Introduction to
(Statistical) Machine Translation

Cristina España i Bonet

MAI–ANLP

Spring 2014
Overview

1. Introduction
2. Basics
3. Components
4. The log-linear model
5. Beyond standard SMT

Part I: SMT background

∼ 120 min
Overview

6 MT Evaluation basics

7 Translation system

Part II: MT evaluation
30 min

Part III: Exercise
30 min
Part I

SMT background
Goal

echo 'das ist ein kleines haus' | moses -f moses.ini
Outline

1. Introduction
2. Basics
3. Components
4. The log-linear model
5. Beyond standard SMT
Introduction

Machine Translation Taxonomy

Machine Translation systems

- Human Translation with Machine Support
- Machine Translation with Human Support
- Fully Automated Translation
Introduction

Machine Translation Taxonomy

Machine Translation systems

- Human Translation with Machine Support
- Machine Translation with Human Support
- Fully Automated Translation
  - Empirical systems
  - Rule-based systems
Introduction

Machine Translation Taxonomy

- Machine Translation systems
  - Human Translation with Machine Support
  - Machine Translation with Human Support
  - Fully Automated Translation
    - Empirical systems
    - Rule-based systems
Introduction

Machine Translation Taxonomy

Machine Translation systems

- Human Translation with Machine Support
- Machine Translation with Human Support
- Fully Automated Translation

Empirical systems

- Statistical Machine Translation
- Example-based Translation

Rule-based systems
Introduction

Machine Translation Taxonomy

Machine Translation systems

- Human Translation with Machine Support
- Machine Translation with Human Support
- Fully Automated Translation
  - Empirical systems
  - Rule-based systems
  - Statistical Machine Translation
  - Example-based Translation
Introduction
Empirical Machine Translation

Empirical MT relies on aligned corpora
Empirical MT relies on aligned corpora
The year is 50 B.C. Gaul is entirely occupied by the Romans. Well, not entirely... One small village of indomitable Gauls still holds out against the invaders. And life is not easy for the Roman legionaries who garrison the fortified camps of Totorum, Aquarium, Laudanum and Compendium...

Som a l’any 50 abans de Crist. Tota la Gàl·lia és ocupada pels romans... Tota? No! Un llogaret del Nord habitat per gals indomables rebutja una i altra vegada ferotgement l’invassor. La vida doncs no és gens planera per als legionaris romans dels petits campaments de Babaòrum, Aquàrium, Laundànum i Petibònum...
Empirical MT relies on large parallel aligned corpora

The year is 50 B.C. Gaul is entirely occupied by the Romans. Yes, not entirely. One small village of indomitable Gauls still holds out against the invaders. And life is not easy for the Roman legionaries who garrison the fortified camps of Totorum, Aquarium, Laudanum and Compendium...

Som a l’any 50 abans de Crist. Tota la Gàl·lia és ocupada pels romans... Tota? No! Un llogaret del Nord habitat per gals indomables rebutja una i altra vegada ferotgement l’invassor. La vida doncs no és gens planera per als legionaris romans dels petits campaments de Babaòrum, Aquàrium, Laundànun i Petibònun...
Empirical MT relies on large parallel aligned corpora

The year is 50 B.C. Gaul is entirely occupied by the Romans. Well, not entirely... One small village of indomitable Gauls still holds out against the invaders. And life is not easy for the Roman legionaries who garrison the fortified camps of Totorum, Aquarium, Laudanum and Compendium...

Asterix, the hero of these adventures. A shrewd, cunning little warrior; all perilous missions are immediately entrusted to him. Asterix gets his superhuman strength from the magic potion brewed by the druid Getafix...

Obelix, Asterix’s inseparable friend. A menhir delivery-man by trade; addicted to wild boar. Obelix is always ready to drop everything and go off on a new adventure with Asterix - so long as there’s wild boar to eat, and plenty of fighting.

Finally, Vitalstitistix, the chief of the tribe. Majestic, brave and hot-tempered, the old warrior is respected by his men and feared by his enemies. Vitalstitistix himself has only one fear; he is afraid the sky may fall on his head tomorrow. But as he always says, “Tomorrow never comes”. 

Som a l’any 50 abans de Crist. Tota la Gàl·lia és ocupada pels romans... Tota? No! Un llogaret del Nord habitat per gals indomables rebutja una i altra vegada ferotgament l’invassor. La vida doncs no és gens planera per als legionaris romans dels petits campaments de Babaòrum, Aquàrium, Laundànnum i Petibònnum... 


Obèlix. És l’antic inseparable d’Astèrix. Fa de repartidor de menhirs i li agraça d’allò més la carn de porc senglar. És capaç d’abandonar-ho tot per tal de seguir Astèrix en una nova aventura. Sobretot si no hi manquen els senglars i fortes batisses.

Copdegarròtix. És el cap de la tribu. Majestuós, valent i desconfiat alhora, el vell guerrer és respectat pels seus homes i temut pels seus enemics. Tan sols una cosa li fa por: que el cel li pugui caure damunt del cap! Però, tal com ell mateix acostuma a dir, “Qui dia passa, any empeny!”.


## Introduction

Empirical Machine Translation

### Aligned parallel corpora: Numbers

<table>
<thead>
<tr>
<th>Corpus</th>
<th># segments (app.)</th>
<th># words (app.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-Acquis</td>
<td>$1.0 \cdot 10^6$</td>
<td>$30 \cdot 10^6$</td>
</tr>
<tr>
<td>Europarl</td>
<td>$2.0 \cdot 10^6$</td>
<td>$55 \cdot 10^6$</td>
</tr>
<tr>
<td>United Nations</td>
<td>$10.7 \cdot 10^6$</td>
<td>$300 \cdot 10^6$</td>
</tr>
</tbody>
</table>

### Books

<table>
<thead>
<tr>
<th>Title</th>
<th># words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bible</td>
<td>$0.8 \cdot 10^6$</td>
</tr>
<tr>
<td>The Dark Tower series</td>
<td>$1.2 \cdot 10^6$</td>
</tr>
<tr>
<td>Encyclopaedia Britannica</td>
<td>$44 \cdot 10^6$</td>
</tr>
</tbody>
</table>
## Aligned parallel corpora: Numbers

<table>
<thead>
<tr>
<th>Corpus</th>
<th># segments (app.)</th>
<th># words (app.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-Acquis</td>
<td>$1.0 \cdot 10^6$</td>
<td>$30 \cdot 10^6$</td>
</tr>
<tr>
<td>Europarl</td>
<td>$2.0 \cdot 10^6$</td>
<td>$55 \cdot 10^6$</td>
</tr>
<tr>
<td>United Nations</td>
<td>$10.7 \cdot 10^6$</td>
<td>$300 \cdot 10^6$</td>
</tr>
</tbody>
</table>

## Books

<table>
<thead>
<tr>
<th>Title</th>
<th># words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bible</td>
<td>$0.8 \cdot 10^6$</td>
</tr>
<tr>
<td>The Dark Tower series</td>
<td>$1.2 \cdot 10^6$</td>
</tr>
<tr>
<td>Encyclopaedia Britannica</td>
<td>$44 \cdot 10^6$</td>
</tr>
</tbody>
</table>
**In practice**

### WMT13 parallel data

<table>
<thead>
<tr>
<th>Corpus</th>
<th># segments</th>
<th># tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl ENG</td>
<td>1,928,274</td>
<td>52,048,855</td>
</tr>
<tr>
<td>Europarl SPA</td>
<td>1,928,274</td>
<td>53,996,661</td>
</tr>
<tr>
<td>News Commentary ENG</td>
<td>155,615</td>
<td>3,901,839</td>
</tr>
<tr>
<td>News Commentary SPA</td>
<td>155,615</td>
<td>4,364,802</td>
</tr>
<tr>
<td>United Nations ENG</td>
<td>10,749,388</td>
<td>283,672,192</td>
</tr>
<tr>
<td>United Nations SPA</td>
<td>10,749,388</td>
<td>318,045,340</td>
</tr>
<tr>
<td>Total (ENG+SPA)</td>
<td>25,666,554</td>
<td>716,029,689</td>
</tr>
</tbody>
</table>

In practice

Shows real examples of the previous theory, always from freely available data/software:

- Data: www.statmt.org/wmt13/
- Software: SRILM, GIZA++, & Moses

Standard tools, but not exclusive

More on the hands-on!
Outline

1. Introduction
2. Basics
3. Components
4. The log-linear model
5. Beyond standard SMT
SMT, basics
The beginnings, summarised timeline

1950  1975  2000

1988
Candide

IBM
CANDIDE system
SMT, basics
The beginnings, summarised timeline

1946 ENIAC
1949 Weaver memo
1966 Alpac report
1988 Candide

SMT, basics
The beginnings, summarised timeline

1946 ENIAC
1949 Weaver memo
1966 Alpac report
1988 Candide

1950 1975 2000

Dictionary MT systems
Rule-based MT systems

Empirical MT systems
BLEU
SMT, basics
The beginnings, summarised timeline

1946 ENIAC

1949 Weaver memo

1950

1966 Alpac report

1975

1988 Candide

2000 BLEU

Dictionary MT systems

Rule-based MT systems

Empirical MT systems
The Noisy Channel as a statistical approach to translation:

Good morning! →
The Noisy Channel as a statistical approach to translation:

Good morning! → Bon jour!
The Noisy Channel as a statistical approach to translation:

- **Language** $E$ ($e \in E$)
- **Language** $F$ ($f \in F$)

**e:** Good morning!  
**f:** Bon jour!
The Noisy Channel approach

Mathematically:

\[ P(e|f) \]
SMT, basics
The Noisy Channel approach

Mathematically:

\[ P(e|f) = \frac{P(e) P(f|e)}{P(f)} \]

\[ T(f) = \hat{e} = \arg\max_e P(e|f) = \arg\max_e P(e) P(f|e) \]
SMT, basics

Components

$$T(f) = \hat{e} = \arg\max_e P(e) P(f|e)$$

Language Model
- Takes care of fluency in the target language
- Data: corpora in the target language

Translation Model
- Lexical correspondence between languages
- Data: aligned corpora in source and target languages

argmax
- Search done by the decoder
SMT, basics

Components

\[ T(f) = \hat{e} = \arg\max_e P(e) P(f|e) \]

Language Model
- Takes care of fluency in the target language
- Data: corpora in the target language

Translation Model
- Lexical correspondence between languages
- Data: aligned corpora in source and target languages

argmax
- Search done by the decoder
SMT, basics
Components

\[
T(f) = \hat{e} = \argmax_e P(e) P(f|e)
\]

Language Model
- Takes care of fluency in the target language
- Data: corpora in the target language

Translation Model
- Lexical correspondence between languages
- Data: aligned corpora in source and target languages

\texttt{argmax}
- Search done by the \textit{decoder}
Language model

\[ T(f) = \hat{e} = \text{argmax}_e \ P(e) \ P(f|e) \]

Estimation of how probable a sentence is.

Naïve estimation on a corpus with \( N \) sentences:

Frequentist probability of a sentence \( e \):

\[ P(e) = \frac{N_e}{N_{\text{sentences}}} \]

Problem:

- Long chains are difficult to observe in corpora.
  \[ \Rightarrow \] Long sentences may have zero probability!
SMT, components
The language model $P(e)$

Language model

$$T(f) = \hat{e} = \arg\max_e P(e) P(f|e)$$

Estimation of how probable a sentence is.

Naïve estimation on a corpus with $N$ sentences:

Frequentist probability of a sentence $e$: 

$$P(e) = \frac{N_e}{N_{sentences}}$$

Problem:

- Long chains are difficult to observe in corpora.
  ⇒ Long sentences may have zero probability!
### Language model

\[
T(f) = \hat{e} = \arg \max_{e} P(e) P(f|e)
\]

Estimation of how probable a sentence is.

