Introduction to (Statistical) Machine Translation

Cristina España i Bonet

MAI–ANLP

Spring 2015

Overview

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

Part I: SMT background

$\sim$ 120 min

Overview

Part II: MT evaluation

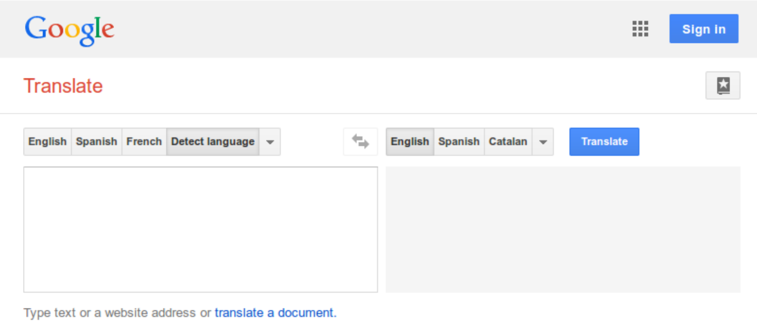
45 min

Part III: Exercise

Part I

SMT background

Goal


Google
Sign in
Translate
English
Spanish
French
Detect language
English
Spanish
Catalan
Translate
Type text or a website address or translate a document.

Goal

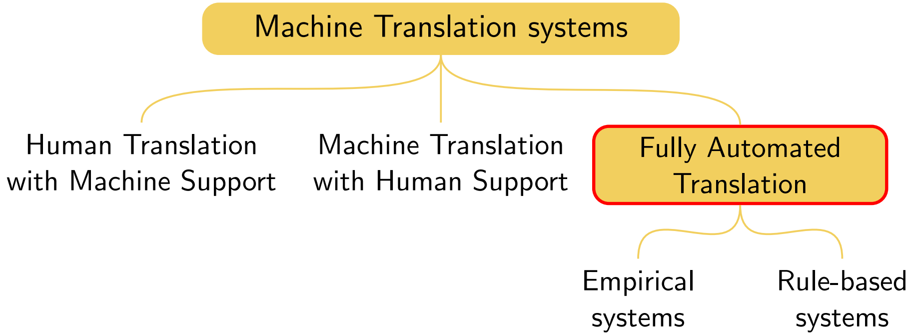
```
echo 'das ist ein kleines haus' | moses -f moses.ini
```

Outline

Introduction

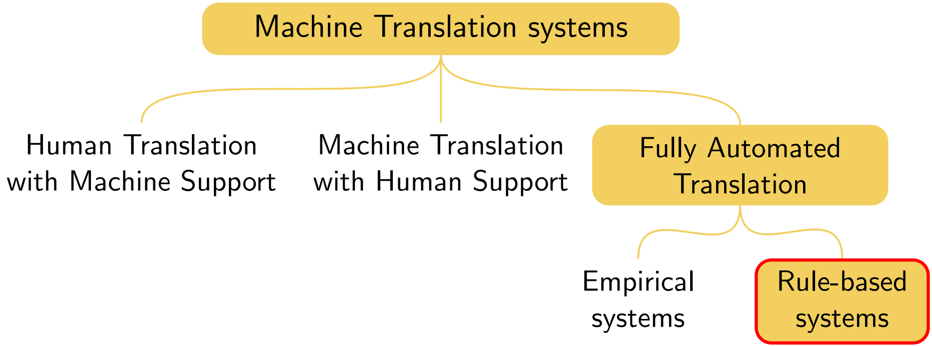
Machine Translation Taxonomy

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


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

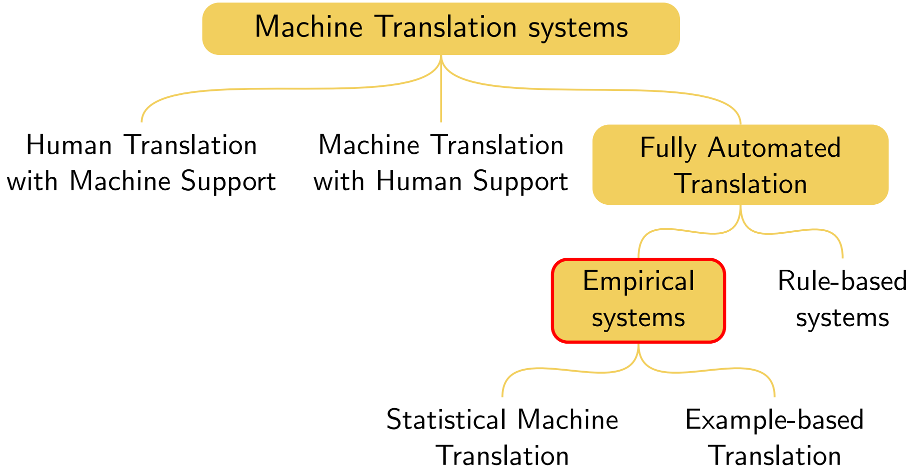
Introduction

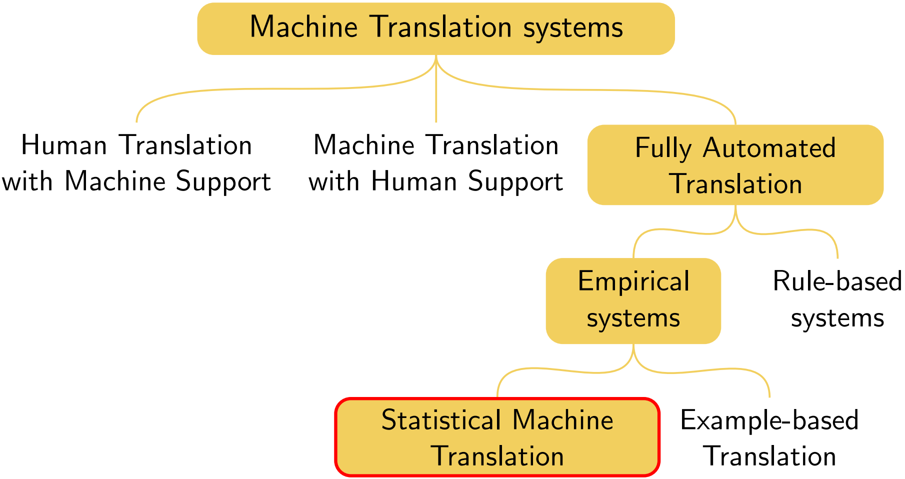
Machine Translation Taxonomy



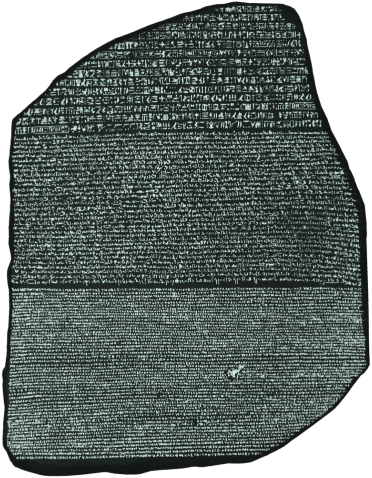
Introduction

Machine Translation Taxonomy

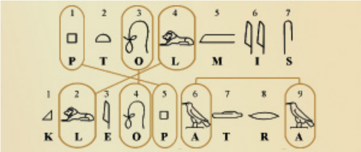



Machine Translation systems
Human Translation with Machine Support
Machine Translation with Human Support
Fully Automated Translation
Empirical systems
Rule-based systems
Statistical Machine Translation
Example-based Translation

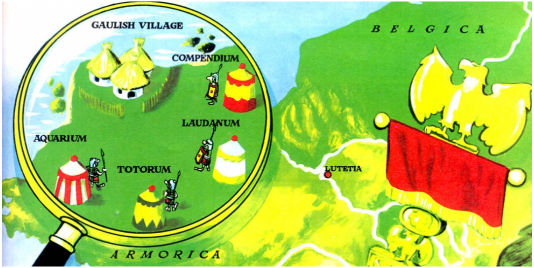

Empirical MT
relies on
aligned
corpora



Empirical MT relies on aligned corpora



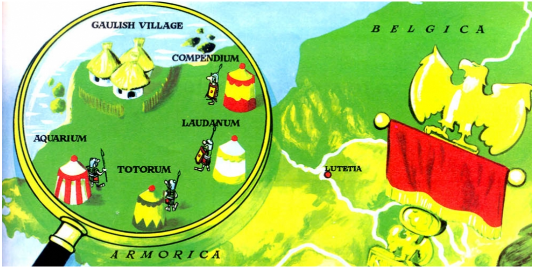
Empirical MT relies on large parallel aligned corpora



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...

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...

Empirical MT relies on large parallel aligned corpora



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...

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...

Empirical MT relies on large parallel aligned corpora

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...

Astèrix. És l'heroic petit guerrer d'aquestes aventures, viu com una centella i enginyosament astut. Per això sempre li són encomanades les missions més perilloses. Extrau la seva terrorífica força de la beguda màgica inventada pel druida Panoràmix.

