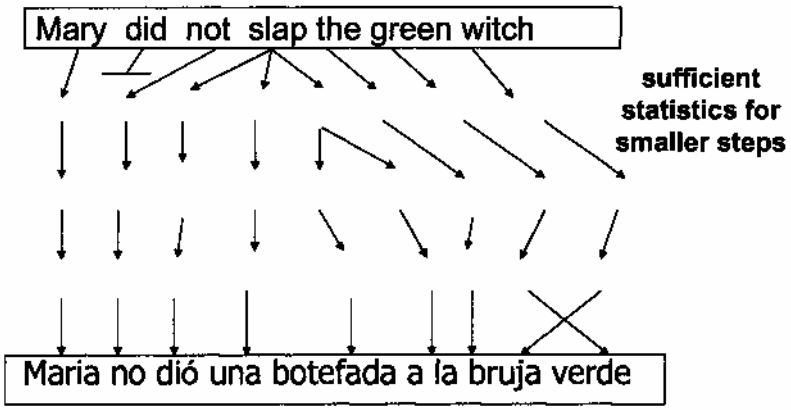
Learn How to Translate from Data

**Direct Estimation:**



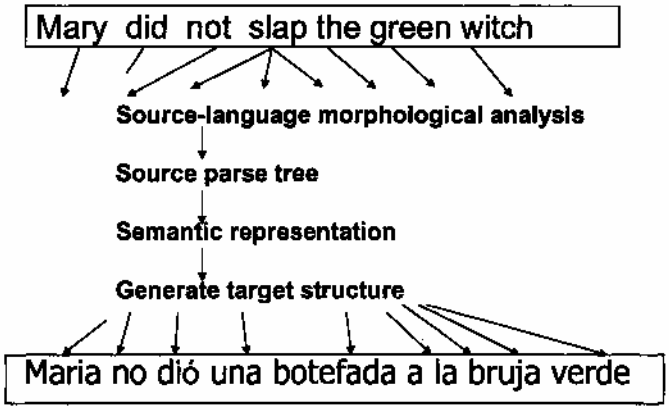

# Generative Model

**Break up process into smaller steps:**



# What kind of Translation Model?

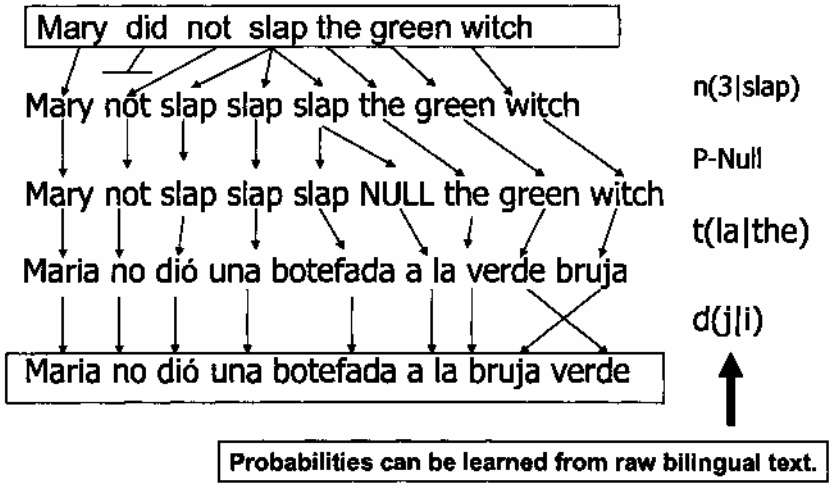
**May use syntactic and semantic representations:**




# The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

**Generative story:**




**Probabilities can be learned from raw bilingual text.**

## Probabilistic Method

- Generative story for how an English string e gets to be a French string f
  - Choices in the story are decided by reference to parameters -- e.g., t(juste | correct)
- Formula for P(f | e) in terms of parameters
  - Usually long and hairy, but mechanical to extract from the story
- Training to obtain parameter estimates from possibly-incomplete data
  - Off-the-shelf EM ("Expectation Maximization")

## How do we Learn the Parameters?

- Incomplete training data:
  - we know input and output words
  - we do not know the connections ("alignments") between them
- Chicken and egg problem:
  - if we knew the alignments, we could estimate the model
  - if we knew the model, we could find the best alignments

# Finding Word Alignments

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

All word alignments equally likely

All P(french-word | english-word) equally likely

# Finding Word Alignments

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

"la" and "the" observed to co-occur frequently,
so P(la | the) is increased.

# Finding Word Alignments

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

"house" co-occurs with both "la" and "maison", but P(maison | house) can be raised without limit, to 1.0, while P(la | house) is limited because of "the"

(pigeonhole principle)

# Finding Word Alignments

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

settling down after another iteration

# Finding Word Alignments

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

**Inherent hidden structure revealed by EM training!**

For details, see

- "A Statistical MT Tutorial Workbook" (Knight, 1999).
  - 37 easy sections, final section promises a free beer.
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

# Statistical Machine Translation

… la maison … la maison bleue … la fleur …

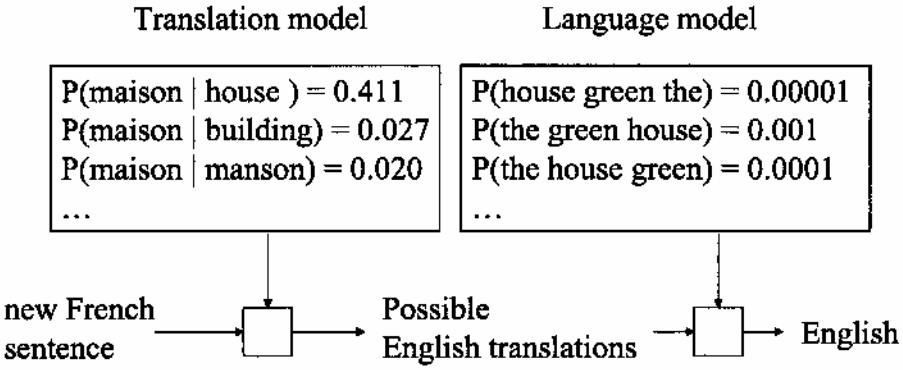
… the house … the blue house … the flower …

| | |
|---|---|
| P(maison \| house ) = 0.411 | |
| P(maison \| building) = 0.027 | |
| P(maison \| manson) = 0.020 | |
| … | |

Estimating the model from training data

# Statistical Machine Translation

Applying the model to new input

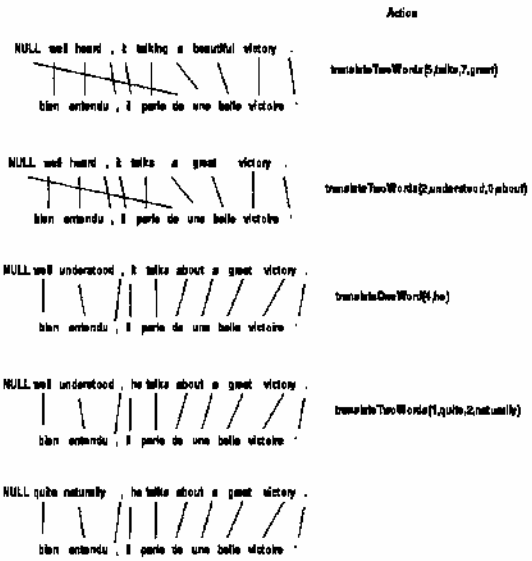



# Decoding for "Classic" Models

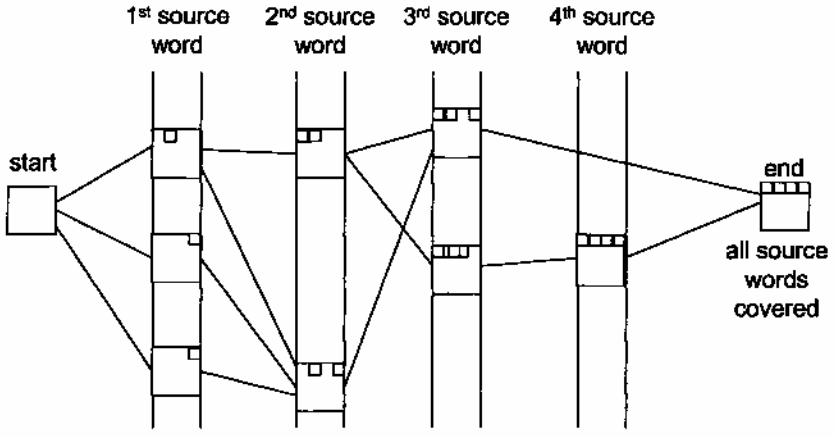
- Of all conceivable English word strings, find the one maximizing P(e) x P(f | e)
- Decoding is an NP-complete challenge
  - (Knight, 1999)
- Several search strategies are available
- Each potential English output is called a *hypothesis*.