Naïve estimation on a corpus with \( N \) sentences:

- Frequentist probability of a sentence \( e \):
  \[
  P(e) = \frac{N_e}{N_{sentences}}
  \]

Problem:
- Long chains are difficult to observe in corpora.
  \( \Rightarrow \) Long sentences may have zero probability!
The n-gram approach

The language model assigns a probability \( P(e) \) to a sequence of words \( e \Rightarrow \{w_1, \ldots, w_m\} \).

\[
P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1})
\]

- The probability of a sentence is the product of the conditional probabilities of each word \( w_i \) given the previous ones.
- Independence assumption: the probability of \( w_i \) is only conditioned by the \( n \) previous words.
Example, a 4-gram model

e: All work and no play makes Jack a dull boy

\[ P(e) = P(All|\phi,\phi,\phi) \cdot P(work|\phi,\phi,All) \cdot P(\text{and}|\phi,All,work) \]
\[ P(\text{no}|All,work,\text{and}) \cdot P(\text{play}|work,\text{and},\text{no}) \]
\[ P(\text{makes}|\text{and},\text{no},\text{play}) \cdot P(\text{Jack}|\text{no},\text{play},\text{makes}) \]
\[ P(\text{a}|\text{play},\text{makes},\text{Jack}) \cdot P(\text{dull}|\text{makes},\text{Jack},\text{a}) \]
\[ P(\text{boy}|\text{Jack},\text{a},\text{dull}) \]

where, for each factor,
\[ P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})} \]
Example, a 4-gram model

$e$: All work and no play makes Jack a dull boy

$$P(e) = P(\text{All}|\phi,\phi,\phi) P(\text{work}|\phi,\phi,\text{All}) P(\text{and}|\phi,\text{All},\text{work})$$

$$P(\text{no}|\text{All},\text{work},\text{and}) P(\text{play}|\text{work},\text{and},\text{no})$$

$$P(\text{makes}|\text{and},\text{no},\text{play}) P(\text{Jack}|\text{no},\text{play},\text{makes})$$

$$P(\text{a}|\text{play},\text{makes},\text{Jack}) P(\text{dull}|\text{makes},\text{Jack},\text{a})$$

$$P(\text{boy}|\text{Jack},\text{a},\text{dull})$$

where, for each factor,

$$P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
The language model $P(e)$

Example, a 4-gram model

$e$: All work and no play makes Jack a dull boy

$$P(e) = P(\text{All}|\phi,\phi,\phi) \cdot P(\text{work}|\phi,\phi,\text{All}) \cdot P(\text{and}|\phi,\text{All},\text{work}) \cdot P(\text{no}|\text{All,work,and}) \cdot P(\text{play}|\text{work, and, no}) \cdot P(\text{makes}|\text{and, no, play}) \cdot P(\text{Jack}|\text{no, play, makes}) \cdot P(\text{a}|\text{play, makes, Jack}) \cdot P(\text{dull}|\text{makes, Jack, a}) \cdot P(\text{boy}|\text{Jack, a, dull})$$

where, for each factor,

$$P(\text{and}|\phi,\text{All, work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
Example, a 4-gram model

\[ e: \textcolor{red}{\underline{\text{All}}} \textcolor{blue}{\text{work}} \text{ and } \textcolor{red}{\underline{\text{no play}}} \text{ makes } \text{Jack} \text{ a dull boy} \]

\[ P(e) = P(\text{All}|\phi,\phi,\phi) \cdot P(\text{work}|\phi,\phi,\text{All}) \cdot P(\text{and}|\phi,\text{All},\text{work}) \]
\[ \cdot P(\text{no}|\text{All,work,and}) \cdot P(\text{play}|\text{work,and,}\text{no}) \]
\[ \cdot P(\text{makes}|\text{and,}\text{no,play}) \cdot P(\text{Jack}|\text{no,play,makes}) \]
\[ \cdot P(\text{a}|\text{play,makes,Jack}) \cdot P(\text{dull}|\text{makes,Jack,a}) \]
\[ \cdot P(\text{boy}|\text{Jack,a,dull}) \]

where, for each factor,

\[ P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})} \]
Example, a 4-gram model

\[ e: \underline{All \ work} \ and \ \underline{no \ play} \ makes \ Jack \ a \ dull \ boy \]

\[ P(e) = P(\text{All}|\phi,\phi,\phi) \ P(\text{work}|\phi,\phi,\text{All}) \ P(\text{and}|\phi,\text{All},\text{work}) \]
\[ P(\text{no}|\text{All},\text{work},\text{and}) \ P(\text{play}|\text{work},\text{and},\text{no}) \]
\[ P(\text{makes}|\text{and},\text{no},\text{play})P(\text{Jack}|\text{no},\text{play},\text{makes}) \]
\[ P(\text{a}|\text{play},\text{makes},\text{Jack})P(\text{dull}|\text{makes},\text{Jack},\text{a}) \]
\[ P(\text{boy}|\text{Jack},\text{a},\text{dull}) \]

where, for each factor,

\[ P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})} \]
SMT, components
The language model $P(e)$

Example, a 4-gram model

$$e: \text{All work and no play makes Jack a dull boy}$$

$$P(e) = P(\text{All}|\phi, \phi, \phi) \cdot P(\text{work}|\phi, \phi, \text{All}) \cdot P(\text{and}|\phi, \text{All}, \text{work}) \cdot P(\text{no}|\text{All}, \text{work}, \text{and}) \cdot P(\text{play}|\text{work}, \text{and}, \text{no}) \cdot P(\text{makes}|\text{and}, \text{no}, \text{play}) \cdot P(\text{Jack}|\text{no}, \text{play}, \text{makes}) \cdot P(\text{a}|\text{play}, \text{makes}, \text{Jack}) \cdot P(\text{dull}|\text{makes}, \text{Jack}, \text{a}) \cdot P(\text{boy}|\text{Jack}, \text{a}, \text{dull})$$

where, for each factor,

$$P(\text{and}|\phi, \text{All}, \text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
SMT, components

The language model $P(e)$

Example, a 4-gram model

$e$: All work and no play makes Jack a dull boy

$$P(e) = P(\text{All}|\phi,\phi,\phi) P(\text{work}|\phi,\phi,\text{All}) P(\text{and}|\phi,\text{All},\text{work})$$
$$P(\text{no}|\text{All,work,and}) P(\text{play}|\text{work,and,}\text{no})$$
$$P(\text{makes}|\text{and,}\text{no,play}) P(\text{Jack}|\text{no,play,}\text{makes})$$
$$P(\text{a}|\text{play,}\text{makes,Jack}) P(\text{dull}|\text{makes,Jack,a})$$
$$P(\text{boy}|\text{Jack,a,dull})$$

where, for each factor,

$$P(\text{and}|\phi,\text{All,work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
The language model $P(e)$

Example, a 4-gram model

$$e: \text{All work and no play makes Jack a dull boy}$$

$$P(e) = P(\text{All}|\phi,\phi,\phi) \cdot P(\text{work}|\phi,\phi,\text{All}) \cdot P(\text{and}|\phi,\text{All},\text{work}) \cdot P(\text{no}|\text{All},\text{work},\text{and}) \cdot P(\text{play}|\text{work},\text{and},\text{no}) \cdot P(\text{makes}|\text{and},\text{no},\text{play}) \cdot P(\text{Jack}|\text{no},\text{play},\text{makes}) \cdot P(\text{a}|\text{play},\text{makes},\text{Jack}) \cdot P(\text{dull}|\text{makes},\text{Jack},\text{a}) \cdot P(\text{boy}|\text{Jack},\text{a},\text{dull})$$

where, for each factor,

$$P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
The language model $P(e)$

Example, a 4-gram model

$$P(e) = P(\text{All} | \phi, \phi, \phi) P(\text{work} | \phi, \phi, \text{All}) P(\text{and} | \phi, \text{All}, \text{work})$$
$$P(\text{no} | \text{All}, \text{work}, \text{and}) P(\text{play} | \text{work}, \text{and}, \text{no})$$
$$P(\text{makes} | \text{and}, \text{no}, \text{play}) P(\text{Jack} | \text{no}, \text{play}, \text{makes})$$
$$P(\text{a} | \text{play}, \text{makes}, \text{Jack}) P(\text{dull} | \text{makes}, \text{Jack}, \text{a})$$
$$P(\text{boy} | \text{Jack}, \text{a}, \text{dull})$$

where, for each factor,

$$P(\text{and} | \phi, \text{All}, \text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}$$
Example, a 4-gram model

\[ e: \text{All work and no play makes Jack a dull boy} \]

\[
P(e) = P(\text{All} | \phi, \phi, \phi) P(\text{work} | \phi, \phi, \text{All}) P(\text{and} | \phi, \text{All}, \text{work})
\]
\[
P(\text{no} | \text{All}, \text{work}, \text{and}) P(\text{play} | \text{work}, \text{and}, \text{no})
\]
\[
P(\text{makes} | \text{and}, \text{no}, \text{play}) P(\text{Jack} | \text{no}, \text{play}, \text{makes})
\]
\[
P(\text{a} | \text{play}, \text{makes}, \text{Jack}) P(\text{dull} | \text{makes}, \text{Jack}, \text{a})
\]
\[
P(\text{boy} | \text{Jack}, \text{a}, \text{dull})
\]

where, for each factor,

\[
P(\text{and} | \phi, \text{All}, \text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}
\]
Example, a 4-gram model

e: All work and no play makes Jack a dull boy

\[
P(e) = P(\text{All}|\phi,\phi,\phi) P(\text{work}|\phi,\phi,\text{All}) P(\text{and}|\phi,\text{All},\text{work}) \\
P(\text{no}|\text{All},\text{work},\text{and}) P(\text{play}|\text{work},\text{and},\text{no}) \\
P(\text{makes}|\text{and},\text{no},\text{play}) P(\text{Jack}|\text{no},\text{play},\text{makes}) \\
P(\text{a}|\text{play},\text{makes},\text{Jack}) P(\text{dull}|\text{makes},\text{Jack},\text{a}) \\
P(\text{boy}|\text{Jack},\text{a},\text{dull})
\]

where, for each factor,

\[
P(\text{and}|\phi,\text{All},\text{work}) = \frac{N(\text{All work and})}{N(\text{All work})}
\]
Example, a 4-gram model

\[ e: \text{All work and no play makes Jack a dull boy} \]

\[ P(e) = P(\text{All} | \phi, \phi, \phi) \cdot P(\text{work} | \phi, \phi, \text{All}) \cdot P(\text{and} | \phi, \text{All}, \text{work}) \]
\[ \cdot P(\text{no} | \text{All, work, and}) \cdot P(\text{play} | \text{work, and, no}) \]
\[ \cdot P(\text{makes} | \text{and, no, play}) \cdot P(\text{Jack} | \text{no, play, makes}) \]
\[ \cdot P(\text{a} | \text{play, makes, Jack}) \cdot P(\text{dull} | \text{makes, Jack, a}) \]
\[ \cdot P(\text{boy} | \text{Jack, a, dull}) \]

where, for each factor,

\[ P(\text{and} | \phi, \text{All, work}) = \frac{N(\text{All work and})}{N(\text{All work})} \]
Main problems and criticisms:

- Long-range dependencies are lost.
- Still, some $n$-grams can be not observed in the corpus.

Solution

Smoothing techniques:

- Linear interpolation.

$$P(\text{and}|\text{All, work}) = \frac{N_{(\text{All, work, and})}}{N_{(\text{All, work})}} + \lambda_2 \frac{N_{(\text{work, and})}}{N_{(\text{work})}} + \lambda_1 \frac{N_{(\text{and})}}{N_{\text{words}}} + \lambda_0$$
SMT, components
The language model $P(e)$

Main problems and criticisms:
- Long-range dependencies are lost.
- Still, some $n$-grams can be not observed in the corpus.