Obèlix. És l'antic inseparable d'Astèrix. Fa de repartidor de menhirs i li agrada 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 batusses.

Copdegarròtix. És el cap de la tribu. Majestuós, valent i desconfiat alhora, el vell guerrer ès 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!".

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".

Aligned parallel corpora: Numbers

Corpora

Corpus	**# segments (app.)**	**# words (app.)**
JRC-Acquis	$1.0 \cdot 10^6$	$30 \cdot 10^6$
Europarl	$2.0 \cdot 10^6$	$55 \cdot 10^6$
United Nations	$10.7 \cdot 10^6$	$300 \cdot 10^6$

Books

Title	**# words (approx.)**
The Bible	$0.8 \cdot 10^6$
The Dark Tower series	$1.2 \cdot 10^6$
Encyclopaedia Britannica	$44 \cdot 10^6$

Aligned parallel corpora: Numbers

Corpora

Corpus	# segments (app.)	# words (app.)
JRC-Acquis	$1.0 \cdot 10^6$	$30 \cdot 10^6$
Europarl	$2.0 \cdot 10^6$	$55 \cdot 10^6$
United Nations	$10.7 \cdot 10^6$	$300 \cdot 10^6$

Books

Title	# words (approx.)
The Bible	$0.8 \cdot 10^6$
The Dark Tower series	$1.2 \cdot 10^6$
Encyclopaedia Britannica	$44 \cdot 10^6$

In practice

WMT13 parallel data

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

http://www.statmt.org/wmt13/translation-task.html

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

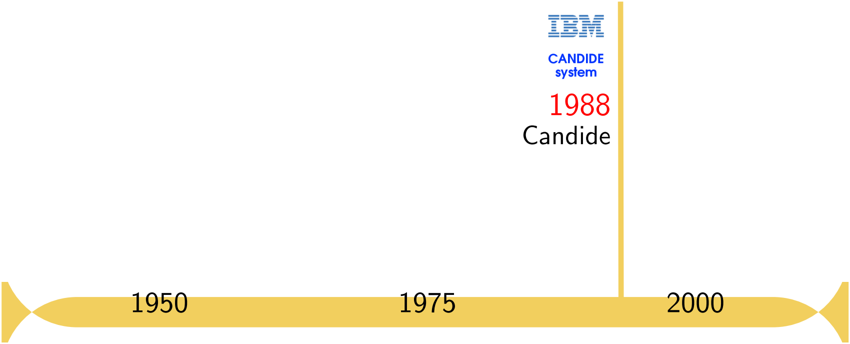
Use it for the exercise!

Outline

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

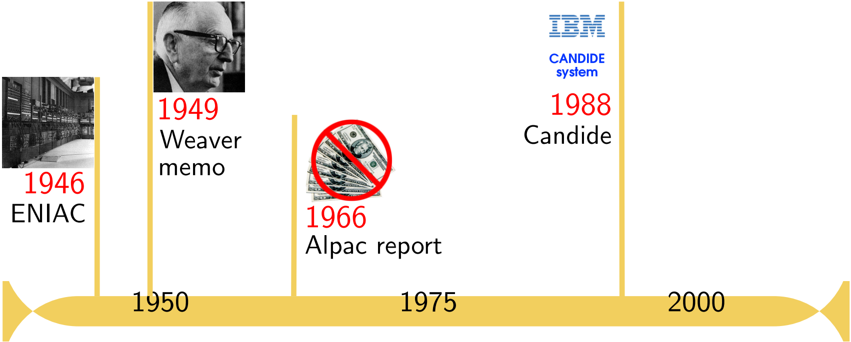
SMT, basics

The beginnings, summarised timeline



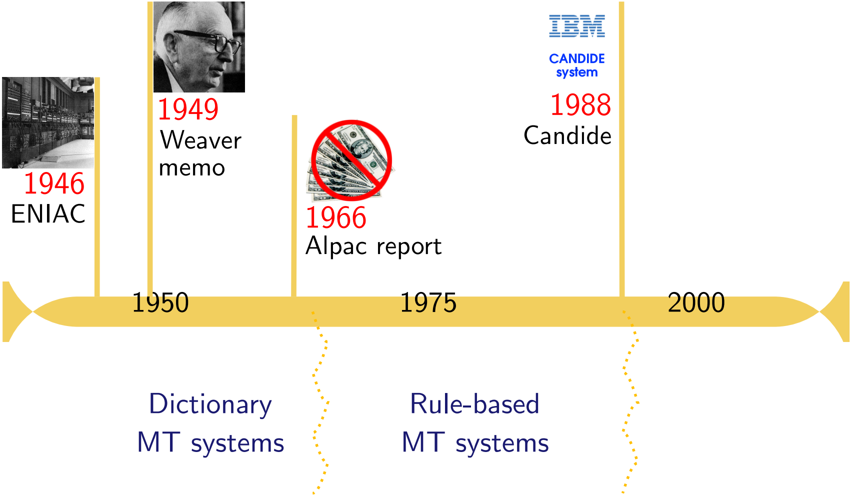
SMT, basics

The beginnings, summarised timeline



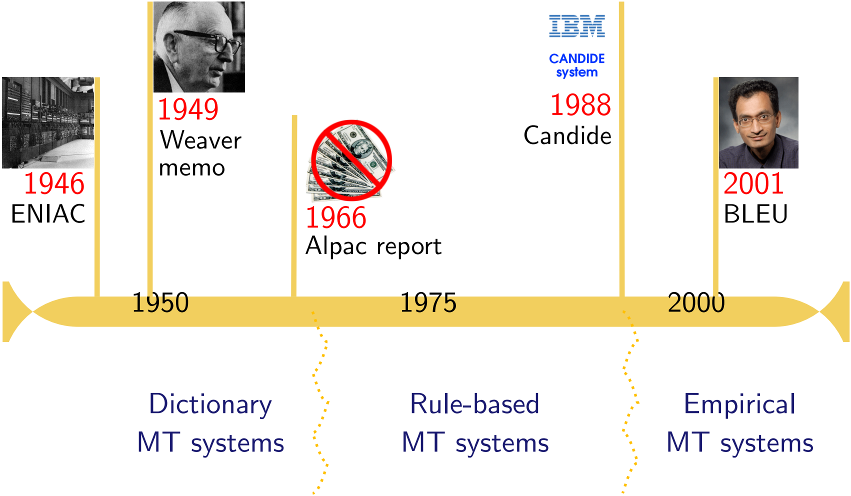
SMT, basics

The beginnings, summarised timeline

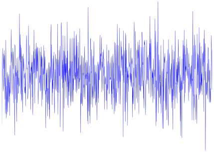


SMT, basics

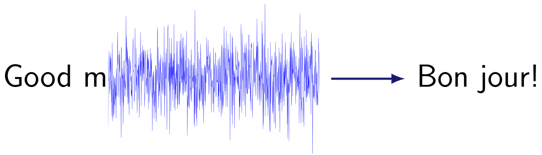
The beginnings, summarised timeline



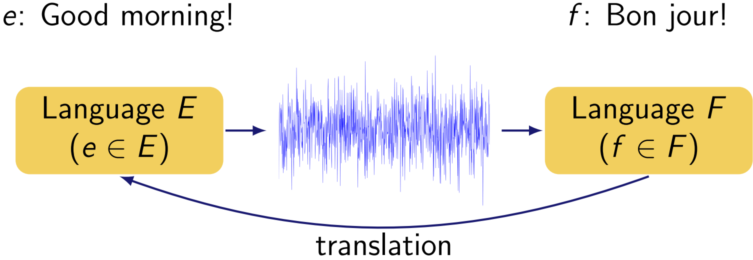
The Noisy Channel as a statistical approach to translation:

Good morning! $\longrightarrow$ 

The Noisy Channel as a statistical approach to translation:

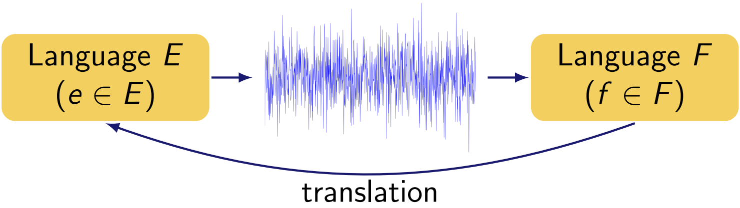


The Noisy Channel as a statistical approach to translation:



SMT, basics

The Noisy Channel approach

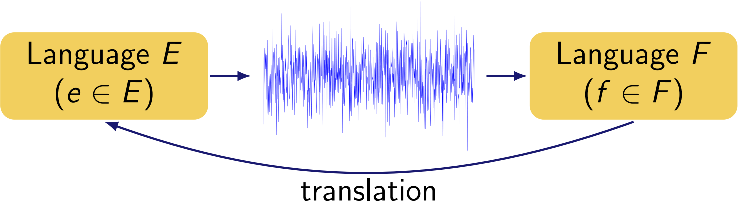


Mathematically:

$$P(e|f)$$

SMT, basics

The Noisy Channel approach



Mathematically:

$$P(e|f) = \frac{P(e)\,P(f|e)}{P(f)}$$

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e|f) = \mathrm{argmax_e}\ P(e)\,P(f|e)$$

$$T(f) = \hat{e} = \mathrm{argmax}_{\mathrm{e}}\, P(e)\, P(f|e)$$

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*

$$T(f) = \hat{e} = \mathrm{argmax}_{\mathrm{e}}\ P(e)\, P(f|e)$$

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

$$T(f) = \hat{e} = \mathrm{argmax}_e \; P(e)\, P(f|e)$$

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

argmax

- Search done by the *decoder*

Outline

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

Language model

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e)\, P(f|e)$$

Estimation of how probable a sentence is.