# Greedy decoding

(Germann et al, ACL-2001)



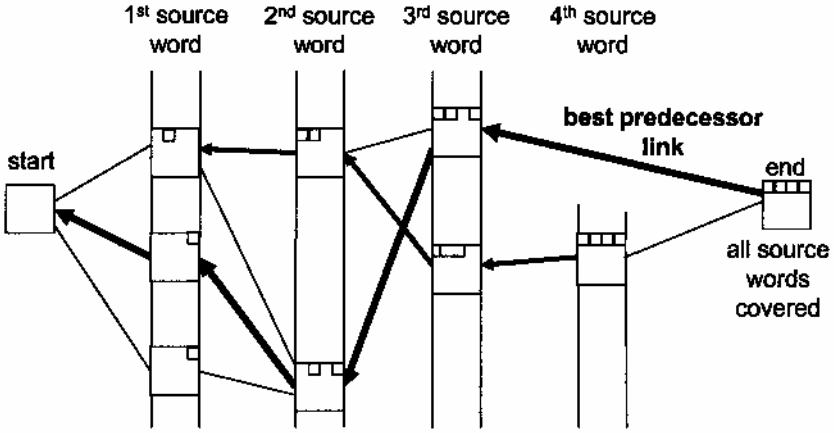

# Dynamic Programming Beam Search




Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

# Dynamic Programming Beam Search




Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

# The Art of Beam Search

- Coalesce hypotheses with the same future behavior, throw away weaker ones
  - Cover same source words
  - End in same two target words
- When pruning, use hypotheses' estimated-future-cost as well as cost-so-far
- Make multiple passes to get estimated-future-cost

# Post-Processing

- If an MT system translates to English, then…

  We have to change this kind of output:

  ```
  " there , " said bob .
  ```

  into something that looks better:

  ```
  "There," said Bob.
  ```

- For English output, this involves *de-tokenization* and *re-capitalization* (latter is nontrivial!).
- For Chinese output, we need to remove the spaces between the words.
- For German output, etc., etc.
- In general: need to reverse the pre-processing.

# The Classic Results

| | |
|---|---|
| • *la politique de la haine .* | *(Foreign Original)* |
| • politics of hate . | (Reference Translation) |
| • the policy of the hatred . | (IBM4+N-grams+Stack) |
| • *nous avons signé le protocole .* | *(Foreign Original)* |
| • we did sign the memorandum of agreement . | (Reference Translation) |
| • we have signed the protocol . | (IBM4+N-grams+Stack) |
| • *où était le plan solide ?* | *(Foreign Original)* |
| • but where was the solid plan ? | (Reference Translation) |
| • where was the economic base ? | (IBM4+N-grams+Stack) |

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

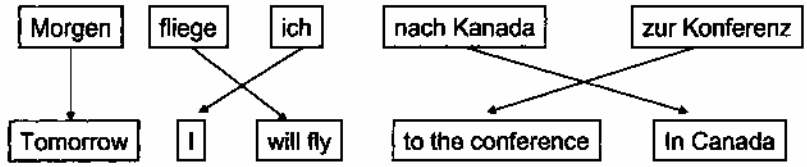
the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

# Flaws of Word-Based MT

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - "real estate", "note that", "interest in"
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

- **Introduction**
- **Data for Statistical MT**
- **MT Evaluation**
- **Word-Based Statistical MT**
- **Phrase-Based Statistical MT**
- **Advanced Training Methods**
- **Syntax and Semantics in Statistical MT**

# Phrase-Based Statistical MT




- Foreign input segmented in to phrases
  - "phrase" is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered
- See [Koehn et al, 2003] for an intro.

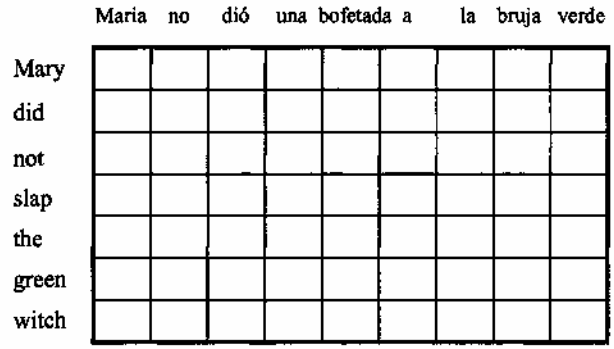
**This is state-of-the-art!**

# Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguating
  - "Interest rate" → …
  - "Interest in" → …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

# How to Learn the Phrase Translation Table?

- One method: "alignment templates" (Och et al, 1999)
- Start with word alignment, build phrases from that.

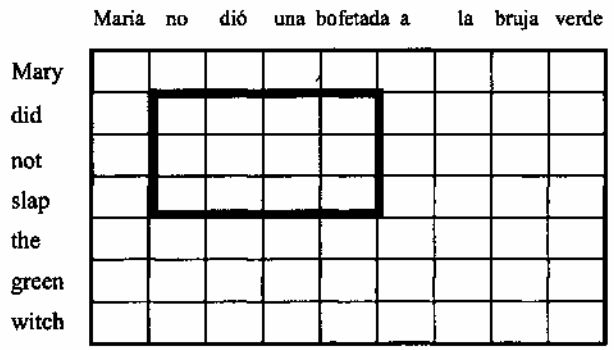



This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or "Viterbi") alignment.

# How to Learn the Phrase Translation Table?

- One method: "alignment templates" (Och et al, 1999)
- Start with word alignment, build phrases from that.

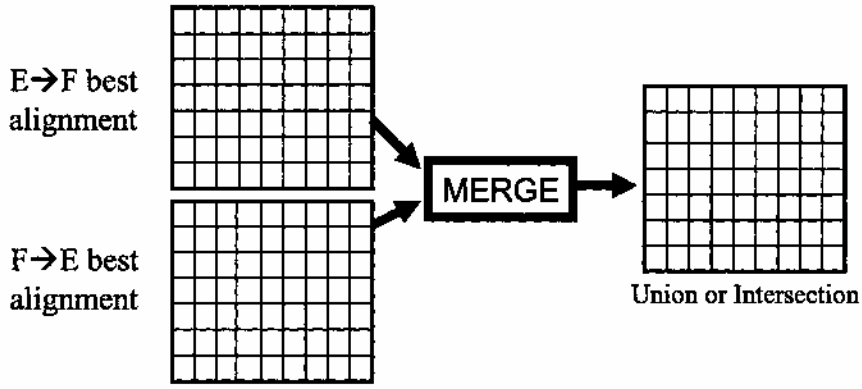



This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or "Viterbi") alignment.