Solution

Smoothing techniques:
- Linear interpolation.
- Back-off models.

$$P(\text{and}|\text{All, work}) = \frac{N(\text{All, work, and})}{N(\text{All, work})} + \lambda_2 \frac{N(\text{work, and})}{N(\text{work})} + \lambda_1 \frac{N(\text{and})}{N(\text{words})} + \lambda_0$$
Main problems and criticisms:

- Long-range dependencies are lost.
- Still, some \( n \)-grams can be not observed in the corpus.

Solution

Smoothing techniques:

- Linear interpolation.

\[
P(\text{and}|\text{All, work}) = \frac{N_{(\text{All, work, and})}}{N_{(\text{All, work})}} + \lambda_2 \frac{N_{(\text{work, and})}}{N_{(\text{work})}} + \lambda_1 \frac{N_{(\text{and})}}{N_{\text{words}}} + \lambda_0
\]
Main problems and criticisms:

- Long-range dependencies are lost.
- Still, some \( n \)-grams can be not observed in the corpus.

Solution

Smoothing techniques:

- Linear interpolation.

\[
P(\text{and}|\text{All, work}) = \lambda_3 \frac{N(\text{All, work, and})}{N(\text{All, work})} + \lambda_2 \frac{N(\text{work, and})}{N(\text{work})} + \lambda_1 \frac{N(\text{and})}{N_{\text{words}}} + \lambda_0
\]
In practice,

cluster:/home/quest/corpus/lm> ls -lkh

-rw-r--r-- 1 emt ia  507M  mar  3  15:28 europarl.lm
-rw-r--r-- 1 emt ia  50M   mar  3  15:29   nc.lm
-rw-r--r-- 1 emt ia  3,1G  mar  3  15:33  un.lm

cluster:/home/quest/corpus/lm> wc -l

      15,181,883   europarl.lm
        1,735,721    nc.lm
      82,504,380   un.lm
SMT, components

The language model $P(e)$

```
cluster:/home/quest/corpus/lm> more nc.lm

data
ngram 1=655770
ngram 2=11425501
ngram 3=10824125
ngram 4=13037011
ngram 5=12127575

1-grams:
-3.142546 ! -1.415594
-1.978775 " -0.9078496
-4.266428 # -0.2729652
-3.806078 $ -0.3918373
-3.199419 % -1.139753
-3.613416 & -0.6046973
-2.712332 ' -0.6895114
```
SMT, components

The language model $P(e)$

\[2\text{-grams:}\]
\[-1.08232 \text{ concierto },\]
\[-1.093977 \text{ concierto } -0.2378127\]
\[-1.747908 \text{ concierto ad}\]
\[-1.748422 \text{ concierto cobraria}\]
\[-0.8927398 \text{ concierto de}\]
\[-1.744176 \text{ concierto europeo}\]
\[-1.740879 \text{ concierto internacional}\]
\[-1.635606 \text{ concierto para}\]
\[-1.744787 \text{ concierto regional}\]

...

\[5\text{-grams:}\]
\[-0.8890668 \text{ no son los unicos culpables}\]
\[-1.396196 \text{ no son los unicos problemas}\]
\[-0.7550655 \text{ no son los unicos que}\]
\[-1.240193 \text{ no son los unicos responsables}\]
Language model: keep in mind

- Statistical LMs estimate the probability of a sentence from its n-gram frequency counts in a monolingual corpus.
- Within an SMT system, it contributes to select fluent sentences in the target language.
- Smoothing techniques are used so that not frequent translations are not discarded beforehand.
SMT, components
The translation model $P(f|e)$

Translation model

$$T(f) = \hat{e} = \arg\max_e P(e) P(f|e)$$

Estimation of the lexical correspondence between languages.

How can be $P(f|e)$ characterised?

NULL Quan tornes a casa ?

When are you coming back home ?
SMT, components

The translation model $P(f|e)$

**Translation model**

$$T(f) = \hat{e} = \arg\max_e P(e) P(f|e)$$

Estimation of the lexical correspondence between languages.

How can be $P(f|e)$ characterised?

NULL Quan tornes a casa ?

When are you coming back home ?
SMT, components
The translation model $P(f|e)$

**Translation model**

$$T(f) = \hat{e} = \operatorname{argmax}_e P(e) P(f|e)$$

Estimation of the lexical correspondence between languages.

How can be $P(f|e)$ characterised?

NULL Quan tornes a casa ?

When are you coming back home ?
SMT, components
The translation model \( P(f|e) \)

One should at least model for each word in the source language:

- Its translation,
- the number of necessary words in the target language,
- the position of the translation within the sentence,
- and, besides, the number of words that need to be generated from scratch.
Word-based models: the IBM models

They characterise \( P(f|e) \) with 4 parameters: \( t, n, d \) and \( p_1 \).

- Lexical probability \( t \)
  \( t(\text{Quan}|\text{When}) \): the probability that Quan translates into When.

- Fertility \( n \)
  \( n(3|\text{tornes}) \): the probability that tornes generates 3 words.
Word-based models: the IBM models

They characterise $P(f|e)$ with 4 parameters: $t$, $n$, $d$ and $p_1$.

- Distortion $d$
  
  $d(j|i, m, n)$: the prob. that the word in the $j$ position generates a word in the $i$ position. $m$ and $n$ are the length of the source and target sentences.

- Probability $p_1$
  
  $p(\text{you}|\text{NULL})$: the prob. that the spurious word you is generated (from NULL).
SMT, components

The translation model $P(f|e)$

Back to the example:

```
NULL  Quan tornes a casa ?
|     |     /              |
NULL  Quan tornes tornes tornes casa ?
|     |     |     /             |
NULL  When are coming back home ?
|     |     |     |     /             |
you  When are coming back home ?
|     |     |     |     |     /             |
When are you coming back home ?
```

- **Fertility**
- **Translation**
- **Insertion**
- **Distortion**
SMT, components

The translation model $P(f|e)$

Back to the example:

```
NULL Quan tornes a casa ?

Fertility

NULL Quantornestornestornes casa ?

Translation

NULL When are coming back home ?

Insertion

you When are coming back home ?

Distortion

When are you coming back home ?
```
SMT, components

The translation model $P(f|e)$

Back to the example:

NULL Quan tornes a casa ?

NULL Quan tornestornestornes casa ?

NULL When are coming back home ?

you When are coming back home ?

When are you coming back home ?
SMT, components
The translation model $P(f|e)$

Back to the example:

```
NULL  Quan tornes a casa ?
      |     |     |     |     |
NULL  Quan tornes tornes tornes casa ?
      |     |     |     |     |     |     |
NULL  When are coming back home ?
      |     |     |     |     |     |     |     |
you  When are coming back home ?
      |     |     |     |     |     |     |     |     |
When are you coming back home ?
```

- **Fertility**
- **Translation**
- **Insertion**
- **Distortion**
SMT, components

The translation model $P(f|e)$

Back to the example:

NULL Quan tornes a casa?

NULL Quan tornes tornes tornes tornes casa?

NULL When are coming back home?

you When are coming back home?

When are you coming back home?
Word-based models: the IBM models

How can \( t, n, d \) and \( p_1 \) be estimated?

- Statistical model \( \Rightarrow \) counts in a (huge) corpus!

But...
- Corpora are aligned at sentence level, not at word level.

Alternatives
- Pay someone to align 2 million sentences word by word.
- Estimate word alignments together with the parameters.
Word-based models: the IBM models

How can $t$, $n$, $d$ and $p_1$ be estimated?

- Statistical model $\Rightarrow$ counts in a (huge) corpus!

But...
- Corpora are aligned at sentence level, not at word level.

Alternatives
- Pay someone to align 2 milion sentences word by word.
- Estimate word alignments together with the parameters.
SMT, components
The translation model $P(f|e)$

**Word-based models: the IBM models**

How can $t$, $n$, $d$ and $p_1$ be estimated?

- Statistical model $\Rightarrow$ counts in a (huge) corpus!

But...
- Corpora are aligned at sentence level, not at word level.

**Alternatives**
- Pay someone to align 2 million sentences word by word.
- Estimate word alignments together with the parameters.
SMT, components
The translation model $P(f|e)$

Expectation-Maximisation algorithm

- Parameter initialisation
- Alignment probability calculation

Converged?

NO

YES

Final parameters and alignments
SMT, components
The translation model $P(f|e)$

Expectation-Maximisation algorithm

Parameter initialisation

Alignment probability calculation

Parameter reestimation

Alignment probability recalculation
Expectation-Maximisation algorithm

Parameter initialisation

Alignment probability calculation

Parameter reestimation

Alignment probability recalculation

Converged?

Final parameters and alignments
Alignment’s asymmetry

The definitions in IBM models make the alignments asymmetric:

- each target word corresponds to only one source word, but the opposite is not true due to the definition of fertility.

Catalan to English

NULL Quan tornes a casa ?

When are you coming back home ?

English to Catalan

NULL When are you coming back home ?

Quan tornes a casa ?
**Alignment’s asymmetry**

The definitions in IBM models make the alignments asymmetric:
- Each target word corresponds to only one source word, but the opposite is not true due to the definition of **fertility**.

**Catalan to English**

NULL Quan tornes a casa ?

When are you coming back home ?

**English to Catalan**

(NULL When are you coming back home ?)

Quan tornes a casa ?
### SMT, components

The translation model $P(f|e)$

**Visually:**

<table>
<thead>
<tr>
<th></th>
<th>NULL</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tbody>
</table>

**Catalan to English**
**SMT, components**

The translation model $P(f|e)$

Visually:

<table>
<thead>
<tr>
<th></th>
<th>NULL</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tbody>
</table>

English to Catalan
Alignment symmetrisation

- Intersection: high-confidence, high precision.

Catalan to English $\cap$ English to Catalan
Alignment symmetrisation

- **Union**: lower confidence, **high recall**.

SMT, components
The translation model $P(f|e)$

Catalan to English $\cup$ English to Catalan
SMT, components
The translation model $P(f|e)$

In practice,

cluster:/home/moses/giza.en-es> zmore en-es.A3.final.gz

# Sentence pair (1) source length 5 target length 4 alignment score: 0.00015062
resumption of the session
NULL ({ }) reanudacion ({ 1 }) del ({ 2 3 }) periodo ({ }) de ({ }) sesiones ({ 4 })

# Sentence pair (2) source length 33 target length 40 alignment score: 3.3682e-61
i declare resumed the session of the european parliament adjourned on friday 17
december 1999 , and i would like once again to wish you a happy new year in the
hope that you enjoyed a pleasant festive period .
NULL ({ 31 }) declaro ({ 1 }) reanudado ({ 2 3 }) el ({ 4 }) periodo ({ }) de ({ }) sesiones ({ 5 }) del ({ 6 7 }) parlamento ({ 9 }) europeo ({ 8 }), ({ })
interrumpido ({ 10 }) el ({ }) viernes ({ 12 14 }) 17 ({ 11 13 }) de ({ }) diciembre ({ 15 }) pasado ({ }), ({ 16 }) y ({ 17 }) reitero ({ 21 }) a ({ 23 }) sus ({ 30 })
señorías ({ }) mi ({ 18 }) deseo ({ 24 }) de ({ }) que ({ 33 }) hayan ({ 25 34 35 })
tenido ({ }) unas ({ 19 20 }) buenas ({ 26 36 }) vacaciones ({ 22 27 28 29 32 37 38 39 }) . ({ 40 })
In practice,


# Sentence pair (1) source length 4 target length 5 alignment score: 1.08865e-07
reanudacion del periodo de sesiones
NULL ({ 4 }) resumption ({ 1 }) of ({ 2 }) the ({ }) session ({ 3 5 })