Naïve estimation on a corpus with N sentences:

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

Problem:

- Long chains are difficult to observe in corpora.
  $\Rightarrow$ Long sentences may have zero probability!

Language model

$$T(f) = \hat{e} = \mathrm{argmax}_{\mathrm{e}}\; P(e)\, P(f|e)$$

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!

Language model

$$T(f) = \hat{e} = \mathrm{argmax}_{\mathrm{e}}\ P(e)\, P(f|e)$$

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

$$\begin{aligned} P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\ & P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\ & P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\ & P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\ & P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull}) \end{aligned}$$

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

Example, a 4-gram model

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

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

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

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

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

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

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

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

where, for each factor,

$$P(\text{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$
P(\text{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}
$$

Example, a 4-gram model

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

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

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\; P(\texttt{work}|\phi,\phi,\texttt{All})\; P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\; P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}$$

SMT, components

The language model $P(e)$

Example, a 4-gram model

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

$$
\begin{aligned}
P(e) = {} & P(\texttt{All}|\phi,\phi,\phi)\ P(\texttt{work}|\phi,\phi,\texttt{All})\ P(\texttt{and}|\phi,\texttt{All},\texttt{work}) \\
& P(\texttt{no}|\texttt{All},\texttt{work},\texttt{and})\ P(\texttt{play}|\texttt{work},\texttt{and},\texttt{no}) \\
& P(\texttt{makes}|\texttt{and},\texttt{no},\texttt{play}) P(\texttt{Jack}|\texttt{no},\texttt{play},\texttt{makes}) \\
& P(\texttt{a}|\texttt{play},\texttt{makes},\texttt{Jack}) P(\texttt{dull}|\texttt{makes},\texttt{Jack},\texttt{a}) \\
& P(\texttt{boy}|\texttt{Jack},\texttt{a},\texttt{dull})
\end{aligned}
$$

where, for each factor,

$$
P(\texttt{and}|\phi,\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All work and})}}{N_{(\texttt{All work})}}
$$

Main problems and criticisims:

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

Solution

Smoothing techniques:

- Linear interpolation.

$$P(\text{and}|\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All},\texttt{work},\texttt{and})}}{N_{(\texttt{All},\texttt{work})}} + \lambda_2 \frac{N_{(\texttt{work},\texttt{and})}}{N_{(\texttt{work})}} + \lambda_1 \frac{N_{(\texttt{and})}}{N_{words}} + \lambda_0$$

Main problems and criticisims:

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

Solution

Smoothing techniques:

- Linear interpolation.
- Back-off models.

$$P(\texttt{and}|\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All},\texttt{work},\texttt{and})}}{N_{(\texttt{All},\texttt{work})}} + \lambda_2 \frac{N_{(\texttt{work},\texttt{and})}}{N_{(\texttt{work})}} + \lambda_1 \frac{N_{(\texttt{and})}}{N_{words}} + \lambda_0$$

SMT, components

The language model $P(e)$

Main problems and criticisims:

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

Solution

Smoothing techniques:

- Linear interpolation.

$$P(\texttt{and}|\texttt{All},\texttt{work}) = \frac{N_{(\texttt{All},\texttt{work},\texttt{and})}}{N_{(\texttt{All},\texttt{work})}} + \lambda_2 \frac{N_{(\texttt{work},\texttt{and})}}{N_{(\texttt{work})}} + \lambda_1 \frac{N_{(\texttt{and})}}{N_{words}} + \lambda_0$$

Main problems and criticisims:

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

Solution

Smoothing techniques:

- Linear interpolation.

$$P(\mathtt{and}|\mathtt{All},\mathtt{work}) = \lambda_3 \frac{N_{(\mathtt{All},\mathtt{work},\mathtt{and})}}{N_{(\mathtt{All},\mathtt{work})}} + \lambda_2 \frac{N_{(\mathtt{work},\mathtt{and})}}{N_{(\mathtt{work})}} + \lambda_1 \frac{N_{(\mathtt{and})}}{N_{words}} + \lambda_0$$

SMT, components

The language model $P(e)$

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 ' -0.6271471
-2.268107 ( -0.6895114
```

SMT, components

The language model $P(e)$

```
\2-grams:
 -1.08232 concierto ,
-1.093977 concierto . -0.2378127
-1.747908 concierto ad
-1.748422 concierto cobraria
-0.8927398 concierto de
-1.744176 concierto europeo
-1.740879 concierto internacional
-1.635606 concierto para
-1.744787 concierto regional

...

\5-grams:
-0.8890668 no son los unicos culpables
-1.396196 no son los unicos problemas
-0.7550655 no son los unicos que
-1.240193 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

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e)\, P(f|e)$$

Estimation of the lexical correspondence between languages.

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

SMT, components

The translation model $P(f|e)$

Translation model

$$T(f) = \hat{e} = \mathrm{argmax}_{\mathrm{e}}\ P(e)\, P(f|e)$$

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

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e)\, P(f|e)$$

Estimation of the lexical correspondence between languages.