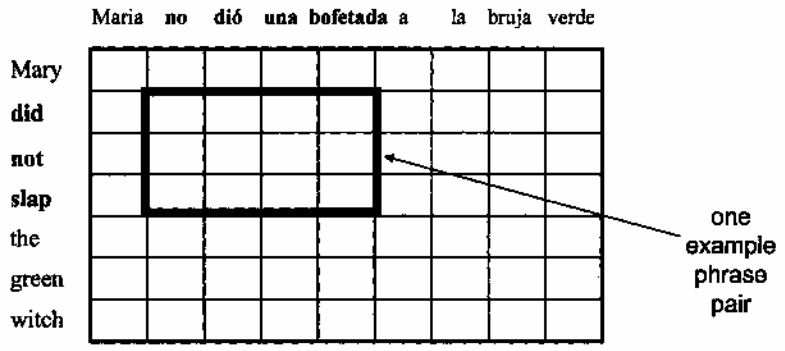
# IBM Models are 1-to-Many

- Run IBM-style aligner both directions, then merge:

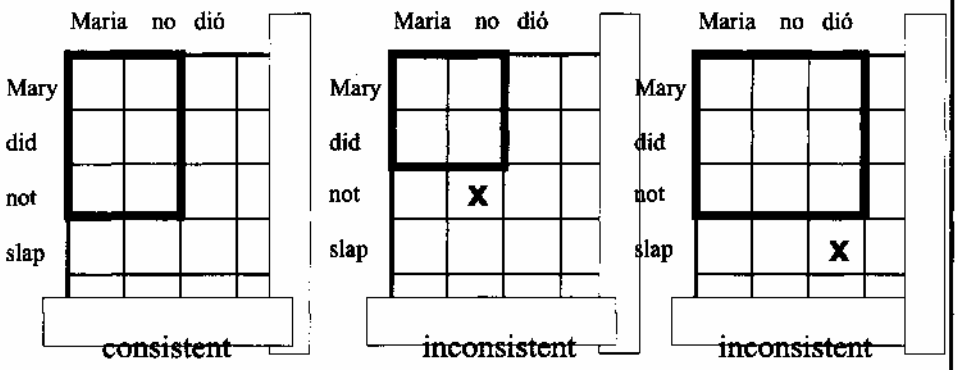



# How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*

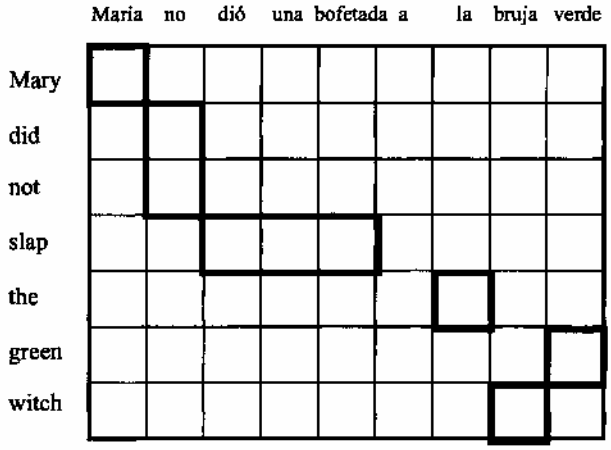


# Consistent with Word Alignment



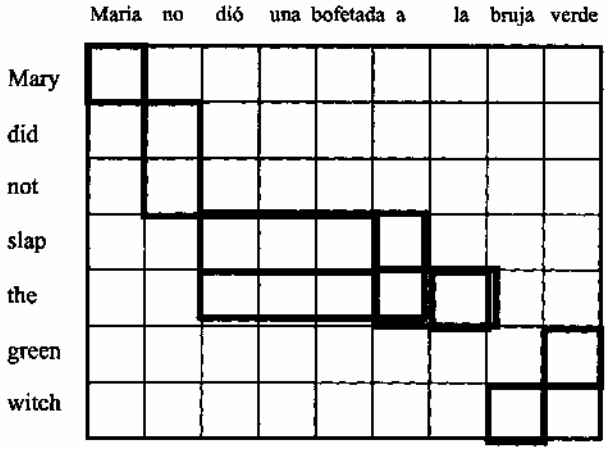

Phrase alignment must contain all alignment points for all the words in both phrases!

# Word Alignment Induced Phrases



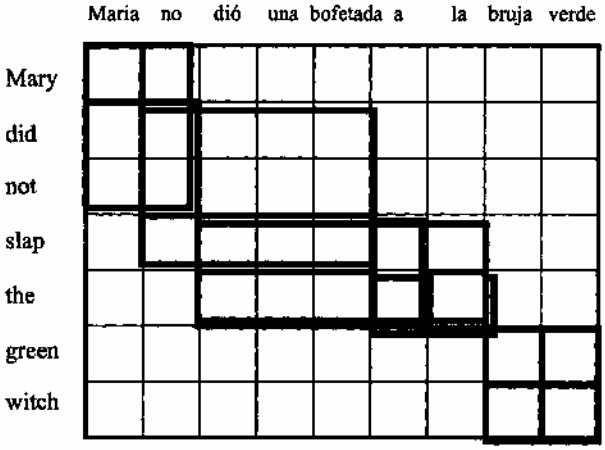

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

# Word Alignment Induced Phrases



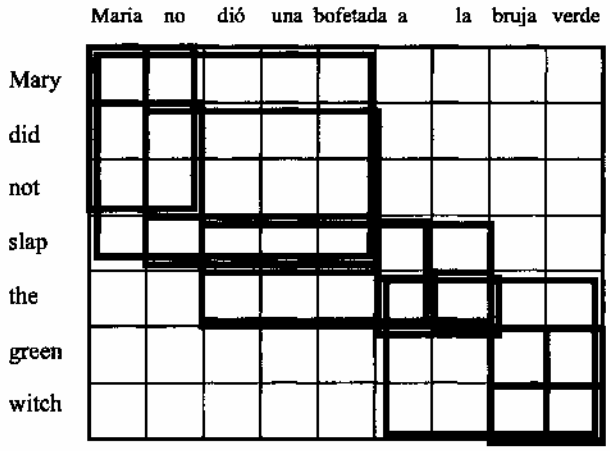

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)

# Word Alignment Induced Phrases



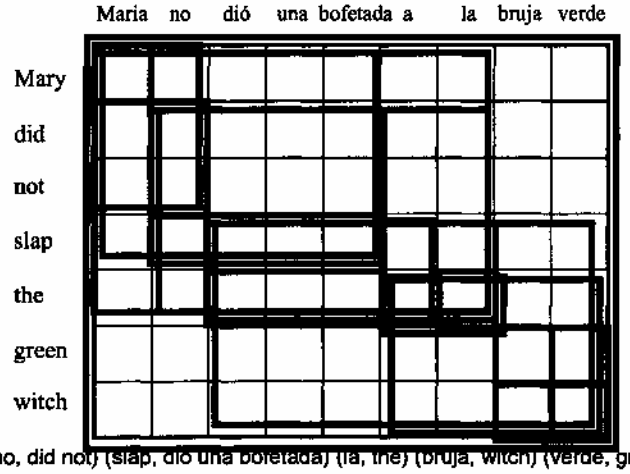

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)
(bruja verde, green witch)

# Word Alignment Induced Phrases




(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)
(bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap)
(a la bruja verde, the green witch) ...

# Word Alignment Induced Phrases




(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)
(bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap)
(a la bruja verde, the green witch) ...
(Maria no dió una bofetada a la bruja verde, Mary did not slap the green witch)

# Phrase Pair Probabilities

- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.
  - We hope so!
- So, now we have a vast list of phrase pairs and their frequencies – how to assign probabilities?

# Phrase Pair Probabilities

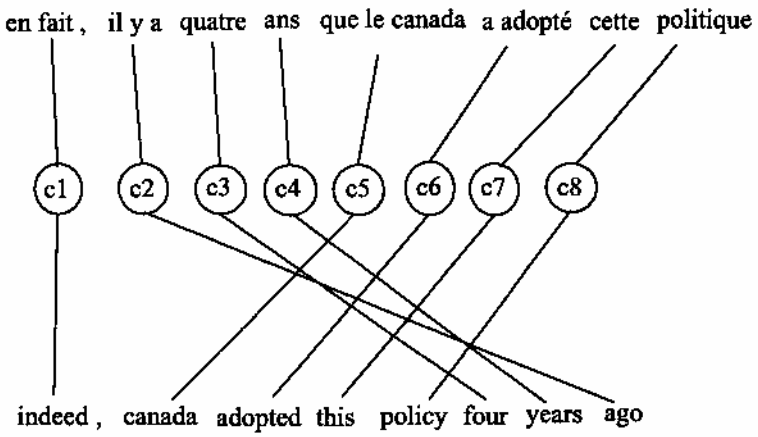
- Basic idea:
  - No EM training
  - Just relative frequency:

    P(f-f-f | e-e-e) = count(f-f-f, e-e-e) / count(e-e-e)
- Important refinements:
  - Smooth using word probs P(f | e) for individual words connected in the word alignment
    - Some low count phrase pairs now have high probability, others have low probability
  - Discount for ambiguity
    - If phrase e-e-e can map to 5 different French phrases, due to the ambiguity of unaligned words, each pair gets a 1/5 count
  - Count BAD events too
    - If phrase e-e-e doesn't map onto *any* contiguous French phrase, increment event count(BAD, e-e-e)