# Sentence pair (2) source length 40 target length 33 alignment score: 1.88268e-50
declaro reanudado el periodo de sesiones del parlamento europeo, interrumpido el
viernes 17 de diciembre pasado, y reitero a sus senorias mi dese o de que hayan
tenido unas buenas vacaciones.
NULL ({ 5 10 }) i ({ }) declare ({ 1 }) resumed ({ 2 }) the ({ 3 }) session ({ 4 6 })
of ({ 7 }) the ({ }) european ({ 9 }) parliament ({ 8 12 }) adjourned ({ 11 }) on
({ 15 }) friday ({ 13 }) 17 ({ 14 }) december ({ 16 17 }) 1999 ({ }) , ({ 18 }) and
({ 19 }) i ({ }) would ({ }) like ({ }) once ({ }) again ({ }) to ({ 21 }) wish ({ })
you ({ }) a ({ }) happy ({ }) new ({ }) year ({ }) in ({ 26 }) the ({ }) hope ({ }
) that ({ 27 }) you ({ }) enjoyed ({ 20 }) a ({ }) pleasant ({ 22 23 24 25 28 29 })
festive ({ 30 31 32 }) period ({ }). ({ 33 })
SMT, components

The translation model $P(f|e)$

cluster:/home/moses/model> more aligned.grow-diag-final

0-0 1-1 1-2 2-3 4-3

0-0 0-1 1-1 1-2 2-3 3-4 5-4 6-5 6-6 8-7 7-8 11-8 10-9 13-10 14-10 12-11
SMT, components
The translation model $P(f|e)$

```bash
cluster:/home/moses/model> more lex.e2f

tuneles tunnels 0.7500000
tuneles transit 0.2000000
estructuralmente weak 1.0000000
estructuralmente structurally 0.5000000
destruido had 0.0454545
para tunnels 0.2500000
sean transit 0.2000000
transito transit 0.6000000
...

cluster:/home/moses/model> more lex.f2e

tunnels tuneles 0.7500000
transit tuneles 0.2500000
weak estructuralmente 0.5000000
structurally estructuralmente 0.5000000
...
```
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
SMT, components

The translation model \( P(f|e) \)

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.

e: \( \phi \)
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.

e: David reads
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.

e: David reads the
SMT, components
The translation model \( P(f|e) \)

From Word-based to Phrase-based models

\[
f: \text{En David llegeix el llibre nou.} \\
e: \text{David reads the book}
\]
From Word-based to Phrase-based models

\( f: \) En David llegeix el llibre \textcolor{red}{
ou}.\n
\( e: \) David reads the book \textcolor{red}{
new}.\n
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the book new. ∼
SMT, components
The translation model \( P(f|e) \)

From Word-based to Phrase-based models

\( f: \) En David llegeix el llibre nou.
\( e: \) David reads the new book. ✓
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
From Word-based to Phrase-based models

\[
f: \text{En David llegeix el llibre nou.} \\
e: \text{David reads the new book.} \quad \checkmark
\]

\[
f: \text{En David llegeix el llibre de nou.} \\
e: \phi
\]
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.

e: David reads the new book. ✓

f: En David llegeix el llibre de nou.

e: David reads the book again.

f: En David llegeix el llibre de nou.

e: David
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David *llegeix* el llibre de nou.
e: David *reads*
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the the
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book.  ✓

f: En David llegeix el \textcolor{red}{llibre} de nou.
e: David reads the \textcolor{red}{book}
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of
SMT, components
The translation model \( P(f|e) \)

From Word-based to Phrase-based models

\[
\begin{align*}
\text{f: } & \text{En David llegeix el llibre nou.} \\
\text{e: } & \text{David reads the new book.} \quad \checkmark \\
\text{f: } & \text{En David llegeix el llibre de nou.} \\
\text{e: } & \text{David reads the book of new.}
\end{align*}
\]
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. X
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. ✗
e: φ
From Word-based to Phrase-based models

\[ f: \text{En David llegeix el llibre nou.} \]
\[ e: \text{David reads the new book. } \checkmark \]

\[ f: \text{En David llegeix el llibre de nou.} \]
\[ e: \text{David reads the book of new. } \times \]
\[ e: \text{David} \]
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. ✗
e: David reads
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.

e: David reads the new book. ✓

f: En David llegeix el llibre de nou.

e: David reads the book of new. ✗
e: David reads the
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. ✗
e: David reads the book
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. ✗
e: David reads the book again.
From Word-based to Phrase-based models

f: En David llegeix el llibre nou.
e: David reads the new book. ✓

f: En David llegeix el llibre de nou.
e: David reads the book of new. ✗
e: David reads the book again. ✓
SMT, components
The translation model $P(f|e)$

From Word-based to Phrase-based models

- f: En David llegeix el llibre nou.
- e: David reads the new book. ✓
- f: En David llegeix el llibre de nou.
- e: David reads the book of new. ✗
- e: David reads the book again. ✓

- Some sequences of words usually translate together.
- Approach: take sequences (phrases) as translation units.
What can be achieved with phrase-based models (as compared to word-based models)

- Allow to translate from several to several words and not only from one to several.

- Some local and short range context is used.

- Idioms can be caught.
SMT, components
The translation model $P(f|e)$

With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.
SMT, components
The translation model $P(f|e)$

With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.
SMT, components
The translation model $P(f|e)$

With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.
SMT, components
The translation model $P(f|e)$

NULL    Quan    tornes    a casa    ?
When    are you coming back    home    ?

But...
- Alignments need to be done at phrase level

Options
- Calculate phrase-to-phrase alignments $\Rightarrow$ hard!
- Obtain phrase alignments from word alignments $\Rightarrow$ how?
Questions to answer:

- How do we obtain phrase alignments from word alignments?
- And, by the way, what’s exactly a phrase?!

A **phrase is** a sequence of words consistent with word alignment. That is, no word is aligned to a word outside the phrase. But a phrase **is not** necessarily a linguistic element.

---

1 We do not use the term phrase here in its linguistic sense: a phrase can be any sequence of words, even if they are not a linguistic constituent.
Questions to answer:

- How do we obtain phrase alignments from word alignments?
- And, by the way, what’s exactly a phrase?!

A phrase is a sequence of words consistent with word alignment. That is, no word is aligned to a word outside the phrase. But a phrase is not necessarily a linguistic element.

1 We do not use the term phrase here in its linguistic sense: a phrase can be any sequence of words, even if they are not a linguistic constituent.
Questions to answer:

- How do we obtain phrase alignments from word alignments?
- And, by the way, what’s exactly a phrase?!

A **phrase is** a sequence of words consistent with word alignment. That is, no word is aligned to a word outside the phrase. But a phrase **is not** necessarily a linguistic element.
Questions to answer:

- How do we obtain phrase alignments from word alignments?
- And, by the way, what’s exactly a phrase?!

A phrase is a sequence of words consistent with word alignment. That is, no word is aligned to a word outside the phrase. But a phrase is not necessarily a linguistic element.¹

¹We do not use the term phrase here in its linguistic sense: a phrase can be any sequence of words, even if they are not a linguistic constituent.
**Phrase extraction** through an example:

<table>
<thead>
<tr>
<th>When</th>
<th>are</th>
<th>you</th>
<th>coming</th>
<th>back</th>
<th>home</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quan</td>
<td>tornes</td>
<td>tu</td>
<td>a</td>
<td>casa</td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>

(Quan tornes, When are you coming back)
**Phrase extraction** through an example:

<table>
<thead>
<tr>
<th>When are you coming back?</th>
<th>Quan tornes</th>
<th>tu</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
</table>

(Quan tornes, When are you coming back)
Phrase extraction through an example:

Quan tornes tu a casa ?

(Quan tornes, When are you coming back)
(Quan tornes tu, When are you coming back)
SMT, components

The translation model $P(f|e)$

**Intersection**

<table>
<thead>
<tr>
<th>When are you coming back home?</th>
<th>Quan tornes a casa?</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>Quan</td>
<td>a</td>
</tr>
<tr>
<td>are</td>
<td>tornes</td>
<td>casa</td>
</tr>
<tr>
<td>you</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>coming</td>
<td>casa</td>
<td>?</td>
</tr>
<tr>
<td>back</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>home</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
**SMT, components**

The translation model $P(f|e)$

**Intersection**

<table>
<thead>
<tr>
<th></th>
<th>Quan tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>When</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>are</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>you</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>coming</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>back</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>home</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
SMT, components

The translation model $P(f|e)$

Intersection

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>are</td>
<td>you</td>
<td>coming</td>
<td>back</td>
<td>home</td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (? , ?) 10 phrases
SMT, components
The translation model $P(f|e)$

Intersection

<table>
<thead>
<tr>
<th>When are you coming back home</th>
<th>Quan tornes</th>
<th>a casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
SMT, components
The translation model \( P(f|e) \)

Intersection

<table>
<thead>
<tr>
<th></th>
<th>Quan tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
SMT, components

The translation model $P(f|e)$

**Intersection**

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>When</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>are</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>you</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>coming</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>back</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>home</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (? , ?) 10 phrases
SMT, components
The translation model \( P(f|e) \)

Intersection

<table>
<thead>
<tr>
<th>When</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (? , ?)  10 phrases
### SMT, components

The translation model $P(f|e)$

#### Intersection

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
**SMT, components**

The translation model $P(f|e)$

**Intersection**

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
SMT, components
The translation model $P(f|e)$

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases
SMT, components
The translation model $P(f|e)$

<table>
<thead>
<tr>
<th>Union</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (? , ?) 21 phrases
SMT, components
The translation model $P(f|e)$

<table>
<thead>
<tr>
<th>Union</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>are</td>
<td>you</td>
<td>coming</td>
<td>back</td>
<td>home</td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases
SMT, components

The translation model $P(f|e)$

### Union

<table>
<thead>
<tr>
<th>When</th>
<th>are</th>
<th>you</th>
<th>coming</th>
<th>back</th>
<th>home</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quan</td>
<td>tornes</td>
<td>a</td>
<td>casa</td>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases
SMT, components

The translation model $P(f|e)$

<table>
<thead>
<tr>
<th>Union</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ...

(tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases
SMT, components
The translation model $P(f|e)$

<table>
<thead>
<tr>
<th>Union</th>
<th>Quan</th>
<th>tornes</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</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>
</tr>
</tbody>
</table>

(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (? , ?) 21 phrases
Phrase extraction

- The number of extracted phrases depends on the symmetrisation method.
  - Intersection: few precise phrases.
  - Union: lots of (less?) precise phrases.

- Usually, neither intersection nor union are used, but something in between.
  - Start from the intersection and add points belonging to the union according to heuristics.
Phrase extraction

- For each phrase-pair \((f_i, e_i)\), \(P(f_i|e_i)\) is estimated by frequency counts in the parallel corpus.

- The set of possible phrase-pairs conforms the set of translation options.