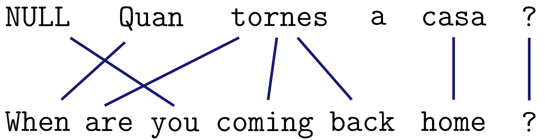
How can be $P(f|e)$ characterised?

```
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
  $t(\texttt{Quan}|\texttt{When})$: the prob. that `Quan` translates into `When`.

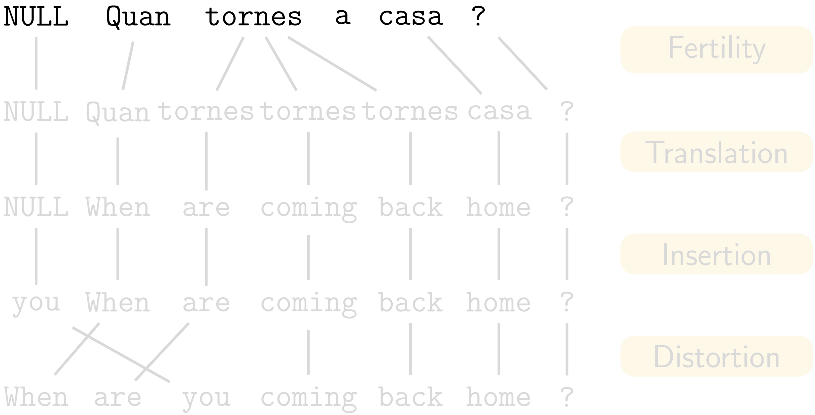
- Fertility n
  $n(3|\texttt{tornes})$: the prob. that `tornes` generates 3 words.

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
  $p(\texttt{you}|\texttt{NULL})$: the prob. that the spurious word `you` is generated (from `NULL`).

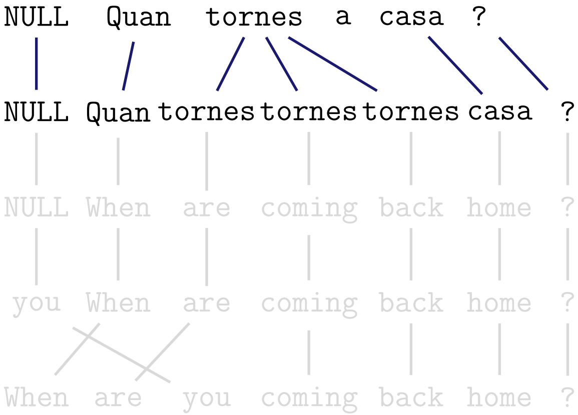
Back to the example:



SMT, components

The translation model $P(f|e)$

Back to the example:



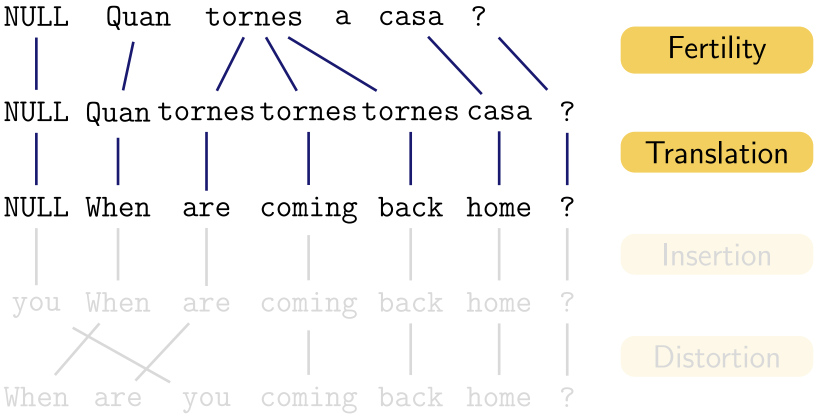
Fertility

Translation

Insertion

Distortion

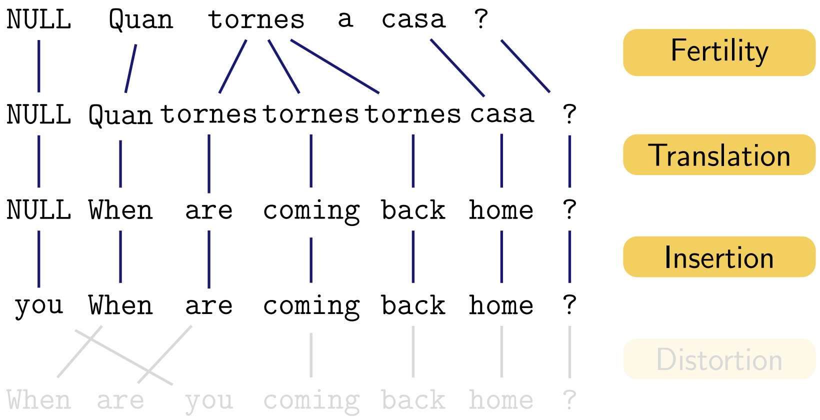
Back to the example:



SMT, components

The translation model $P(f|e)$

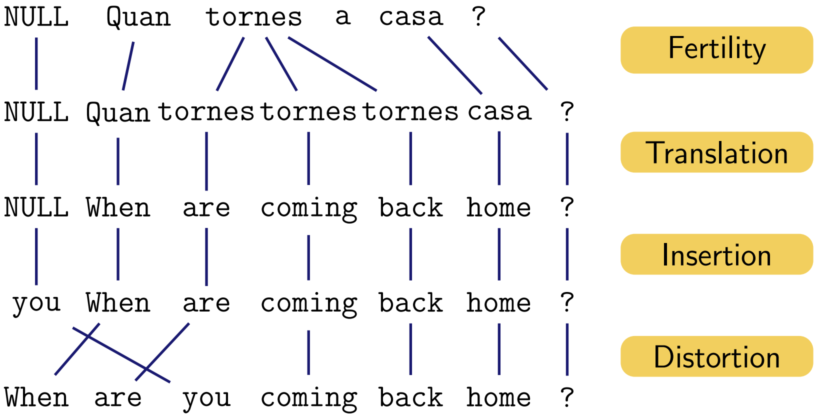
Back to the example:



SMT, components

The translation model $P(f|e)$

Back to the example:



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.

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 milion sentences word by word.
- Estimate word alignments together with the parameters.

SMT, components

The translation model $P(f|e)$

Expectation-Maximisation algorithm

Parameter initialisation

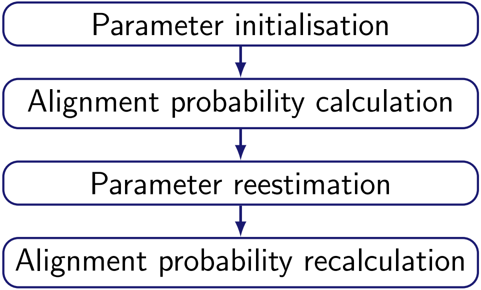
↓

Alignment probability calculation

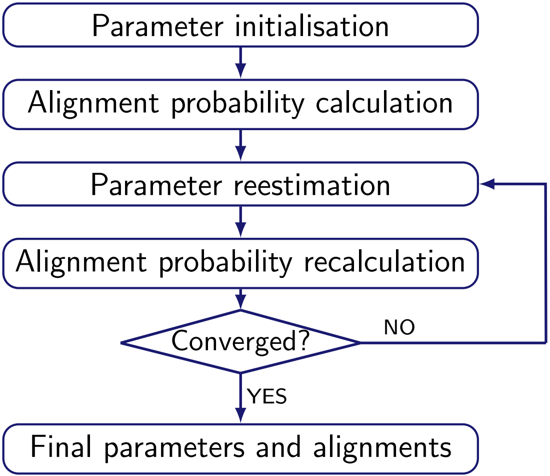
SMT, components

The translation model $P(f|e)$

Expectation-Maximisation algorithm



Expectation-Maximisation algorithm

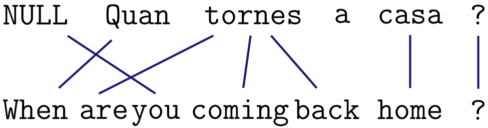


Alignment's asymmetry

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



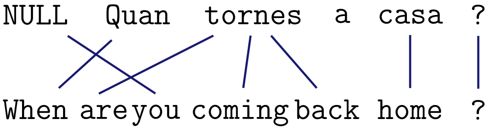
English to Catalan

Alignment's asymmetry

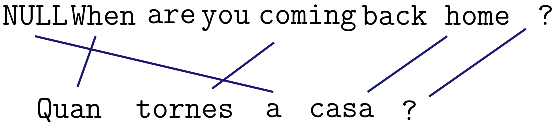
The definitions in IBM models make the alignments asymmetric

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

Catalan to English



English to Catalan



SMT, components

The translation model $P(f|e)$

Visually:

Catalan to English

SMT, components

The translation model $P(f|e)$

Visually:

English to Catalan

SMT, components

The translation model $P(f|e)$

Alignment symmetrisation

- Intersection: high-confidence, high precision.

Catalan to English $\bigcap$ English to Catalan

Alignment symmetrisation

- Union: lower confidence, high recall.

Catalan to English $\bigcup$ English to Catalan

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 })
senorias ({ }) 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 })
```

SMT, components

The translation model $P(f|e)$

In practice,

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

```
# 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 deseo 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 })
```

```
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
13-12 12-13 15-14 17-15 18-16 23-17 19-20 20-22 24-23 21-29 26-32 27-33
27-34 30-35 28-36 31-36 29-37 30-37 31-37 31-38 32-39
```