# Phrase Pair Probabilities

- Lots of variations on this theme:
  - [Tillmann, 2003]
  - [Zens and Ney, 2004]
  - [Venugopal et al, 2003]
  - [Zhang et al, 2003]
  - [Marcu and Wong, 2002]

# Joint Phrase Model (Marcu and Wong, 2002)
## EM-based trained model



# Phrase Table Learned Examples

- possessives:
  - ( la réaction de : 's response ) :5.30995e-06
  - ( la réponse de : 's response ) :5.29405e-06
- complex nominals:
  - ( monsieur le président : mr. speaker ) :4.07544e-05
- negation:
  - ( ne est pas : is not ) :0.000173962
  - ( est inadmissible : is not good enough ) :2.31636e-06
  - ( ne est pas ici : is not here ) :3.57434e-06
- adj-noun order:
  - ( fièvre typhoïde : paratyphoid fever ) :1.77204e-06
- paraphrase:
  - ( fonction de les plus importantes : most important responsibility ) :1.77e-06

# Phrase-Based Decoding

- Monotonic version:
  - Substitute phrases for phrases, left to right
  - Word order can change within a phrase, but phrases themselves don't change order
  - Allows a dynamic programming solution
  - Monotonic assumption sometimes does not severely damage BLEU score (surprisingly)
    - For Arabic/Chinese → English, about 3-4%
- Non-monotonic version
  - Explore re-ordering of phrases themselves

- Introduction
- Data for Statistical MT
- MT Evaluation
- Word-Based Statistical MT
- Phrase-Based Statistical MT
- **Advanced Training Methods**
- Syntax and Semantics in SMT Models

## Basic Model, Revisited

$$\text{argmax}_{e} \; P(e \mid f) =$$

$$\text{argmax}_{e} \; P(e) \times P(f \mid e) / P(f) =$$

$$\text{argmax}_{e} \; P(e) \times P(f \mid e)$$

## Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\ P(e) \times P(f \mid e) / P(f) =$$

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e)$$ … works better!

## Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\ P(e) \times P(f \mid e) / P(f)$$

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1}$$

Rewards longer hypotheses, since these are unfairly punished by P(e)

# Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times KS^{3.7} \ldots$$

Lots of knowledge sources vote on any given hypothesis.

"Knowledge source" = "feature function" = "score component".

Feature function simply scores a hypothesis with a real value.

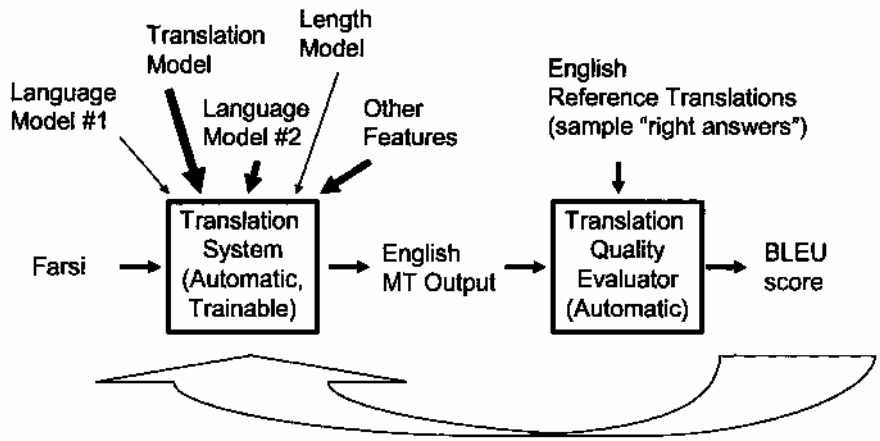
(May be binary, as in "e has a verb").

**Problem: How to set the exponent weights?**

# Discriminative Training (Och & Ney, 2002)

- Basic Idea:
  - Set parameters to get the right translation, rather than to maximize the likelihood of the data
- Application:
  - Learn handful of weights for language model, phrase-translation model, word-translation-model, length model, re-ordering model.
  - Don't try to set (the millions of) word and phrase translation probabilities this way.

# Maximum BLEU Training (Och, 2003)




Learning Algorithm for Directly Reducing Translation Error
Yields big improvements in quality.

# Learn Model Weights

1. Start with initial weights
2. Generate list of n-best translations for each source sentence in a development set
3. Score all those translations with BLEU
4. Figure out which model weights push the best-BLEU translations to the top
5. Change model weights
6. Go to Step 2

# Discriminative Training

Find model weights that make the correct translations score best.
Methods vary depending on how they implement "find", "correct", and "score best."

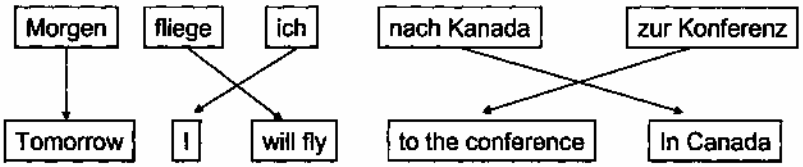
| | LM | TM | LEN | Correct |
|---|---|---|---|---|
| 1 Mary not give slap witch green . | -17.2 | -5.2 | -7 | no |
| 2 Mary not slap the witch green . | -16.3 | -5.7 | -7 | no |
| 3 Mary not give slap of the green witch . | -18.1 | -4.9 | -9 | no |
| 4 Mary not give of green witch . | -16.5 | -5.1 | -8 | no |
| 5 Mary did not slap the witch green . | -20.1 | -4.7 | -8 | no |
| 6 Mary did not slap green witch . | -15.5 | -3.2 | -7 | no |
| 7 Mary not slap of the witch green . | -19.2 | -5.3 | -8 | no |
| 8 Mary did not give slap of witch green . | -23.2 | -5.0 | -9 | no |
| 9 Mary did not give slap of the green witch . | -21.8 | -4.4 | -10 | no |
| 10 Mary did slap the witch green . | -15.5 | -6.9 | -7 | no |
| **11 Mary did not slap the green witch .** | **-17.4** | **-5.3** | **-8** | **yes** |
| 12 Mary did slap witch green . | -16.9 | -6.9 | -6 | no |
| 13 Mary did slap the green witch . | -14.3 | -7.1 | -7 | no |
| 14 Mary did not slap the of green witch . | -24.2 | -5.3 | -9 | no |
| 15 Mary did not give slap the witch green . | -25.2 | -5.5 | -9 | no |



# Discriminative Training

- Example work in machine translation:
  - [Och and Ney, 2002]
  - [Och, 2003]
  - [Shen et al, 2004]
- Also becoming popular to use Simplex algorithm to optimize model weights.

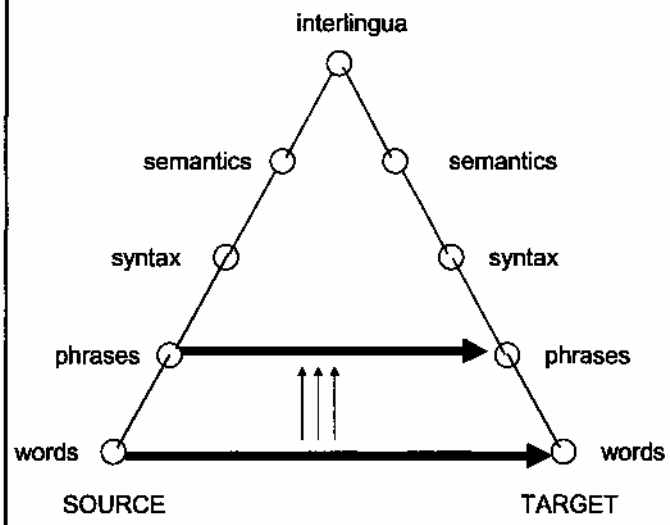
# Summary: Phrase-Based MT




This is state-of-the-art!