- The set of phrase-pairs together with their probabilities conform the translation table.
In practice,

```bash
cluster:/home/moses/model> zmore extract.gz

reanudacion ||| resumption ||| 0-0
reanudacion del ||| resumption of the ||| 0-0 1-1 1-2
reanudacion del periodo de sesiones ||| resumption of the session ||| 0-0 1-1 1-2 2-3 4-3
```

```bash
cluster:/home/moses/model> zmore extract.inv.gz

resumption ||| reanudacion ||| 0-0
resumption of the ||| reanudacion del ||| 0-0 1-1 2-1
resumption of the session ||| reanudacion del periodo de sesiones ||| 0-0 1-1 2-1 3-2 3-4
```

```bash
cluster:/home/moses/model> zmore extract.o.gz

reanudacion ||| resumption ||| mono mono
reanudacion del ||| resumption of the ||| mono mono
reanudacion del periodo de sesiones ||| resumption of the session ||| mono mono
```
SMT, components
The translation model $P(f|e)$

cluster:/home/moses/model> zmore phrase-table.gz

be consistent ||| coherentes ||| 0.0384615 0.146893 0.0833333 0.0116792 2.718 ||| 1-0 ||| 26 12
be consistent ||| sean coherentes ||| 0.2 0.00022714 0.0833333 0.0916808 2.718 ||| 0-0 1-1 ||| 5 12
be consistent ||| sean consistentes ||| 0.5 0.000104834 0.0833333 0.0785835 2.718 ||| 0-0 1-1 ||| 2 12
be consistent ||| ser coherente ||| 0.5 0.0204044 0.166667 0.569957 2.718 ||| 0-0 1-1 ||| 2 12
be consistent ||| ser consecuente ||| 1 0.000340072 0.0833333 0.759942 2.718 ||| 0-0 1-1 ||| 1 12
be consistent ||| ser consistente ||| 1 0.00850183 0.5 0.633285 2.718 ||| 0-0 1-1 ||| 6 12
consistent when ||| coherente cuando se ||| 1 0.00783857 1 0.329794 2.718 ||| 0-0 1-1 1-2 ||| 1 1
consistent ||| adecuado ||| 0.00512821 0.0112994 0.00671141 0.009009 2.718 ||| 0-0 ||| 195 149
consistent ||| coherencia ||| 0.137931 0.0282486 0.0268456 0.0847458 2.718 ||| 0-0 ||| 29 149
consistent ||| constante ||| 0.0333333 0.0112994 0.0134228 0.0307692 2.718 ||| 0-0 ||| 60 149
consistent ||| constantes ||| 0.0625 0.0056497 0.00671141 0.047619 2.718 ||| 0-0 ||| 16 149
...
Translation model: keep in mind

- Statistical TMs estimate the probability of a translation from a parallel aligned corpus.
- Its quality depends on the quality of the obtained word (phrase) alignments.
- Within an SMT system, it contributes to select semantically adequate sentences in the target language.
Decoder

\[ T(f) = \hat{e} = \arg\max_e P(e) P(f|e) \]

Responsible for the search in the space of possible translations.

Given a model (LM+TM+...), the decoder constructs the possible translations and looks for the most probable one.

In our context, one can find:

- Greedy decoders. Initial hypothesis (word by word translation) refined iteratively using hill-climbing heuristics.
- Beam search decoders.
$T(f) = \hat{e} = \arg\max_e P(e) P(f|e)$

Responsible for the search in the space of possible translations.

Given a model (LM+TM+...), the decoder constructs the possible translations and looks for the most probable one.

In our context, one can find:

- **Greedy decoders.** Initial hypothesis (word by word translation) refined iteratively using hill-climbing heuristics.
- **Beam search decoders.**
Decoder

\[ T(f) = \hat{e} = \arg\max_e P(e) P(f|e) \]

Responsible for the search in the space of possible translations.

Given a model \((LM+TM+...\)), the decoder constructs the possible translations and looks for the most probable one.

In our context, one can find:

- **Greedy decoders.** Initial hypothesis (word by word translation) refined iteratively using hill-climbing heuristics.
- **Beam search decoders.** Let’s see..
Core algorithm

1. Collect translation options
2. Initial state: empty hypothesis
3. Expand hypotheses with all translation options
4. Estimate the cost for each hypothesis
5. Check if all source words are covered
   - If NO, go back to expanding hypotheses
   - If YES, return translation: hypothesis with the lowest cost
Example: Quan tornes a casa

Translation options:

(Quan, When)
(Quan_tornes, When_are_you_coming_back)
(Quan_tornes_a_casa, When_are_you_coming_back_home)
(tornes, come_back)
(tornes_a_casa, come_back_home)
(a_casa, home)

Notation for hypotheses in construction:

Constructed sentence so far: come_back
Source words already translated: - x - -
Example: Quan tornes a casa

- Translation options:
  - (Quan, When)
  - (Quan_tornes, When_are_you_coming_back)
  - (Quan_tornes_a_casa, When_are_you_coming_back_home)
  - (tornes, come_back)
  - (tornes_a_casa, come_back_home)
  - (a_casa, home)

- Notation for hypotheses in construction:
  - Constructed sentence so far: come_back
  - Source words already translated: - X - -
Example: Quan tornes a casa

- Translation options:
  
  (Quan, When)
  (Quan_tornes, When_are_you_coming_back)
  (Quan_tornes_a_casa, When_are_you_coming_back_home)
  (tornes, come_back)
  (tornes_a_casa, come_back_home)
  (a_casa, home)

- Initial hypothesis
  
  Constructed sentence so far: $\phi$
  Source words already translated: ---
SMT, components
Decoding

\phi
\ldots
SMT, components
Decoding

When
x - - -

When_are_you_coming_back
x x - -

When_are_you_coming_back_home
* x x x*

come_back
- x - -

come_back_home
- x x x

home
-- x x
SMT, components
Decoding

\( \phi \)

\[
\begin{align*}
\text{When} & \mid \\
\text{x} & \mid \\
\text{When\_are\_you\_coming\_back} & \mid \\
\text{x} & \mid \\
\text{When\_are\_you\_coming\_back\_home} & \mid \\
\text{*x} & \mid \text{x} & \text{x} & \text{x} & \text{x} & \\
\text{come\_back} & \mid \\
\text{x} & \mid \\
\text{come\_back\_home} & \mid \\
\text{x} & \mid \text{x} & \text{x} & \\
\text{home} & \mid \\
\text{x} & \mid \text{x} & \text{x} & \\
\end{align*}
\]
When are you coming back home
* x x x x*

When come back
x x -

When are you coming back
x x -

When are you coming back home
* x x x x*

come back
- x -

come back home
- x x x

home
- - x x
SMT, components

Decoding

\[ \phi \]

- - - -

When|come_back|home

* x x x *

When|come_back

x - - -

When_are_you_coming_back|come_back

x x - -

When_are_you_coming_back_home

* x x x *

come_back|

- x - -

come_back_home|

- x x x

home|

- - x x
When|come_back_home

When|come_back

When_are_you_coming_back

When_are_you_coming_back_home

come_back|home

come_back|when

come_back_home

home|come_back

home|when
Exhaustive search

- As a result, one should have an estimation of the cost of each hypothesis, being the lowest cost one the best translation.

But...

- The number of hypotheses is exponential with the number of source words.
  (30 words sentence ⇒ $2^{30} = 1,073,741,824$ hypotheses!)

Solution

- Optimise the search by:
  - Hypotheses recombination
  - Beam search and pruning
Exhaustive search

- As a result, one should have an estimation of the cost of each hypothesis, being the lowest cost one the best translation.

But...

- The number of hypotheses is exponential with the number of source words.
  (30 words sentence $\Rightarrow 2^{30} = 1,073,741,824$ hypotheses!)

Solution

- Optimise the search by:
  - Hypotheses recombination
  - Beam search and pruning
Exhaustive search

- As a result, one should have an estimation of the cost of each hypothesis, being the lowest cost one the best translation.

But...

- The number of hypotheses is exponential with the number of source words.
  
  (30 words sentence ⇒ $2^{30} = 1,073,741,824$ hypotheses!)

Solution

- Optimise the search by:
  - Hypotheses recombination
  - Beam search and pruning
Hypotheses recombination

Combine hypotheses with the same source words translated, keep that with a lower cost.

- Risk-free operation. The lowest cost translation is still there.
- But the space of hypothesis is not reduced enough.
Hypotheses recombination

Combine hypotheses with the same source words translated, keep that with a lower cost.

\[
\text{When|come\_back\_home} \quad \leftrightarrow \quad \text{When|come\_back\_home}
\]

- Risk-free operation. The lowest cost translation is still there.
- But the space of hypothesis is not reduced enough.
Hypotheses recombination

Combine hypotheses with the same source words translated, keep that with a lower cost.

- **Risk-free operation.** The lowest cost translation is still there.
- But the space of hypothesis is not reduced enough.
Beam search and pruning (at last!)

Compare hypotheses with the same number of translated source words and prune out the inferior ones.

What is an inferior hypothesis?

- The quality of a hypothesis is given by the cost so far and by an estimation of the future cost.
- Future cost estimations are only approximate, so the pruning is not risk-free.
Beam search and pruning (at last!)

Strategy:

- Define a **beam size** (by threshold or number of hypotheses).
- **Distribute** the hypotheses being generated in stacks according to the number of translated source words, for instance.
- **Prune out** the hypotheses falling outside the beam.
- The hypotheses to be pruned are those with a higher (current + future) cost.
SMT, components

Decoder

Decoding: keep in mind

- Standard SMT decoders translate the sentences from left to right by expanding hypotheses.
- Beam search decoding is one of the most efficient approach.
- But, the search is only approximate, so, the best translation can be lost if one restricts the search space too much.
Outline

1. Introduction
2. Basics
3. Components
4. The log-linear model
5. Beyond standard SMT
SMT, the log-linear model

Motivation

**Maximum likelihood (ML)**

\[
\hat{e} = \arg\max_e P(e|f) = \arg\max_e P(e)P(f|e)
\]

**Maximum entropy (ME)**

\[
\hat{e} = \arg\max_e P(e|f) = \arg\max_e \exp\left\{\sum \lambda_m h_m(f, e)\right\}
\]

\[
\hat{e} = \arg\max_e \log P(e|f) = \arg\max_e \sum \lambda_m h_m(f, e)
\]

Log-linear model
SMT, the log-linear model

Motivation

**Maximum likelihood (ML)**

\[ \hat{e} = \arg\max_e P(e|f) = \arg\max_e P(e) P(f|e) \]

**Maximum entropy (ME)**

\[ \hat{e} = \arg\max_e P(e|f) = \arg\max_e \exp \left\{ \sum \lambda_m h_m(f, e) \right\} \]

\[ \hat{e} = \arg\max_e \log P(e|f) = \arg\max_e \sum \lambda_m h_m(f, e) \]

**Log-linear model**
SMT, the log-linear model

Motivation

**Maximum likelihood (ML)**

\[ \hat{e} = \arg \max_e P(e|f) = \arg \max_e \ P(e) \ P(f|e) \]

**Maximum entropy (ME)**

\[ \hat{e} = \arg \max_e P(e|f) = \arg \max_e \ \exp \left\{ \sum \lambda_m h_m(f, e) \right\} \]

\[ \hat{e} = \arg \max_e \log P(e|f) = \arg \max_e \ \sum \lambda_m h_m(f, e) \]

Log-linear model
SMT, the log-linear model

Motivation

Maximum likelihood (ML)

\[ \hat{e} = \arg\max_e P(e|f) = \arg\max_e P(e) P(f|e) \]

Maximum entropy (ME)

\[ \hat{e} = \arg\max_e \log P(e|f) = \arg\max_e \sum \lambda_m h_m(f, e) \]

Log-linear model with

\[ h_1(f, e) = \log P(e), \ h_2(f, e) = \log P(f|e), \ \text{and} \ \lambda_1 = \lambda_2 = 1 \]

\[ \Rightarrow \text{Maximum likelihood model} \]
What can be achieved with the log-linear model (as compared to maximum likelihood model)

- Extra features $h_m$ can be easily added...
- ... but their weight $\lambda_m$ must be somehow determined.
- Different knowledge sources can be used.
SMT, the log-linear model

Features

Standard feature functions

Eight features are usually used: \( P(e), P(f|e), P(e|f), \text{lex}(f|e), \text{lex}(e|f), ph(e), w(e) \) and \( P_d(e,f) \).

- Language model \( P(e) \)
  \( P(e) \): Language model probability as in ML model.

- Translation model \( P(f|e) \)
  \( P(f|e) \): Translation model probability as in ML model.

- Translation model \( P(e|f) \)
  \( P(e|f) \): Inverse translation model probability to be added to the generative one.
SMT, the log-linear model

Features

**Standard feature functions**

Eight features are usually used: $P(e)$, $P(f|e)$, $P(e|f)$, $\text{lex}(f|e)$, $\text{lex}(e|f)$, $\text{ph}(e)$, $w(e)$ and $P_d(e,f)$.