SMT, components

The translation model $P(f|e)$

```
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.`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: ϕ

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David reads`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David reads the`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David reads the book`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David reads the book new.`

From Word-based to Phrase-based models

f: `En David llegeix el llibre nou.`

e: `David reads the book new.` $\sim$

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: `En David llegeix el llibre nou.`

e: `David reads the new book.` ✓

f: `En David llegeix el llibre de nou.`

e: ϕ

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`

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`

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`

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`

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`

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.`

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.` ✗

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: `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`

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.` ✓

From Word-based to Phrase-based models

f: `En David llegeix el llibre` $\boxed{\texttt{nou}}$.

e: `David reads the` $\boxed{\texttt{new}}$ `book.` ✓

f: `En David llegeix el llibre` $\boxed{\texttt{de nou}}$.

e: `David reads the book of new.` ✗

e: `David reads the book` $\boxed{\texttt{again}}$. ✓

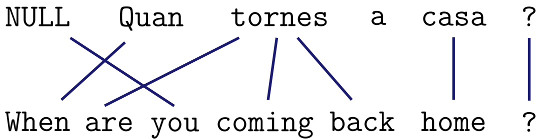
- Some sequences of words usually translate together.
- Approach: take sequences (phrases) as translation units.

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 catched.

SMT, components

The translation model $P(f|e)$

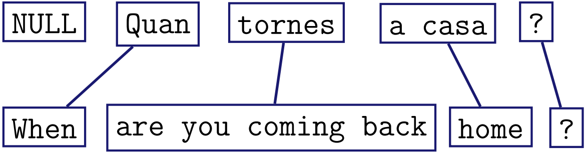


With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.

SMT, components

The translation model $P(f|e)$

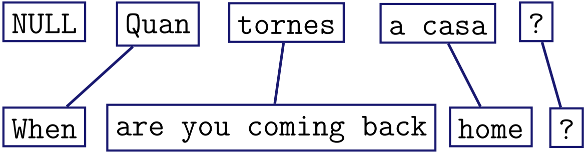


With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.

SMT, components

The translation model $P(f|e)$

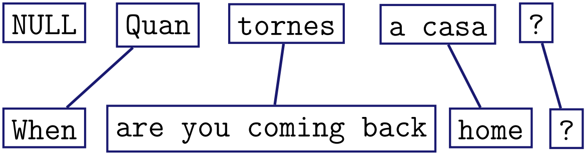


With the new translation units, $P(f|e)$ can be obtained following the same strategy as for word-based models with few modifications:

1. Segment source sentence into phrases.
2. Translate each phrase into the target language.
3. Reorder the output.

SMT, components

The translation model $P(f|e)$



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?

SMT, components

The translation model $P(f|e)$

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.

SMT, components

The translation model $P(f|e)$

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.

SMT, components

The translation model $P(f|e)$

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]

[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.

SMT, components

The translation model $P(f|e)$

Phrase extraction through an example:

(Quan tornes, When are you coming back)

SMT, components

The translation model $P(f|e)$

Phrase extraction through an example:

~~(Quan tornes, When are you coming back)~~

Phrase extraction through an example:

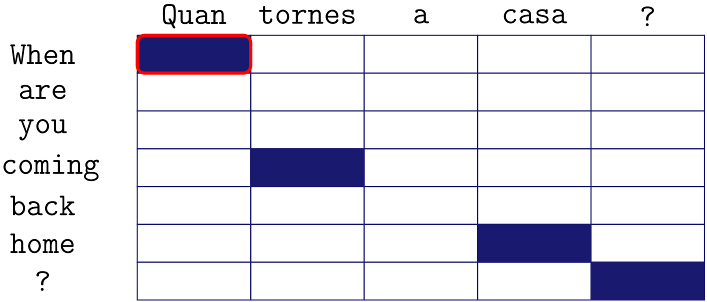
~~(Quan tornes, When are you coming back)~~

(Quan tornes tu, When are you coming back)

SMT, components

The translation model $P(f|e)$

Intersection

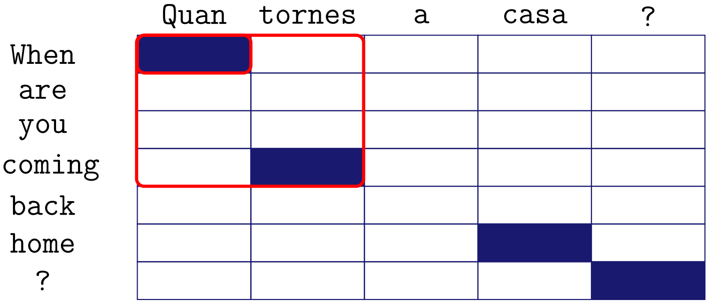


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

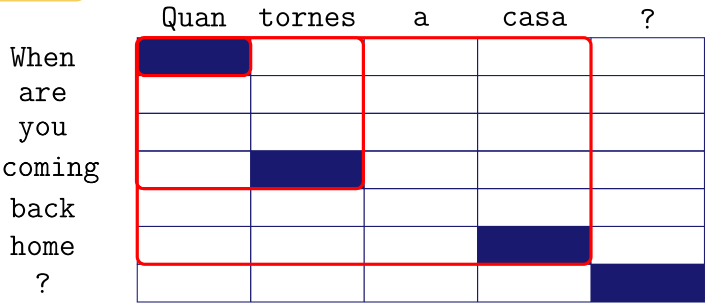


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

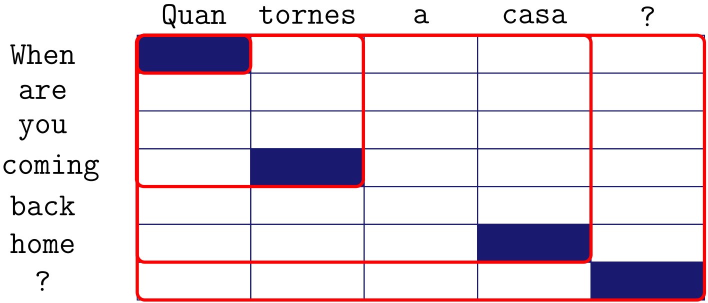


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

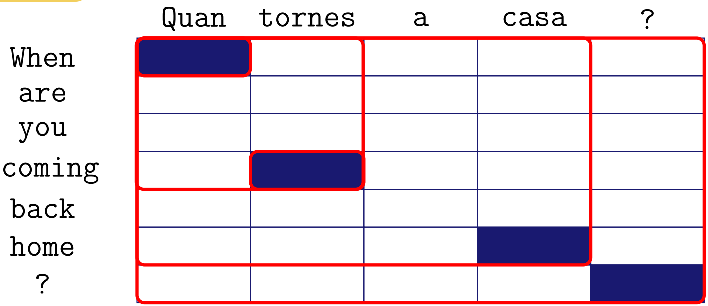


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

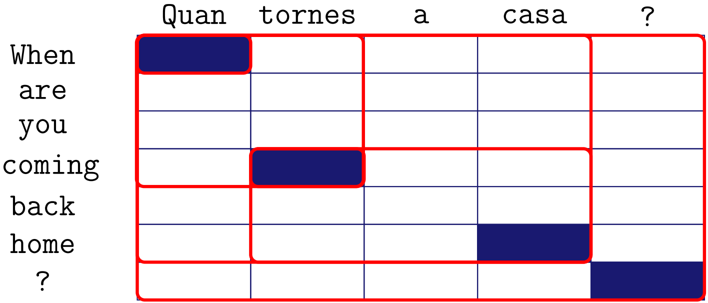


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

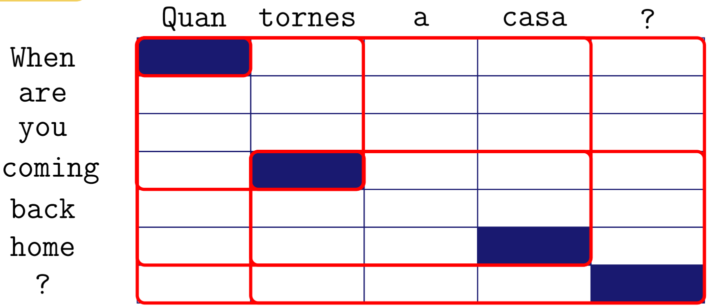


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

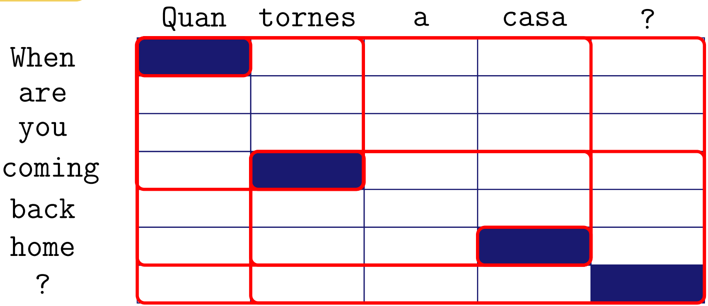


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

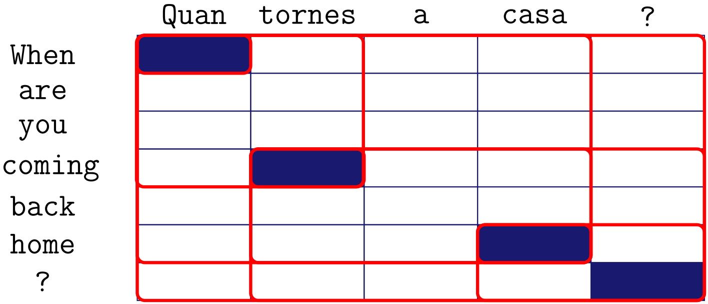


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

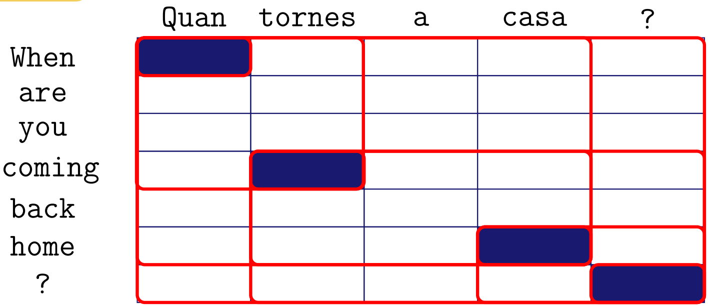


(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 10 phrases

SMT, components

The translation model $P(f|e)$

Intersection

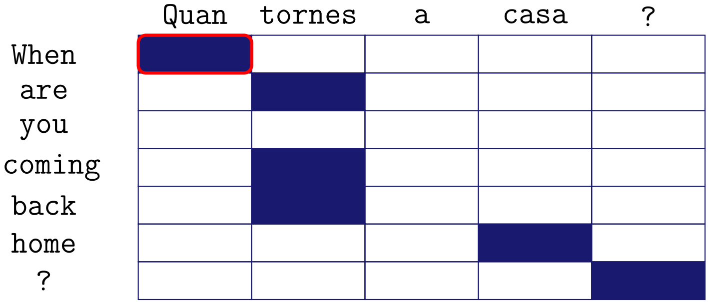


`(Quan, When) (Quan tornes, When are you coming) (Quan tornes a casa, When are you coming back home) (Quan tornes a casa ?, When are you coming back home ?) (tornes, coming) (tornes a casa, coming back home) (tornes a casa ?, coming back home ?) (casa, home) (casa ?, home ?) (?, ?)` 10 phrases

SMT, components

The translation model $P(f|e)$

Union

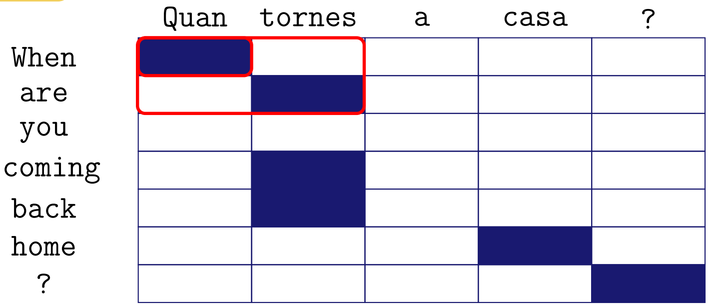


(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases

SMT, components

The translation model $P(f|e)$

Union

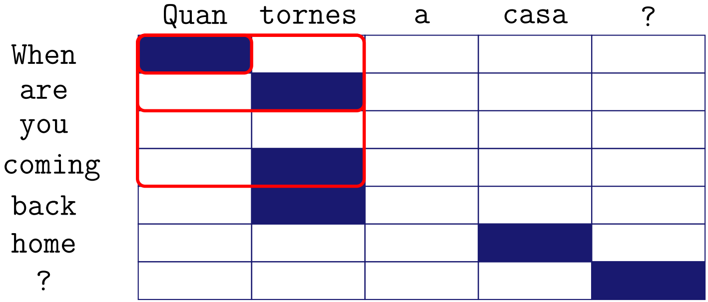


(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases

SMT, components

The translation model $P(f|e)$

Union

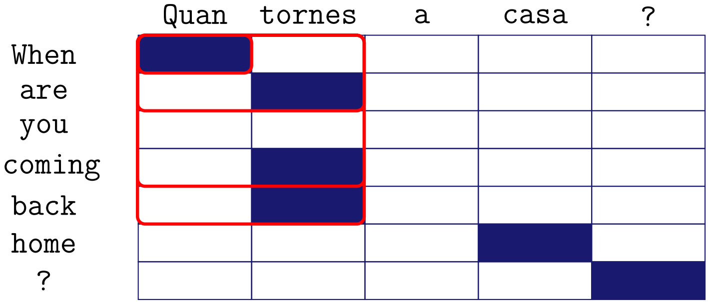


(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home) ... (tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?) 21 phrases

SMT, components

The translation model $P(f|e)$

Union

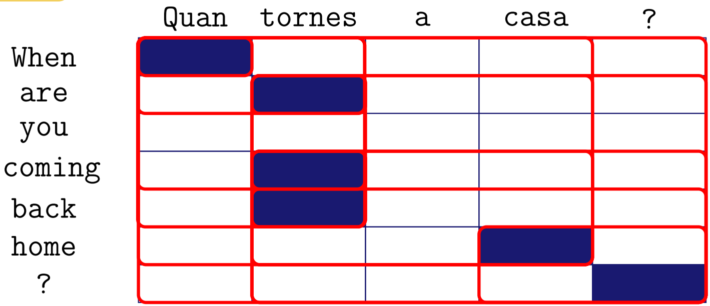


`(Quan, When) (Quan tornes, When are) (Quan tornes, When are you coming) (Quan tornes, When are you coming back) (Quan tornes a casa, When are you coming back home)` ... `(tornes a casa ?, are you coming back home ?) (casa, home) (casa ?, home ?) (?, ?)` 21 phrases

SMT, components

The translation model $P(f|e)$

Union



(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.

SMT, components

The translation model $P(f|e)$

In practice,