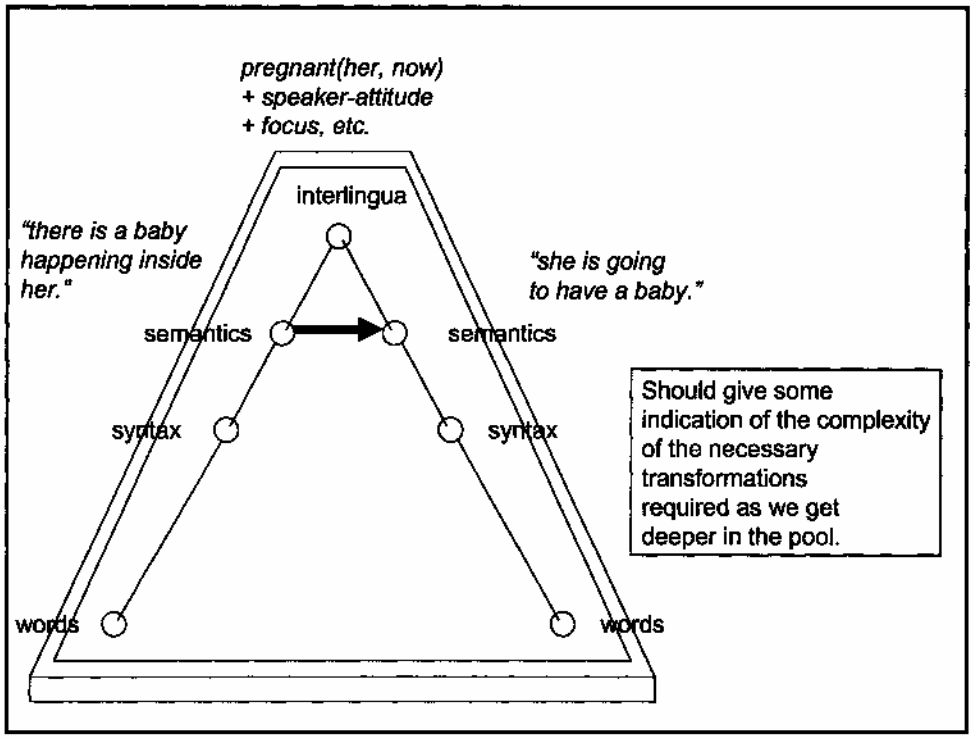
- **Introduction**
- **Data for Statistical MT**
- **MT Evaluation**
- **Word-Based Statistical MT**
- **Phrase-Based Statistical MT**
- **Advanced Training Methods**
- **Syntax and Semantics in Statistical MT**

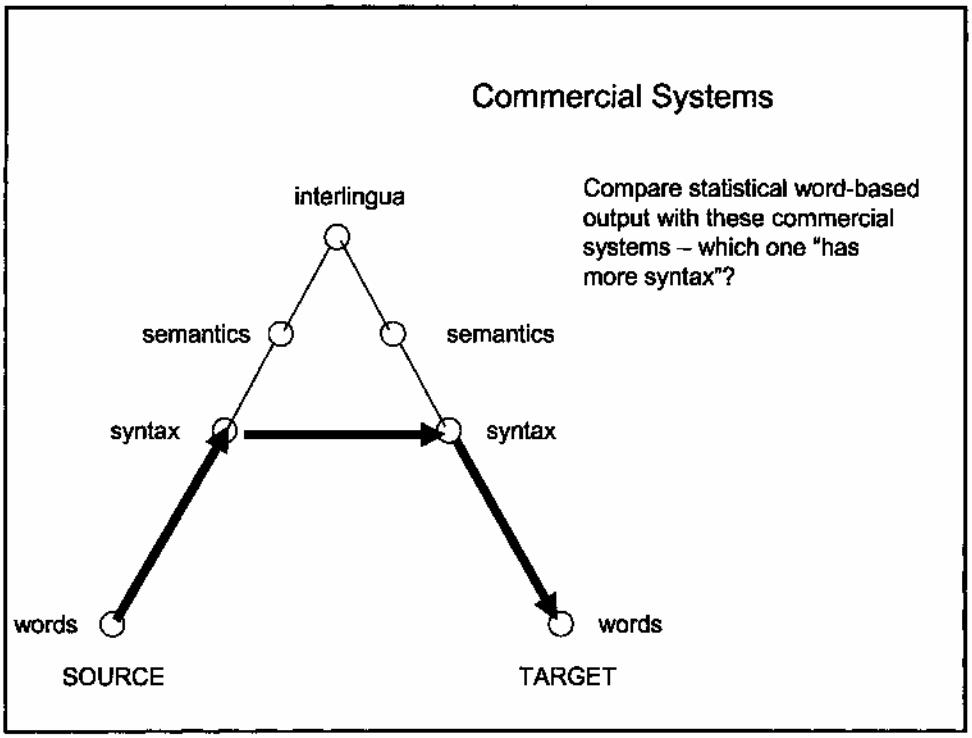
# MT Pyramid



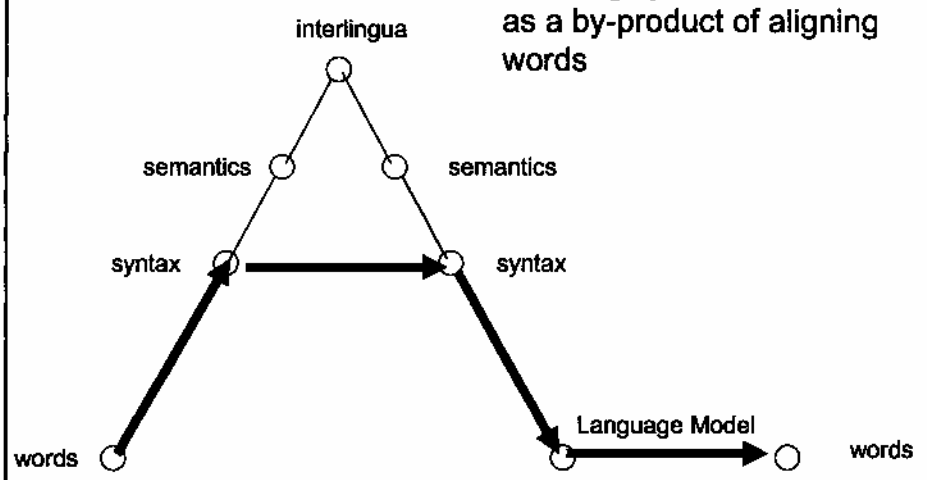

# Why Syntax?

- Need much more grammatical output
- Need accurate control over re-ordering
- Need accurate insertion of function words
- Word translations need to depend on grammatically-related words

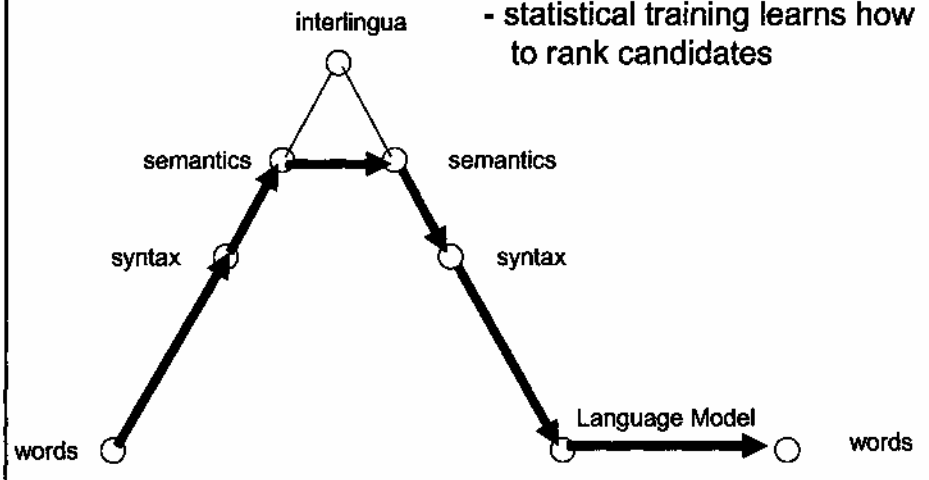

pregnant(her, now)
+ speaker-attitude
+ focus, etc.
interlingua
"there is a baby happening inside her."
"she is going to have a baby."
semantics
semantics
syntax
syntax
words
words
Should give some indication of the complexity of the necessary transformations required as we get deeper in the pool.