- **Translation model** $\text{lex}(f|e)$
  
  $\text{lex}(f|e)$: Lexical translation model probability.

- **Translation model** $\text{lex}(e|f)$
  
  $\text{lex}(e|f)$: Inverse lexical translation model probability.

- **Phrase penalty** $\text{ph}(e)$
  
  $\text{ph}(e)$: A constant cost per produced phrase.
SMT, the log-linear model

Features

Standard feature functions

Eight features are usually used: \( P(e) \), \( P(f|e) \), \( P(e|f) \), \( \text{lex}(f|e) \), \( \text{lex}(e|f) \), \( \text{ph}(e) \), \( w(e) \) and \( P_d(e, f) \).

- **Word penalty** \( w(e) \)

  \( w(e) \): A constant cost per produced word.

- **Distortion** \( P_d(e, f) \)

  \( P_d(\text{ini}\_\text{phrase}_i, \text{end}\_\text{phrase}_{i-1}) \): Relative distortion probability distribution. A simple distortion model:

  \[
P_d(\text{ini}\_\text{phrase}_i, \text{end}\_\text{phrase}_{i-1}) = \alpha |\text{ini}\_\text{phrase}_i - \text{end}\_\text{phrase}_{i-1} - 1|
  \]
SMT, components

The translation model $P(f|e)$

In practice,

cluster:/home/moses/model> zmore phrase-table.gz

```plaintext
be consistent ||| coherentes ||| 0.0384615 0.146893 0.0833333 0.0116792 2.718 ||| 1-0 ||| 26 12
be consistent ||| sean coherentes ||| 0.2 0.00022714 0.0833333 0.0916808 2.718 ||| 0-0 1-1 ||| 5 12
be consistent ||| sean consistentes ||| 0.5 0.000104834 0.0833333 0.0785835 2.718 ||| 0-0 1-1 ||| 2 12
be consistent ||| ser coherente ||| 0.5 0.0204044 0.166667 0.569957 2.718 ||| 0-0 1-1 ||| 4 12
be consistent ||| ser consecuente ||| 1 0.000340072 0.0833333 0.759942 2.718 ||| 0-0 1-1 ||| 1 12
be consistent ||| ser consistente ||| 1 0.00850183 0.5 0.633285 2.718 ||| 0-0 1-1 ||| 6 12
consistent when ||| coherente cuando se ||| 1 0.00783857 1 0.329794 2.718 ||| 0-0 1-1 1-2 ||| 1 1
consistent ||| adecuado ||| 0.00512821 0.0112994 0.00671141 0.009009 2.718 ||| 0-0 ||| 195 149
consistent ||| coherencia ||| 0.137931 0.0282486 0.0268456 0.0847458 2.718 ||| 0-0 ||| 29 149
consistent ||| constante ||| 0.0333333 0.0112994 0.0134228 0.0307692 2.718 ||| 0-0 ||| 60 149
consistent ||| constantes ||| 0.0625 0.0056497 0.00671141 0.047619 2.718 ||| 0-0 ||| 16 149
...```
State of the art?

Software such as Moses makes easy the incorporation of more sophisticated reordering.

From a distance-based reordering (1 feature)

to include orientation information in a lexicalised reordering. (3-6 features)
From where and how can one learn reorders?

<table>
<thead>
<tr>
<th>When</th>
<th>are</th>
<th>you</th>
<th>coming</th>
<th>back</th>
<th>home</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quan</td>
<td>tornes</td>
<td>tu</td>
<td>a</td>
<td>casa</td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>

(are, tornes, monotone)
From where and how can one learn reorders?

<table>
<thead>
<tr>
<th>When</th>
<th>coming back</th>
<th>tornes</th>
<th>tu</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(coming back, tornes, swap)
SMT, the log-linear model

Digression: lexicalised reordering or distortion

From where and how can one learn reorders?

<table>
<thead>
<tr>
<th></th>
<th>Quan</th>
<th>tornes</th>
<th>tu</th>
<th>a</th>
<th>casa</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(home ?, casa ?, discontinuous)
SMT, the log-linear model
Digression: lexicalised reordering or distortion

3 new features estimated by frequency counts:
\( P_{\text{monotone}}, P_{\text{swap}} \) and \( P_{\text{discontinuous}} \) (6 when bidirectional).

\[
P_{\text{or.}}(\text{orientation}|f, e) = \frac{\text{count}(\text{orientation}, e, f)}{\sum_{\text{or.}} \text{count}(\text{orientation}, e, f)}
\]

- Sparse statistics of the orientation types → smoothing.
- Several variations.
In practice,

cluster:/home/moses/model> zmore extract.o.gz

resumption ||| reanudacion ||| mono mono
resumption of the ||| reanudacion del ||| mono mono
resumption of the session ||| reanudacion del periodo de sesiones ||| mono mono
de la union ||| union ‘s ||| swap swap
competencia de la union ||| union ‘s competition ||| swap other
...

cluster:/home/moses/model> zmore reordering-table.wbe-msd-bidirectional-fe.gz

a resumption of the s ||| se reanudara el periodo de s ||| 0.200 0.200 0.600 0.600 0.200 0.200
resumption of the s ||| reanudacion del periodo de s ||| 0.995 0.002 0.002 0.995 0.002 0.002
the resumption of the s ||| la continuacion del periodo de s ||| 0.142 0.142 0.714 0.714 0.142 0.142
the resumption of the s ||| la reanudacion del periodo de s ||| 0.818 0.090 0.090 0.818 0.090 0.090
SMT, components
The translation model $P(f|e)$

```
cluster:/home/moses/model> wc -l *

493,896,818 phrase-table
493,896,818 reordering-table.wbe-msd-bidirectional-fe
```

```
cluster:/home/moses/model> ls -lkh *

-rw-r--r-- 1 emt ia 57G mar 3 14:01 phrase-table
-rw-r--r-- 1 emt ia 55G mar 3 14:08 reordering-table.wbe-msd-bidirectional-fe
```
SMT, the log-linear model

Features

Standard feature functions

13 features may be used:

- $P(e)$;
- $P(f|e)$, $P(e|f)$, $\text{lex}(f|e)$, $\text{lex}(e|f)$;
- $ph(e)$, $w(e)$;
- $P_{mon}(o|e,f)$, $P_{swap}(o|e,f)$, $P_{dis}(o|e,f)$,
- $P_{mon}(o|f,e)$, $P_{swap}(o|f,e)$, $P_{dis}(o|f,e)$. 
Development training, weights optimisation

- Supervised training: a (small) aligned parallel corpus is used to determine the optimal weights.

\[ \hat{e} = \arg\max_e \log P(e|f) = \arg\max_e \sum \lambda_m h_m(f, e) \]
Development training, weights optimisation

Strategies

- **Generative training.** Optimises ME objective function which has a unique optimum. Maximises the likelihood.

- **Discriminative training** only for feature weights (not models), or purely discriminative for the model as a whole. This way translation performance can be optimised.

- Minimum Error-Rate Training (MERT).
Development training, weights optimisation

Strategies

- **Generative training.** Optimises ME objective function which has a unique optimum. Maximises the likelihood.

- **Discriminative training** only for feature weights (not models), or purely discriminative for the model as a whole. This way translation performance can be optimised.

- **Minimum Error-Rate Training (MERT).**
Minimum Error-Rate Training

- Approach: Minimise an error function.

But... what’s the error of a translation?

- There exist several error measures or metrics.
- Metrics not always correlate with human judgements.
- The quality of the final translation on the metric chosen for the optimisation is shown to improve.
- For the moment, let’s say we use BLEU.

(More on MT Evaluation section)
Minimum Error-Rate Training rough algorithm

1. $\lambda_i$ initialisation
2. Select a direction $k$, fix $\lambda_i \neq \lambda_k$
3. Best $\lambda_k$ (line minimisation)
4. all $k$ explored?
   - NO
   - YES: $\lambda_i$ stable?
     - NO
     - YES: Optimal $\lambda_i$
SMT, the log-linear model

Minimum Error-Rate Training (MERT)

Powell’s method (2D: $\lambda_1, \lambda_2$)
SMT, the log-linear model
Minimum Error-Rate Training (MERT)

**Powell’s method (2D: $\lambda_1, \lambda_2$)**
SMT, the log-linear model
Minimum Error-Rate Training (MERT)

Powell’s method (2D: $\lambda_1, \lambda_2$)
SMT, the log-linear model
Minimum Error-Rate Training (MERT)

Powell’s method (2D: $\lambda_1, \lambda_2$)
SMT, the log-linear model
Minimum Error-Rate Training (MERT)

Powell’s method (2D: $\lambda_1, \lambda_2$)
In practice,

# language model weights
[weight-l]
0.102111

# translation model weights
[weight-t]
0.0146796
0.0281078
0.0501881
0.087537
0.128371

# word penalty
[weight-w]
-0.142732
The log-linear model allows to include several weighted features. Standard systems use 8 (13) real features.

The corresponding weights are optimised on a development set, a small aligned parallel corpus.

An optimisation algorithm such as MERT is appropriate for about a dozen of features. For more features, purely discriminative learnings should be used.

For MERT, the choice of the metric that quantifies the error in the translation is an issue.
Word alignment with...

GIZA++
https://code.google.com/p/giza-pp

The Berkeley WordAligner
https://code.google.com/p/berkeleyaligner

Fast Align
https://github.com/clab/fast_align

...
Phrase-based SMT systems
Tools & Choices

Language Model with...

SRILM
http://www.speech.sri.com/projects/srilm

IRSTLM
http://sourceforge.net/projects/irstlm

RandLM
http://sourceforge.net/projects/randlm

KenLM
http://kheafield.com/code/kenlm

...
Try parameter optimisation with...

MERT
Minimum error rate training, Och (2003)

PRO
Pairwise ranked optimization, Hopkins and May (2011)

MIRA
Margin Infused Relaxed Algorithm, Hasler et al. (2011)

...
Decoding with...

Moses
http://www.statmt.org/moses

Phrasal
http://nlp.stanford.edu/software/phrasal

Docent
https://github.com/chardmeier/docent
Outline

1. Introduction
2. Basics
3. Components
4. The log-linear model
5. Beyond standard SMT
   - Factored translation models
   - Syntactic translation models
   - Ongoing research
Considering **linguistic information** in phrase-based models

- Phrase-based log-linear models do not consider linguistic information other than words. This is information should be included.

**Options**

- Use syntactic information as pre- or post-process (for reordering or reranking for example).
- Include linguistic information in the model itself.
  - Factored translation models.
  - Syntactic-based translation models.
Factored translation models

Extension to phrase-based models where every word is substituted by a vector of factors.

\[(\text{word}) \rightarrow (\text{word}, \text{lemma}, \text{PoS}, \text{morphology}, ...)\]

The translation is now a combination of pure translation (T) and generation (G) steps:
SMT, beyond standard SMT
Factored translation models

Factored translation models
Extension to phrase-based models where every word is substituted by a vector of factors.

\[(\text{word}) \implies (\text{word}, \text{lemma}, \text{PoS}, \text{morphology}, ...)\]

The translation is now a combination of pure translation (T) and generation (G) steps:

\[\begin{align*}
\text{lemma}_f & \downarrow T \\
\text{PoS}_f & \downarrow T \\
\text{morphology}_f & \downarrow T \\
\text{word}_f & \\
\text{lemma}_e & \downarrow T \\
\text{PoS}_e & \downarrow T \\
\text{morphology}_e & \downarrow T \\
\text{word}_e & \end{align*}\]
Factored translation models

Extension to phrase-based models where every word is substituted by a vector of factors.

\[(\text{word}) \rightarrow (\text{word, lemma, PoS, morphology, ...})\]

The translation is now a combination of pure translation (T) and generation (G) steps:

\[
\begin{align*}
\text{casa}_f & \quad \text{NN}_f & \quad \text{fem., plural}_f & \quad \text{cases}_f \\
\downarrow T & \quad \downarrow T & \quad \downarrow T & \quad \text{G} \\
\text{house}_e & \quad \text{NN}_e & \quad \text{plural}_e & \quad \text{houses}_e
\end{align*}
\]
SMT, beyond standard SMT
Factored translation models

What differs in factored translation models
(as compared to standard phrase-based models)

- The parallel corpus must be annotated beforehand.
- Extra language models for every factor can also be used.
- Translation steps are accomplished in a similar way.
- Generation steps imply a training only on the target side of the corpus.
- Models corresponding to the different factors and components are combined in a log-linear fashion.
SMT, beyond standard SMT
Syntactic translation models

Syntactic translation models
Incorporate syntax to the source and/or target languages.

Approaches

- Syntactic phrase-based based on tree trasducers:
  - Tree-to-string. Build mappings from target parse trees to source strings.
  - String-to-tree. Build mappings from target strings to source parse trees.
  - Tree-to-tree. Mappings from parse trees to parse trees.
SMT, beyond standard SMT
Syntactic translation models

**Syntactic translation models**
Incorporate syntax to the source and/or target languages.

**Approaches**

- Synchronous grammar formalism which learns a grammar that can simultaneously generate both trees.
  - Syntax-based. Respect linguistic units in translation.
Syntactic models ease reordering. An intuitive example:

*En David llegeix un llibre nou*