```
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
```

```
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
```

```
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 ||| 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
...
```

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

$$T(f) = \hat{e} = \mathrm{argmax}_e \; P(e)\, P(f|e)$$

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

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e)\,P(f|e)$$

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

$$T(f) = \hat{e} = \mathrm{argmax_e}\; P(e)\, P(f|e)$$

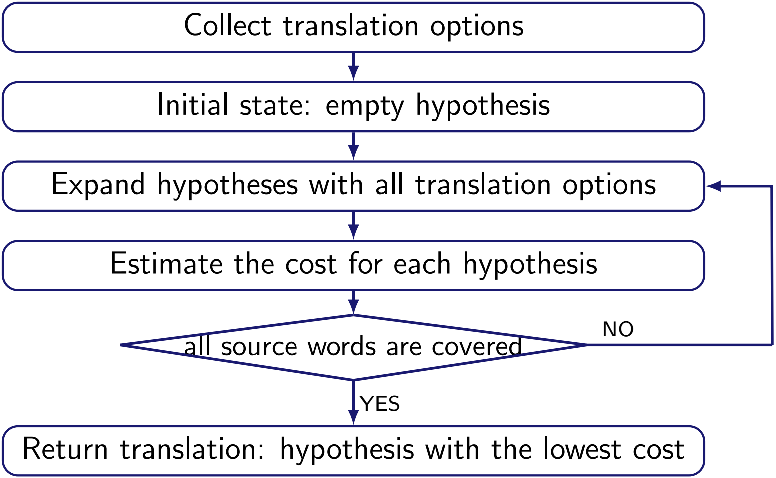
Responsible for the search in the space of possible translations.

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



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)
```

SMT, components
Decoding

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`)