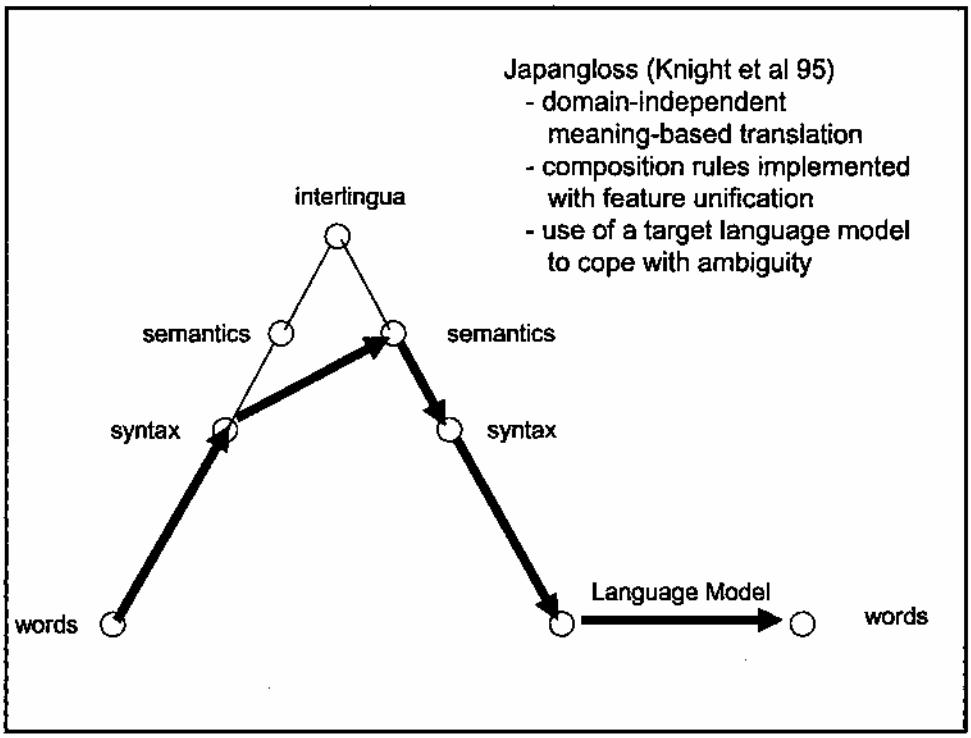

Commercial Systems
interlingua
Compare statistical word-based output with these commercial systems – which one "has more syntax"?
semantics
semantics
syntax
syntax
words
words
SOURCE
TARGET

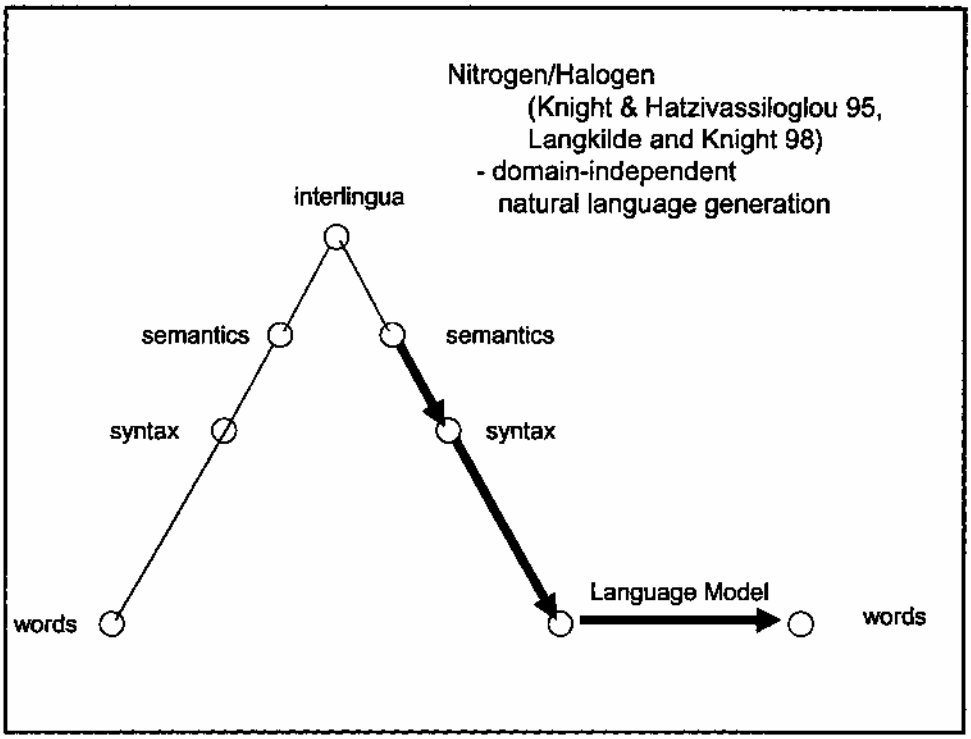
Wu (97), Alshawi et al (98)
- learning from string pairs
- inducing syntactic structure as a by-product of aligning words

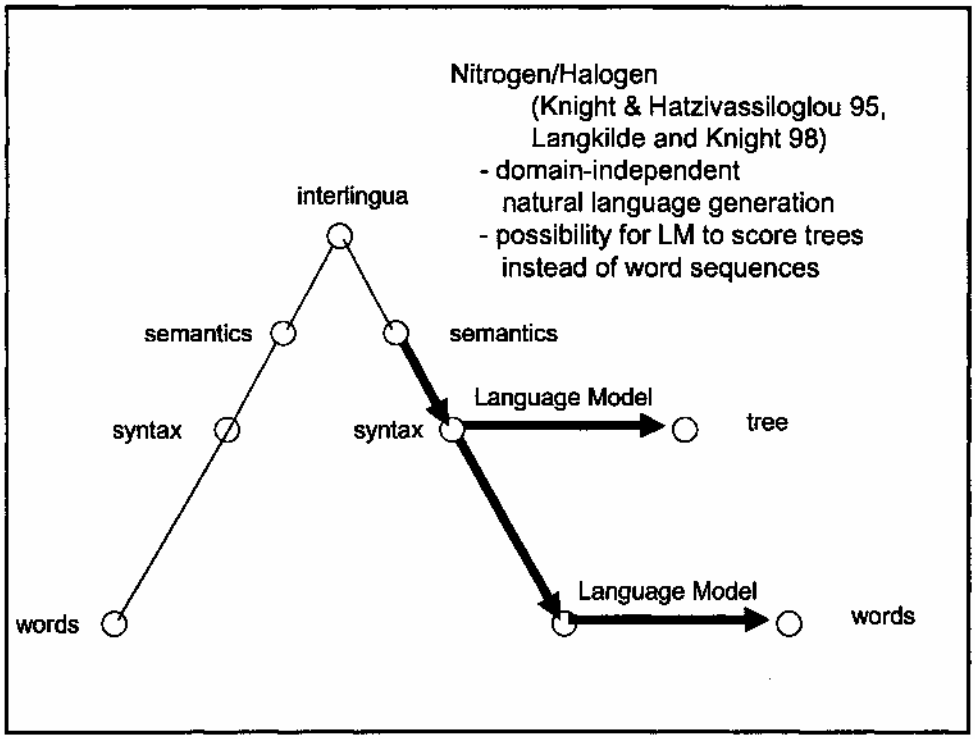


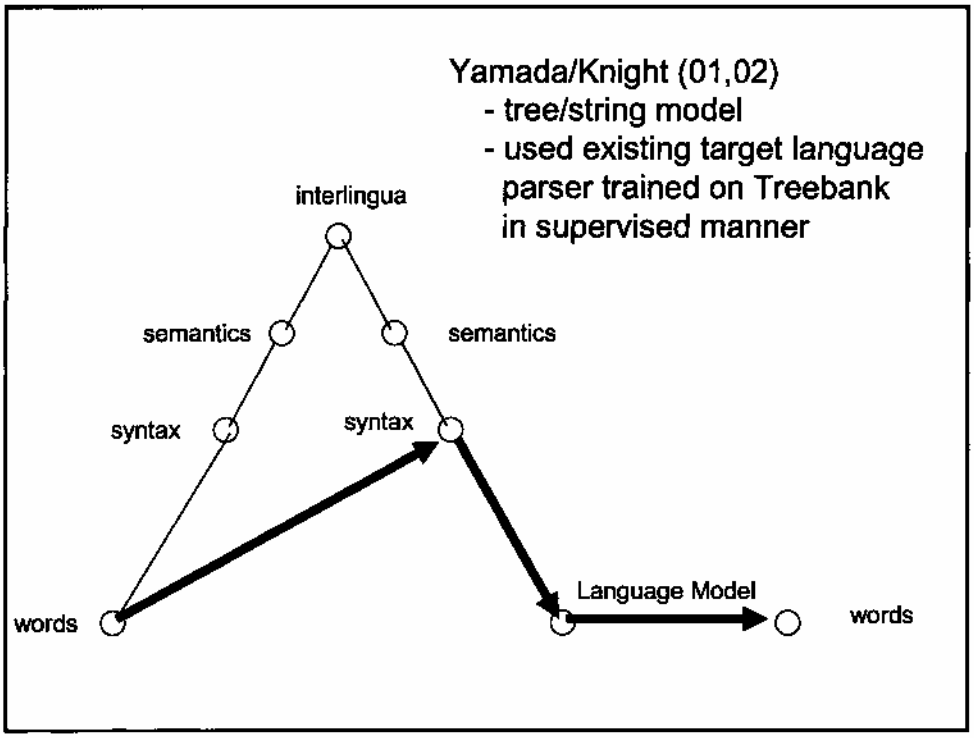

Su 95
- candidate outputs at each step are generated by rules
- statistical training learns how to rank candidates