```latex
S
  | V
  | llegeix
 NP
  | DT
  | NN
  | En David

S
  | V
  | VP
  | DT
  | NN
  | PP
  | un
  | llibre
  | nou

S
  | V
  | VP
  | DT
  | PP
  | NN
```
Syntactic models ease reordering. An intuitive example:

En David llegeix un llibre nou
Syntactic models ease reordering. An intuitive example:

En David llegeix un llibre nou
Syntactic models ease reordering. An intuitive example:

En David llegeix un llibre nou
Syntactic models ease reordering. An intuitive example:

En David llegeix un llibre nou

David reads a new book
SMT, beyond standard SMT

Ongoing research

**Hot research topics**

Current research on SMT addresses known and new problems.

Some *components* of the standard phrase-based model are still under study:

- Automatic alignments.
- Language models and smoothing techniques.
- Parameter optimisation.
Complements to a standard system can be added:

- Reordering as a pre-process or post-process.
- Reranking of n-best lists.
- OOV treatment.
- Domain adaptation.
Development of full systems from scratch or modifications to the standard:

- Using machine learning.
- Including linguistic information.
- Hybridation of MT paradigms.

Or a different strategy:

- Systems combination.
Factored models include linguistic information in phrase-based models and are suitable for morphologically rich languages.

Syntactic models consider somehow syntax and are adequate for language pairs with a different structure of the sentences.

Current research addresses both new models and modifications to the existing ones.
Part II

MT Evaluation
Outline

6 MT Evaluation basics

- Automatic Evaluation
- BLEU
- Limits of lexical similarity
MT Evaluation
Importance for system development

Evaluation methods

Error detection
Error analysis
Refinement
Implementation
Test

OK? YES NO

Unfruitful results
MT Evaluation
Importance for system development

- Error detection
- Error analysis
- Refinement
- Implementation
- Test

Evaluation methods

Unfruitful results
MT Evaluation
Importance for system development

Evaluation methods

Error detection
Error analysis
Refinement
Implementation
Test

Unfruitful results

OK? YES NO
MT Evaluation
Importance for system development

Evaluation methods

Error detection
Error analysis
Refinement
Implementation
Test

OK? YES NO

Unfruitful results
MT Evaluation
Importance for system development

Error detection
Error analysis
Refinement
Implementation
Test

Evaluation methods
Unfruitful results

OK?

YES
NO
MT Evaluation
Importance for system development

Evaluation methods

Error detection
Error analysis
Refinement
Implementation
Test

OK?

Unfruitful results

Evaluation methods
MT Evaluation

Importance for system development

Error detection -> Error analysis
Error analysis -> Refinement
Refinement -> Implementation
Implementation -> Test

Evaluation methods

Unfruitful results

OK? YES NO
MT Evaluation
Automatic evaluation

What can achieved with automatic evaluation
(as compared to manual evaluation)

- Automatic metrics notably accelerate the development cycle of MT systems:
  - Error analysis
  - System optimisation
  - System comparison

Besides, they are

- Costless (vs. costly)
- Objective (vs. subjective)
- Reusable (vs. non-reusable)
MT Evaluation
Lexical similarity

**Metrics based on lexical similarity**
(most of the metrics!)

- **Edit Distance**: WER, PER, TER
- **Precision**: BLEU, NIST, WNM
- **Recall**: ROUGE, CDER
- **Precision/Recall**: GTM, METEOR, BLANC, SIA

Nowadays, BLEU is accepted as the standard metric.
MT Evaluation
Lexical similarity

Metrics based on lexical similarity
(most of the metrics!)

- **Edit Distance**: WER, PER, TER
- **Precision**: BLEU, NIST, WNM
- **Recall**: ROUGE, CDER
- **Precision/Recall**: GTM, METEOR, BLANC, SIA

Nowadays, BLEU is accepted as *the standard* metric.
The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.
Candidate 1:

It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2:

It is to insure the troops forever hearing the activity guidebook that party direct.
Candidate 1:
It is a guide to action which ensures that the military always obeys the commands of the party.

Reference 1:
It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:
It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3:
It is the practical guide for the army always to heed the directions of the party.
MT Evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Candidate 1:

It is a guide to action which ensures that the military always obeys the commands of the party.

Reference 1:

It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:

It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3:

It is the practical guide for the army always to heed the directions of the party.
MT Evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Candidate 2:

It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1:

It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:

It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3:

It is the practical guide for the army always to heed the directions of the party.
Modified n-gram precision (1-gram)

Precision-based measure, but:

Candidate:
   The the the the the the the the.

Reference 1:
   The cat is on the mat.

Reference 2:
   There is a cat on the mat.
Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{1 + 7}{7} \]

Candidate:
\[
\text{The the the the the the the the the.}
\]

Reference 1:
\[
\text{The cat is on the mat.}
\]

Reference 2:
\[
\text{There is a cat on the mat.}
\]
Modified n-gram precision (1-gram)

Precision-based measure, but:

\[ \text{Prec.} = \frac{2 + 7}{7} \]

Candidate:

\text{The the the the the the the the.}

Reference 1:

\text{The cat is on the mat.}

Reference 2:

\text{There is a cat on the mat.}
**Modified n-gram precision** (1-gram)

Precision-based measure, but:  

$$\text{Prec.} = \frac{3 + 7}{7}$$

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{4 + 7}{7}$

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
**Modified n-gram precision** (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{5 + 7}{7} \]

Candidate:

\textbf{The the the the the the the.}

Reference 1:

\textbf{The cat is on the mat.}

Reference 2:

\textbf{There is a cat on the mat.}
Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{6 + 7}{7} \]

Candidate:

The the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
MT Evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{7}{7} \]

Candidate:

\text{The the the the the the the the the.}

Reference 1:

\text{The cat is on the mat.}

Reference 2:

\text{There is a cat on the mat.}
Modified n-gram precision (1-gram)

A reference word should only be matched once.

Algorithm:

1. Count number of times $w_i$ occurs in each reference.
2. Keep the minimum between the maximum of (1) and the number of times $w_i$ appears in the candidate (clipping).
3. Add these values and divide by candidate’s number of words.
Modified n-gram precision (1-gram)

Modified 1-gram precision:

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1 $w_i \rightarrow$ The
   \#$_{w_i,R1} = 2$
   \#$_{w_i,R2} = 1$

2 Max$_{(1)}$=2, \#$_{w_i,C} = 7$
   $\Rightarrow$ Min=2

3 No more distinct words
Modified n-gram precision (1-gram)

Modified 1-gram precision: \[ P_1 = \]

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1. \( w_i \rightarrow \) The
   \( \#w_{i,R1} = 2 \)
   \( \#w_{i,R2} = 1 \)

2. \( \text{Max}(1) = 2, \#w_i,c = 7 \)
   \( \Rightarrow \text{Min} = 2 \)

3. No more distinct words
Modified n-gram precision (1-gram)

Modified 1-gram precision: \( P_1 = \frac{2}{7} \)

Candidate:

The the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1. \( w_i \rightarrow \text{The} \)
   \( \# w_i, R_1 = 2 \)
   \( \# w_i, R_2 = 1 \)

2. \( \text{Max}_{(1)} = 2, \# w_i, C = 7 \)
   \( \Rightarrow \text{Min} = 2 \)

3. No more distinct words
Modified n-gram precision (1-gram)

Modified 1-gram precision:

\[ P_1 = \frac{2}{7} \]

Candidate:

The the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1. \( w_i \rightarrow \text{The} \)
   \#_{w_i,R1} = 2
   \#_{w_i,R2} = 1

2. \( \text{Max}_{(1)} = 2, \#_{w_i,C} = 7 \)
   \Rightarrow \text{Min}=2

3. No more distinct words
Modified n-gram precision

- Straightforward generalisation to \( n \)-grams, \( P_n \).
- Generalisation to multiple sentences:

\[
P_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{\text{ngram} \in C} \text{Count}_{\text{clipped}}(\text{ngram})}{\sum_{C \in \{\text{candidates}\}} \sum_{\text{ngram} \in C} \text{Count}(\text{ngram})}
\]

low \( n \) adequacy  
high \( n \) fluency
Brevity penalty

Candidate:
  of the

Reference 1:
  It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:
  It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3:
  It is the practical guide for the army always to heed the directions of the party.
Brevity penalty

Candidate:

of the

\[ P_1 = \frac{2}{2}, \ P_2 = \frac{1}{1} \]

Reference 1:
It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2:
It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3:
It is the practical guide for the army always to heed the directions of the party.
Brevity penalty

\[ BP = \begin{cases} 
1 & \text{if } c > r \\
\frac{e^{1-r/c}}{c} & \text{if } c \leq r 
\end{cases} \]

- \( c \) candidate length, \( r \) reference length

- Multiplicative factor.
- At sentence level, huge punishment for short sentences.
- Estimated at document level.
BiLingual Evaluation Understudy, BLEU

\[
\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^{N} w_n \log P_n \right)
\]

- Geometric average of \( P_n \) (empirical suggestion).
- \( w_n \) positive weights summing to one.
- Brevity penalty.
MT Evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Paper’s Conclusions

- BLEU correlates with human judgements.
- It can distinguish among similar systems.
- Need for multiple references or a big test with heterogeneous references.
- More parametrisation in the future.
Watch out with BLEU implementations!

There are several widely used implementations of BLEU. (Moses multi-bleu.perl script, NIST mteval-vXX.pl script, etc.)

Results differ because of:

- Different tokenisation approach.
- Different definition of *closest reference* in the brevity penalty estimation.
NIST is based on BLEU but:

- Arithmetic average of $n$-gram counts rather than a geometric average.
- Informative $n$-grams are given more weight.
- Different definition of brevity penalty.
**Limits of lexical similarity**

The reliability of lexical metrics depends very strongly on the heterogeneity/representativeness of reference translations.

\[ e: \text{This sentence is going to be difficult to evaluate.} \]

\[ \text{Ref1: The evaluation of the clause is complicated.} \]
\[ \text{Ref2: The sentence will be hard to qualify.} \]
\[ \text{Ref3: The translation is going to be hard to evaluate.} \]
\[ \text{Ref4: It will be difficult to punctuate the output.} \]

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e: This sentence is going to be difficult to evaluate.

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e: This sentence is going to be difficult to evaluate.

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
**Recent efforts** to go over lexical similarity

**Extend** the reference material:

- Using **lexical variants** such as morphological variations or synonymy lookup or using **paraphrasing** support.

Compare other **linguistic features** than words:

- Syntactic similarity: shallow parsing, full parsing (constituents /dependencies).
- Semantic similarity: named entities, semantic roles, discourse representations.

**Combination** of the existing metrics.
Recent efforts to go over lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- **Syntactic** similarity: shallow parsing, full parsing (constituents/dependencies).
- **Semantic** similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.
Recent efforts to go over lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- Syntactic similarity: shallow parsing, full parsing (constituents /dependencies).
- Semantic similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.
MT Evaluation
Towards Heterogeneous Automatic MT Evaluation
MT Evaluation
Towards Heterogeneous Automatic MT Evaluation
Asiya has been designed to assist both system and metric developers by offering a rich repository of metrics and meta-metrics.

http://nlp.lsi.upc.edu/asiya/
MT Evaluation

Summary

MT Evaluation: keep in mind

- Evaluation is important in the system development cycle. Automatic evaluation accelerates significantly the process.

- Up to now, most (common) metrics rely on lexical similarity, but it cannot assure a correct evaluation.

- Current work is being devoted to go beyond lexical similarity.
Part III

SMT experiments
Outline Part II

7 Translation system
- Demos
- Software
- Steps
- Software
- Steps
- Demo
SMT system
Demo: http://demo.statmt.org/

Moses Machine Translation Demo

Source:
Hello, I want to translate my first sentence into German.

Looking to translate a web page? Then click here

This site is maintained by the Machine Translation Group at the University of Edinburgh.
SMT system
Demo: http://sz.ru/smt/

Введите одно английское предложение или фразу.
Hello, I want to translate my first sentence into Russian.

Перевести одно предложение.

Sergey Protasov
Build your own SMT system

1. Language model with SRILM.

2. Word alignments with GIZA++.
   http://code.google.com/p/giza-pp/downloads/list

3. And everything else with the Moses package.
   https://github.com/moses-smt/mosesdecoder
1. Download and prepare your data

Parallel corpora and some tools can be downloaded for instance from the WMT 2013 web page:
http://www.statmt.org/wmt13/translation-task.html

How to construct a baseline system is also explained there:
http://www.statmt.org/wmt10/baseline.html

We continue with the Europarl corpus Spanish-to-English.
1. Download and prepare your data (cont’d)

2. Tokenise the corpus with WMT10 scripts. (training corpus and development set for MERT)


1. **Download and prepare your data** (cont’d)

3. Filter out long sentences with Moses scripts.  
   (Important for GIZA++)

   ```
   bin/moses-scripts/training/clean-corpus-n.perl eurov4.es-en.TOK es en eurov4.es-en.TOK.clean 1 100
   ```

4. Lowercase training and development with WMT10 scripts.  
   (Optional but recommended)

   ```
   ```
2. Build the language model

1. Run SRILM on the English part of the parallel corpus or on a monolingual larger one. (tokenise and lowercase in case it is not)

   ngram-count -order 5 -interpolate -kndiscount -text eurov4.es-en.en -lm eurov4.en.lm
3. Train the translation model

1. Use the Moses script `train-model.perl`
   This script performs the whole training:

   `train-model.perl --help`

   **Train Phrase Model**

   **Steps:**
   (1) prepare corpus
   (2) run GIZA
   (3) align words
   (4) learn lexical translation
   (5) extract phrases
   (6) score phrases
   (7) learn reordering model
   (8) learn generation model
   (9) create decoder config file
3. Train the translation model (cont’d)

So, it takes a few arguments (and a few time!):