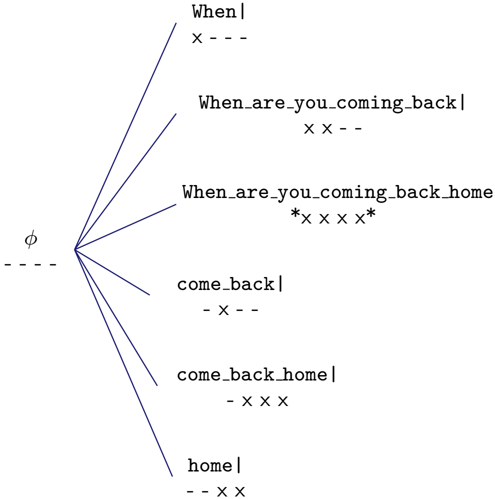
- Initial hypothesis

  Constructed sentence so far: ϕ
  Source words already translated: - - - -

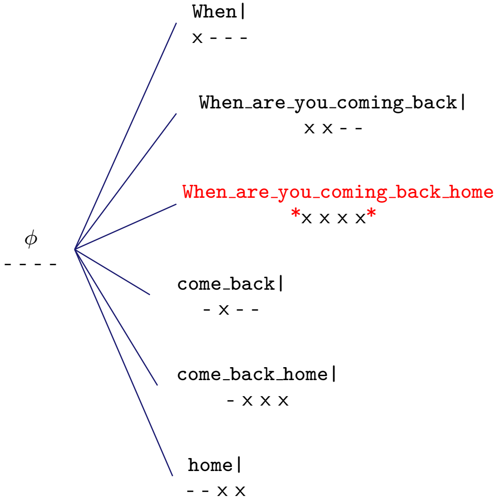
ϕ

- - - -

SMT, components
Decoding

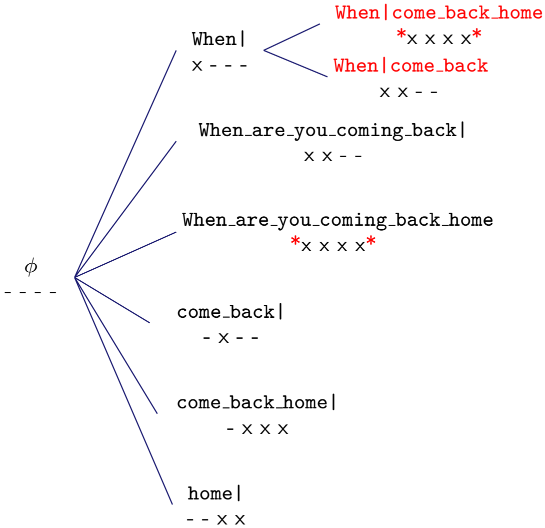


SMT, components
Decoding

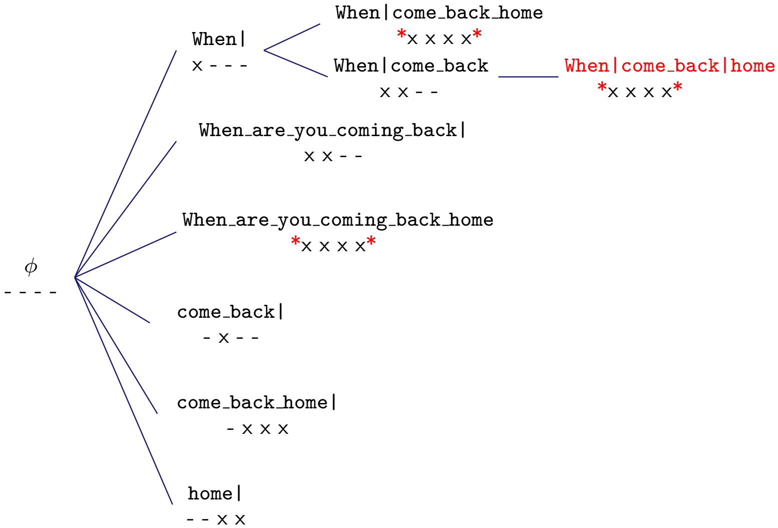


SMT, components

Decoding

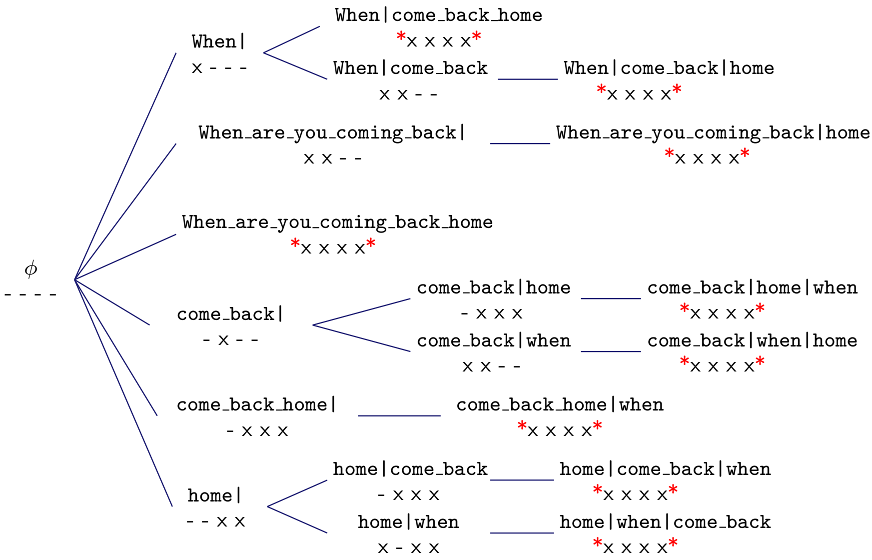


SMT, components
Decoding



SMT, components

Decoding



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 $\Rightarrow$ $2^{30} = 1,073,741,824$ 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 $\Rightarrow$ $2^{30} = 1,073,741,824$ 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.

```
When|come_back_home   ⟺   When|come_back|home
     x x x x                    x x x x
```

- 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.

`When|come_back_home` $\Longleftrightarrow$ `When|come_back|home`

x x x x   x x x x

- 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.

```
When|come_back_home          When|come_back|home
      x x x x          ⟺           x x x x
```

- 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.

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

SMT, the log-linear model

Motivation

Maximum likelihood (ML)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ P(e)\, P(f|e)$$

Maximum entropy (ME)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ \exp\left\{\sum \lambda_m\, h_m(f,e)\right\}$$

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} \log P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ \sum \lambda_m\, h_m(f,e)$$

Log-linear model

SMT, the log-linear model

Motivation

Maximum likelihood (ML)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ P(e)\, P(f|e)$$

Maximum entropy (ME)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ \exp\left\{\sum \lambda_m\, h_m(f,e)\right\}$$

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} \log P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ \sum \lambda_m\, h_m(f,e)$$

Log-linear model

SMT, the log-linear model

Motivation

Maximum likelihood (ML)

$$\hat{e} = \mathrm{argmax_e} P(e|f) = \mathrm{argmax_e}\ P(e)\,P(f|e)$$

Maximum entropy (ME)

$$\hat{e} = \mathrm{argmax_e} P(e|f) = \mathrm{argmax_e}\ \exp\left\{\sum \lambda_m\, h_m(f, e)\right\}$$

$$\hat{e} = \mathrm{argmax_e} \log P(e|f) = \mathrm{argmax_e}\ \sum \lambda_m\, h_m(f, e)$$

Log-linear model

SMT, the log-linear model

Motivation

Maximum likelihood (ML)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ P(e)\,P(f|e)$$

Maximum entropy (ME)

$$\hat{e} = \mathrm{argmax}_{\mathrm{e}} \log P(e|f) = \mathrm{argmax}_{\mathrm{e}}\ \sum \lambda_m\, h_m(f,e)$$

Log-linear model with

$$h_1(f,e) = logP(e),\ h_2(f,e) = logP(f|e),\ \text{and}\ \lambda_1 = \lambda_2 = 1$$

$\Rightarrow$ 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 λ_m must be somehow determined.
- Different knowledge sources can be used.

Standard feature functions

Eight features are usually used: $P(e)$, $P(f|e)$, $P(e|f)$, $lex(f|e)$, $lex(e|f)$, $ph(e)$, $w(e)$ 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.

Standard feature functions

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

- Translation model $lex(f|e)$
  $lex(f|e)$: Lexical translation model probability.
- Translation model $lex(e|f)$
  $lex(e|f)$: Inverse lexical translation model probability.
- Phrase penalty $ph(e)$
  $ph(e)$: A constant cost per produced phrase.

Standard feature functions

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

- Word penalty $w(e)$
  $w(e)$: A constant cost per produced word.
- Distortion $P_d(e,f)$
  $P_d(\mathrm{ini}_{\mathrm{phrase}_i}, \mathrm{end}_{\mathrm{phrase}_{i-1}})$: Relative distortion probability distribution. A simple distortion model:
  $P_d(\mathrm{ini}_{\mathrm{phrase}_i}, \mathrm{end}_{\mathrm{phrase}_{i-1}}) = \alpha |\mathrm{ini}_{\mathrm{phrase}_i} - \mathrm{end}_{\mathrm{phrase}_{i-1}} - 1|$

SMT, components

The translation model $P(f|e)$

In practice,

```
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 ||| 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?

(are, tornes, monotone)

SMT, the log-linear model

Digression: lexicalised reordering or distortion

From where and how can one learn reorders?

(coming back, tornes, swap)

From where and how can one learn reorders?

(home ?, casa ?, discontinuous)

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

$$P_{or.}(\mathrm{orientation}|f,e) = \frac{count(\mathrm{orientation}, e, f)}{\sum_{or.} count(\mathrm{orientation}, e, f)}$$

- Sparse statistics of the orientation types $\rightarrow$ smoothing.
- Several variations.

SMT, components

The translation model $P(f|e)$

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
```

Standard feature functions

13 features may be used:

- $P(e)$;
- $P(f|e)$, $P(e|f)$, $lex(f|e)$, $lex(e|f)$;
- $ph(e)$, $w(e)$;
- $P_{mon}(o|e,f)$, $P_{swap}(o|e,f)$, $P_{dis}(o|e,f)$,
- $P_{mon}(o|f,e)$, $P_{swap}(o|f,e)$, $P_{dis}(o|f,e)$.

Development training, weights optimisation

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

$$\hat{e} = \mathrm{argmax_e} \log P(e|f) = \mathrm{argmax_e} \sum \lambda_m h_m(f, e)$$

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

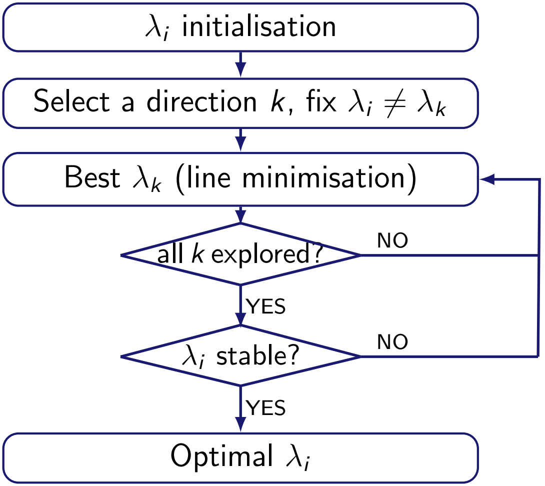
- Approach: Minimise an error function.

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

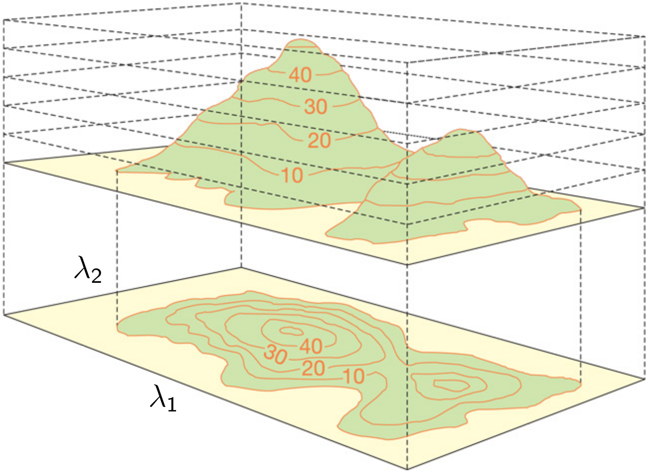
- There exist several error measures or metrics.
- Metrics not always correlate with human judgements.
- The quality of the final translation on the metric choosen for the optimisation is shown to improve.
- For the moment, let's say we use BLEU.