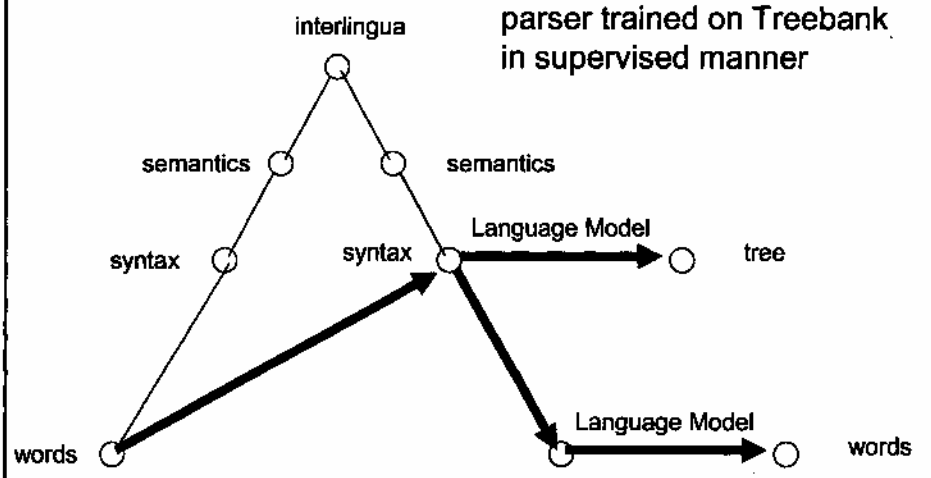

Japangloss (Knight et al 95)
- domain-independent
meaning-based translation
- composition rules implemented
with feature unification
- use of a target language model
to cope with ambiguity
interlingua
semantics
semantics
syntax
syntax
Language Model
words
words


Nitrogen/Halogen
(Knight & Hatzivassiloglou 95,
Langkilde and Knight 98)
- domain-independent
natural language generation
interlingua
semantics
semantics
syntax
syntax
Language Model
words
words


Nitrogen/Halogen
(Knight & Hatzivassiloglou 95,
Langkilde and Knight 98)
- domain-independent
natural language generation
- possibility for LM to score trees
instead of word sequences
interlingua
semantics
semantics
syntax
syntax
Language Model
tree
Language Model
words
words

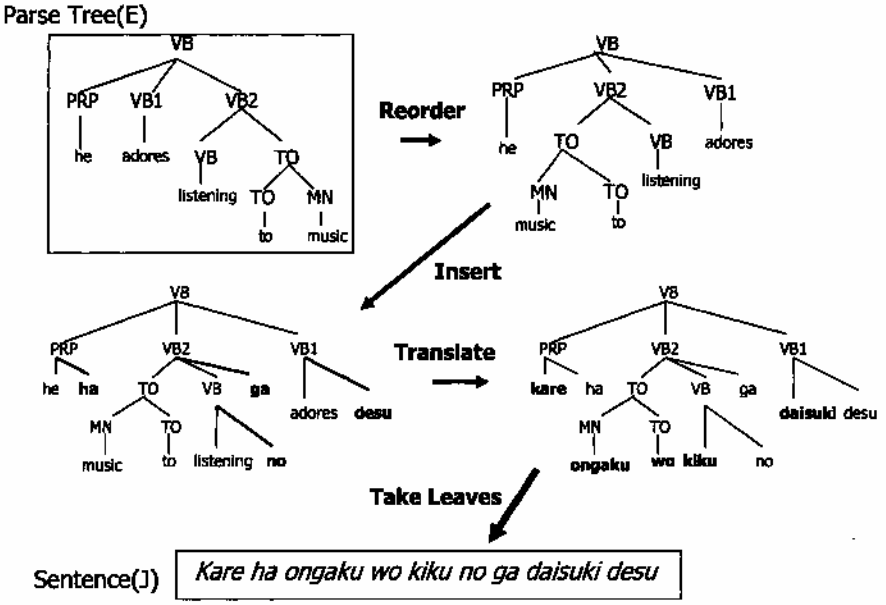

Yamada/Knight (01,02)
- tree/string model
- used existing target language
parser trained on Treebank
in supervised manner
interlingua
semantics
semantics
syntax
syntax
Language Model
words
words

Yamada/Knight (01,02)
- tree/string model
- used existing target language parser trained on Treebank in supervised manner




# Yamada/Knight 01: Modeling and Training

Parse Tree(E)




Sentence(J) *Kare ha ongaku wo kiku no ga daisuki desu*

# Japanese/English Reorder Table

| Original Order | Reordering | P(reorder\|original) |
|---|---|---|
| **PRP VB1 VB2** | PRP VB1 VB2 | 0.074 |
| | **PRP VB2 VB1** | **0.723** |
| | VB1 PRP VB2 | 0.061 |
| | VB1 VB2 PRP | 0.037 |
| | VB2 PRP VB1 | 0.083 |
| | VB2 VB1 PRP | 0.021 |
| **VB TO** | VB TO | 0.107 |
| | **TO VB** | **0.893** |
| **TO NN** | TO NN | 0.251 |
| | **NN TO** | **0.749** |
| | | |

For French/English, useful parameters like P(N ADJ | ADJ N).

# Japanese/English Reorder Table

How well does this re-order model fit the data?

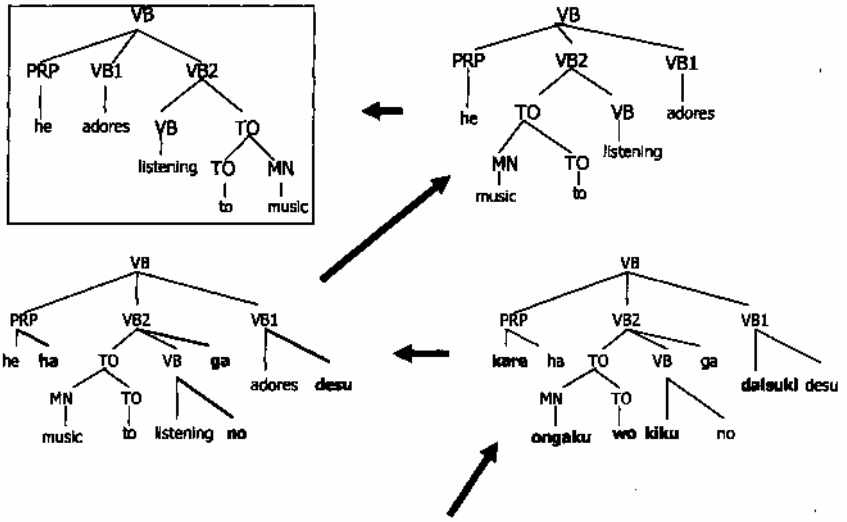
(S (NP1 (VP V NP2)) => V NP1 NP2
- one solution: flattening trees before training

No discontinuous constituents.

See [Galley, Hopkins, Knight, Marcu, 04] for a full analysis of how this model fits the data.