```
moses-scripts/training/train-model.perl -scripts-root-dir
bin/moses-scripts/ -root-dir working-dir -corpus eurov4.es-en -f es -e en
-alignment grow-diag-final-and -reordering msd-bidirectional-fe
-lm 0:5:eurov4.en.lm:0
```

It generates a configuration file moses.ini needed to run the decoder where all the necessary files are specified.
### 4. Tuning of parameters with MERT

1. Run the Moses script `mert-moses.pl`  
   (Another slow step!)

   ```bash
   moses-scripts/training/mert-moses.pl eurov4.es-en.dev.es eurov4.es-en.dev.en mosesdecoder/bin/moses ./model/moses.ini --working-dir ./tuning --rootdir bin/moses-scripts/
   ```

2. Insert weights into configuration file with WMT10 script:

   ```bash
   wmt10scripts/reuse-weights.perl ./tuning/moses.ini < ./model/moses.ini > moses.weight-reused.ini
   ```
5. **Run Moses decoder on a test set**

1. Tokenise and lowercase the test set as before.

2. Filter the model with Moses script.
   *(mandatory for large translation tables)*
   
   moses-scripts/training/filter-model-given-input.pl ./filteredmodel
   moses.weight-reused.ini testset.es

3. Run the decoder:
   
   mosesdecoder/bin/moses -f ./filteredmodel/moses.ini < testset.es >
   testset.translated.en
Translation system

- Demos
- Software
- Steps
- Software
- Steps
- Demo
Evaluate the results

1. With BLEU scoring tool. Available as a Moses script or from NIST:
   http://www.itl.nist.gov/iad/mig/tools/mtevalv13a-20091001.tar.gz

2. With Asiya package:
   http://nlp.lsi.upc.edu/asiya/
1. Evaluate the results

With BLEU scoring tool in Moses:

```bash
moses/scripts/generic/multi-bleu.perl references.en < testset.translated.en
```
MT Evaluation
Steps

1. With the Asiya toolkit:

   Asiya.pl -eval single,ulc -g sys Asiya.config

   input=raw

   SRCLANG=de
   TRGLANG=en
   SRCCASE=cs
   TRGCASE=cs

   #SRC
   src=./data/patsA61P.test.de

   #REF
   ref=./data/patsA61P.test.en

   #OUT
   sys=./data/patsA61P.test.trans.de2en
   sys=./data/patsA61P.test.trad.google.de2en
   sys=./data/patsA61P.test.trad.bing.de2en
With the Asiya toolkit:

Asiya.pl -eval single,ulc -g sys Asiya.config

input=raw
SRCLANG=de
TRGLANG=en
SRCCASE=cs
TRGCASE=cs

#SRC
src=./data/patsA61P.test.de

#REF
ref=./data/patsA61P.test.en

#OUT
sys=./data/patsA61P.test.trans.de2en
sys=./data/patsA61P.test.trad.google.de2en
sys=./data/patsA61P.test.trad.bing.de2en

Asiya.pl -eval single,ulc -m metrSet Asiya.config

SRCLANG=de
TRGLANG=en

#SRC ==================================================
src=./data/patsA61P.test.de
#REF ==================================================
ref=./data/patsA61P.test.en
#OUT ===================================================
sys=./data/patsA61P.test.trans.de2en
#-------------------------------------------------------

metrSet=1-PER 1-TER 1-WER BLEU-4 CP-0c-* CP-Op-* CP-STM-9 DP-HWC-c-4
DP-HWC-r-4 DP-HWC-w-4 DP-0c-* DP-01-* DP-Or-* DR-Or-* DR-Orp-* DR-STM-9
GTM-1 GTM-2 GTM-3 MTR-exact MTR-stem MTR-wnstm MTR-wnsyn NE-Me-* NE-Oe-**
NE-Oe-* NIST-5 RG-L RG-S* RG-SU* RG-W-1.2 SP-0c-* SP-Op-* SP-cNIST-5
SP-iobNIST-5 SP-1NIST-5 SP-pNIST-5 SR-Mr-* SR-Mrv-* SR-Or SR-Or-* SR-Orv
<table>
<thead>
<tr>
<th>METRIC NAMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTE Evaluation Metrics in Asiya (English)</td>
</tr>
<tr>
<td>RAW TEXT END</td>
</tr>
</tbody>
</table>
2. Evaluate the results on-line

1. Asiya Interface

http://asiya.lsi.upc.edu/demo/asiya_online.php
3. Analyse the results on-line

1. t-Search Interface

http://asiya.lsi.upc.edu/demo/tsearch_upload.php
MT Evaluation

Demo: http://asiya.lsi.upc.edu/demo/asiya_online.php
Part IV

Appendix: References
History of SMT

- Weaver, 1949 [Wea55]
- Alpac Memorandum [Aut66]
- Hutchins, 1978 [Hut78]
- Slocum, 1985 [Slo85]

The beginnings, word-based SMT

- Brown et al., 1990 [BCP+90]
- Brown et al., 1993 [BPPM93]
References

Phrase-based model

- Och et al., 1999 [OTN99]
- Koehn et al., 2003 [KOM03]

Log-linear model

- Och & Ney, 2002 [ON02]
- Och & Ney, 2004 [ON04]

Factored model

- Koehn & Hoang, 2007 [KH07]
Syntax-based models

- Yamada & Knight, 2001 [YK01]
- Chiang, 2005 [Chi05]
- Carreras & Collins, 2009 [CC09]

Discriminative models

- Carpuat & Wu, 2007 [CW07]
- Bangalore et al., 2007 [BHK07]
- Giménez & Màrquez, 2008 [GM08]
References

Language model

- Kneser & Ney, 1995 [KN95]

MERT

- Och, 2003 [Och03]

Domain adaptation

- Bertoldi and Federico, 2009 [Och03]
Reordering

- Crego & Mariño, 2006 [Cn06]
- Bach et al., 2009 [BGV09]
- Chen et al., 2009 [CWC09]

Systems combination

- Du et al., 2009 [DMW09]
- Li et al., 2009 [LDZ⁺09]
- Hildebrand & Vogel, 2009 [HV09]
Alternative systems in development

- Blunsom et al., 2008 [BCO08]
- Canisius & van den Bosch, 2009 [CvdB09]
- Chiang et al., 2009 [CKW09]
- Finch & Sumita, 2009 [FS09]
- Hassan et al., 2009 [HSW09]
- Shen et al., 2009 [SXZ+09]
Evaluation

- Papineni, 2002 [PRWZ02]
- Doddington, 2002 [Dod02]
- Banerjee & Alon Lavie, 2005 [BL05]
- Giménez & Amigó, 2006 [GA06]
Surveys, theses and tutorials

- Knight, 1999
  http://www.isi.edu/natural-language/mt/wkbk.rtf

- Knight & Koehn, 2003

- Koehn, 2006

- Way & Hassan, 2009

- Lopez, 2008 [Lop08]

- Giménez, 2009 [Gim08]


David Chiang.
A hierarchical phrase-based model for statistical machine translation.

David Chiang, Kevin Knight, and Wei Wang.
11,001 new features for statistical machine translation.

Josep Maria Crego and José B. Mari no.
Improving smt by coupling reordering and decoding.

Sander Canisius and Antal van den Bosch.
A constraint satisfaction approach to machine translation.


Andrew Finch and Eiichiro Sumita.  
Bidirectional phrase-based statistical machine translation.  

Jesús Giménez and Enrique Amigó.  

Jesús Giménez.  

Jesús Giménez and Lluís Màrquez.  
Discriminative Phrase Selection for SMT, pages 205–236.  

Hany Hassan, Khalil Sima’an, and Andy Way.  
A syntactified direct translation model with linear-time decoding.  
W. J. Hutchins.  

Almut Silja Hildebrand and Stephan Vogel.  
CMU system combination for WMT’09.  

Philipp Koehn and Hieu Hoang.  
Factored Translation Models.  

R. Kneser and H. Ney.  
Improved backing-off for m-gram language modeling.  

Philipp Koehn, Franz Josef Och, and Daniel Marcu.  
Statistical phrase-based translation.  


Franz Josef Och and Hermann Ney.  
The alignment template approach to statistical machine translation.  

Franz Josef Och, Christoph Tillmann, and Hermann Ney.  
Improved alignment models for statistical machine translation.  

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.  

Jonathan Slocum.  