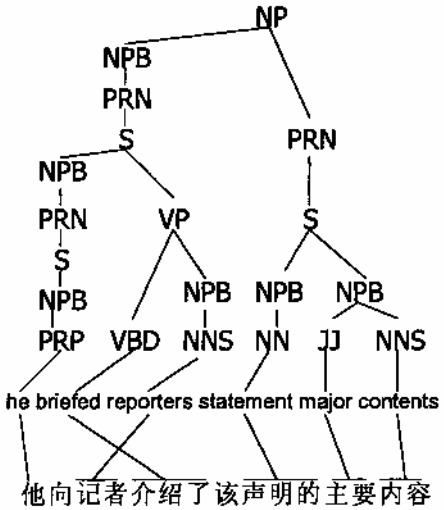
## Decoding Direction: Reverse of the Model Direction

**OUTPUT:** Many such trees, packed into a forest.




**INPUT:** *Kare ha ongaku wo kiku no ga daisuki desu*

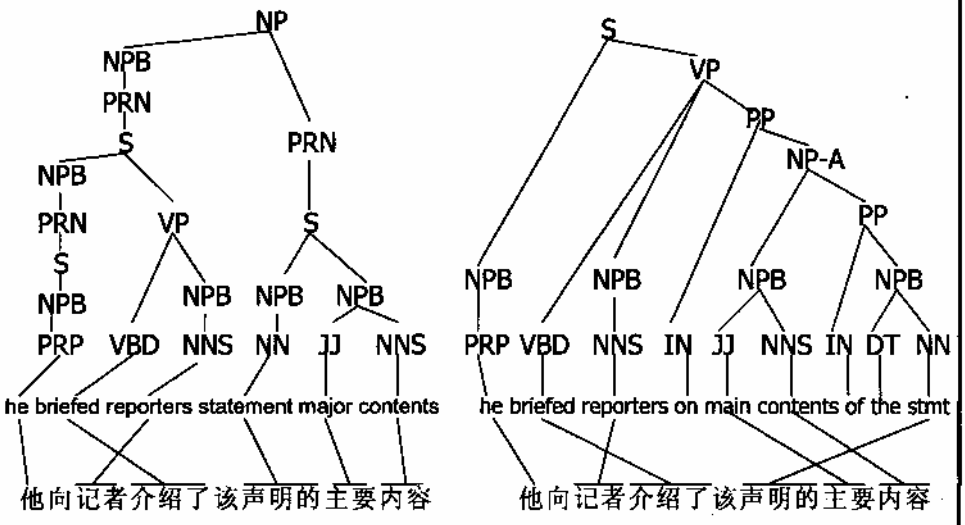
## Decoded Tree




This tree is disgusting!

Decoding with Trigram Language Model

# Decoded Tree




Decoding with Trigram LM

Decoding with Charniak Tree-Based LM

# Analysis of Charniak/Knight/Yamada

- Big increase in the number of grammatically correct outputs
- No increase in the number of semantically correct outputs
- No increase in BLEU score

"If you get it almost all the way there, then I can make it perfect." -- Charniak

(Though for dissemination applications, perfection matters).

## Analysis of Charniak/Knight/Yamada

- Only short sentences can be decoded
- Missing the phrasal translations
  - the only phrasal entries are for English constituents
- Missing integrated search
  - TM hands forest of possibilities to LM
  - pruning often eliminates good translation
- Distortion is too uncontrolled

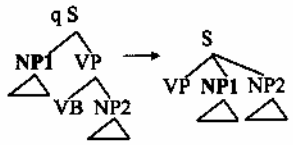
- **Lots of interesting problems ... !**

## Some Generative Statistical Models Using Syntax

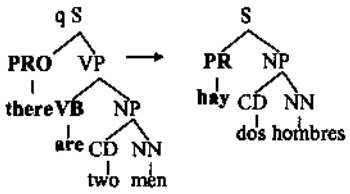
- Su 95+, Wu 96+, Alshawi et al 98
- Richardson et al 01, Yamada & Knight 01
- Cmejrek et al 03, Gildea 03, Melamed 03, Cherry & Lin 03, Schafer & Yarowsky 03
- Tree-to-tree mappings, dependency mappings, better word alignments using syntactic dependencies, projection of English syntax tools onto other languages...
- Parallel treebanks coming from UPenn & LDC
- Synchronous grammars
- Tree transducers

# Casting Syntax MT Models As Tree Transducer Automata [Graehl & Knight 04]

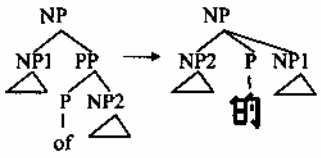
**Non-local Re-Ordering (*English/Arabic*)**



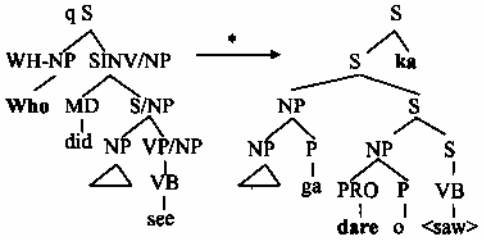

**Non-constituent Phrasal Translation (*English/Spanish*)**




**Lexicalized Re-Ordering (*English/Chinese*)**




**Long-distance Re-Ordering (*English/Japanese*)**




# Summary

- Phrase-based models are state-of-the-art
  - Word alignments
  - Phrase pair extraction & probabilities
  - N-gram language models
  - Beam search decoding
  - Feature functions & learning weights
- But the output is not English
  - Fluency must be improved
  - Better translation of person names, organizations, locations
  - More automatic acquisition of parallel data, exploitation of monolingual data across a variety of domains/languages
  - Need good accuracy across a variety of domains/languages

# Available Resources

- Bilingual corpora
  - 100m+ words of Chinese/English and Arabic/English, LDC (www.ldc.upenn.edu)
  - Lots of French/English, Spanish/French/English, LDC
  - European Parliament (sentence-aligned), 11 languages, Philipp Koehn, ISI
    - (www.isi.edu/~koehn/publications/europarl)
  - 20m words (sentence-aligned) of English/French, Ulrich Germann, ISI
    - (www.isi.edu/natural-language/download/hansard/)
- Sentence alignment
  - Dan Melamed, NYU (www.cs.nyu.edu/~melamed/GMA/docs/README.htm)
  - Xiaoyi Ma, LDC (Champollion)
- Word alignment
  - GIZA, JHU Workshop '99 (www.clsp.jhu.edu/ws99/projects/mt/)
  - GIZA++, RWTH Aachen (www-i6.Informatik.RWTH-Aachen.de/web/Software/GIZA++.html)
  - Manually word-aligned test corpus (500 French/English sentence pairs), RWTH Aachen
  - Shared task, NAACL-HLT'03 workshop
- Decoding
  - ISI ReWrite Model 4 decoder (www.isi.edu/licensed-sw/rewrite-decoder/)
  - ISI Pharaoh phrase-based decoder
- Statistical MT Tutorial Workbook, ISI (www.isi.edu/~knight/)
- Annual common-data evaluation, NIST (www.nist.gov/speech/tests/mt/index.htm)

# Some Papers Referenced on Slides

- ACL
  - [Och, Tillmann, & Ney, 1999]
  - [Och & Ney, 2000]
  - [Germann et al, 2001]
  - [Yamada & Knight, 2001, 2002]
  - [Papineni et al, 2002]
  - [Alshawi et al, 1998]
  - [Collins, 1997]
  - [Koehn & Knight, 2003]
  - [Al-Onaizan & Knight, 2002]
  - [Och & Ney, 2002]
  - [Och, 2003]
  - [Koehn et al, 2003]
- EMNLP
  - [Marcu & Wong, 2002]
  - [Fox, 2002]
  - [Munteanu & Marcu, 2002]
- AI Magazine
  - [Knight, 1997]
- www.isi.edu/~knight
  - [MT Tutorial Workbook]
- AMTA
  - [Soricut et al, 2002]
  - [Al-Onaizan & Knight, 1998]
- EACL
  - [Cmejrek et al, 2003]
- Computational Linguistics
  - [Brown et al, 1993]
  - [Knight, 1999]
  - [Wu, 1997]
- AAAI
  - [Koehn & Knight, 2000]
- IWNLG
  - [Habash, 2002]
- MT Summit
  - [Charniak, Knight, Yamada, 2003]
- NAACL
  - [Koehn, Marcu, Och, 2003]
  - [Germann, 2003]
  - [Graehl & Knight, 2004]
  - [Galley, Hopkins, Knight, Marcu, 2004]