(More on MT Evaluation section)

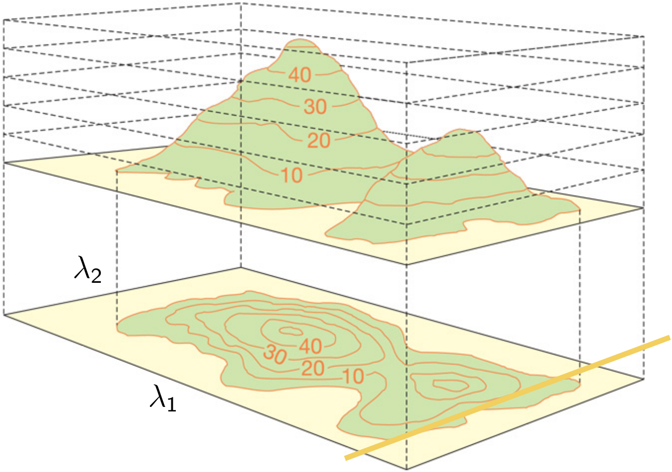
Minimum Error-Rate Training rough algorithm



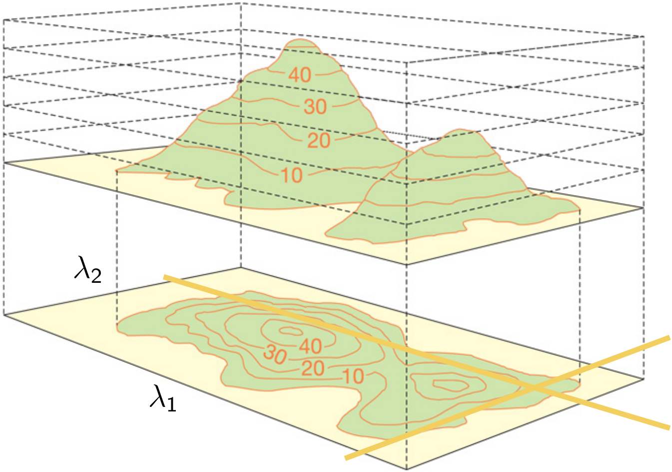
Powell's method (2D: λ_1,λ_2)



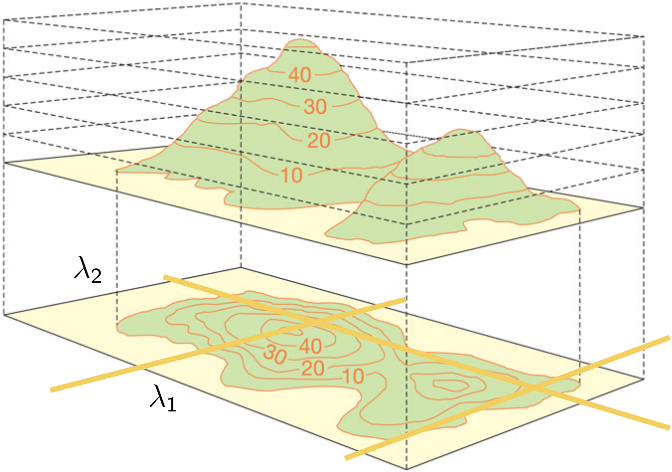
Powell's method (2D: λ_1,λ_2)



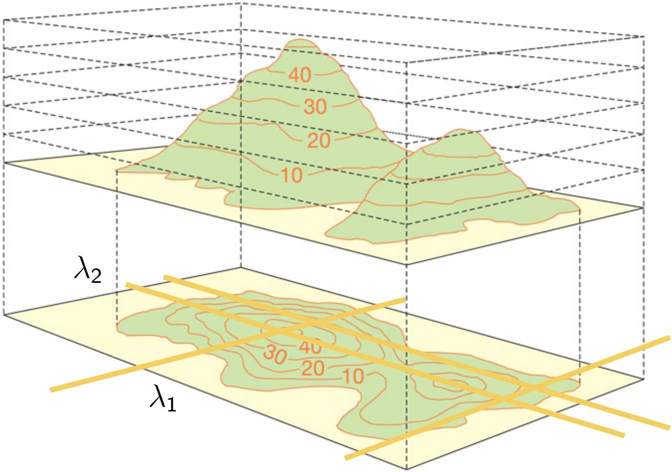
Powell's method (2D: λ_1,λ_2)



Powell's method (2D: λ_1,λ_2)



Powell's method (2D: λ_1,λ_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
```

Log-linear model: keep in mind

- 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 Word Aligner
`https://code.google.com/p/berkeleyaligner`

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

...

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.

$$(\texttt{word}) \Longrightarrow (\texttt{word}, \texttt{lemma}, \texttt{PoS}, \texttt{morphology}, \ldots)$$

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

Factored translation models

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

$$(\texttt{word}) \Longrightarrow (\texttt{word}, \texttt{lemma}, \texttt{PoS}, \texttt{morphology}, \ldots)$$

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

$$\begin{array}{cccc}
\texttt{lemma}_f & \texttt{PoS}_f & \texttt{morphology}_f & \qquad \texttt{word}_f \\
\downarrow \mathsf{T} & \downarrow \mathsf{T} & \downarrow \mathsf{T} & \\
\texttt{lemma}_e & \texttt{PoS}_e & \texttt{morphology}_e & \xrightarrow{\mathsf{G}} \texttt{word}_e
\end{array}$$

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.

$$(\texttt{word}) \Longrightarrow (\texttt{word}, \texttt{lemma}, \texttt{PoS}, \texttt{morphology}, \ldots)$$

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

$$\begin{array}{ccccc}
\texttt{casa}_f & \texttt{NN}_f & \texttt{fem., plural}_f & & \texttt{cases}_f \\
\downarrow \mathsf{T} & \downarrow \mathsf{T} & \downarrow \mathsf{T} & & \\
\texttt{house}_e & \texttt{NN}_e & \texttt{plural}_e & \xrightarrow{\mathsf{G}} & \texttt{houses}_e
\end{array}$$

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.

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.

Syntactic translation models

Incorporate syntax to the source and/or target languages.

Approaches

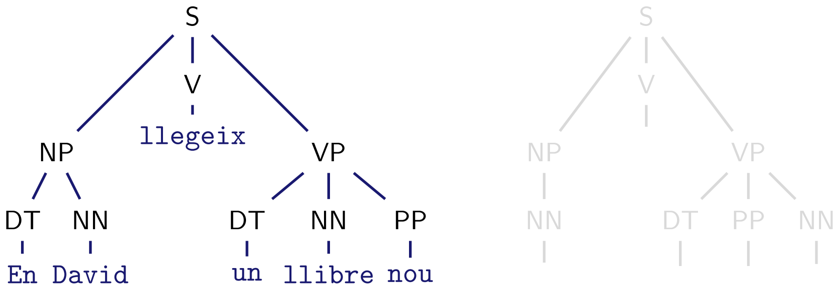
- Synchronous grammar formalism which learns a grammar that can simultaneously generate both trees.
  - Syntax-based. Respect linguistic units in translation.
  - Hierarchical phrase-based. Respect phrases in translation.

Syntactic models ease reordering. An intuitive example:

`En David llegeix un llibre nou`

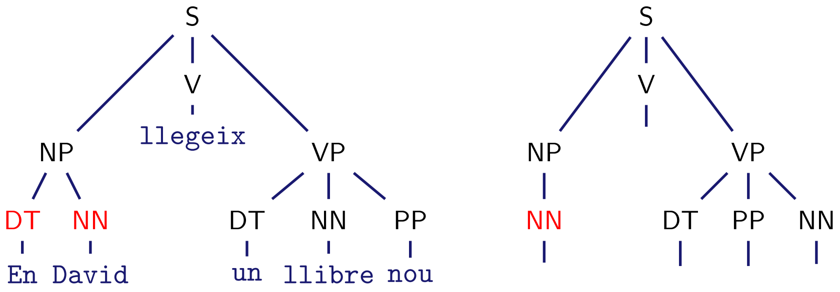
Syntactic models ease reordering. An intuitive example:

`En David llegeix un llibre nou`



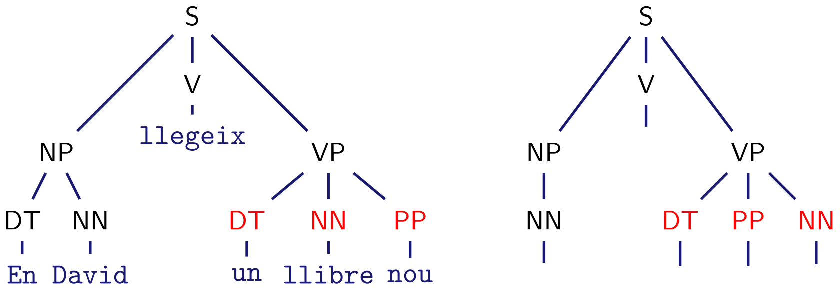
Syntactic models ease reordering. An intuitive example:

`En David llegeix un llibre nou`



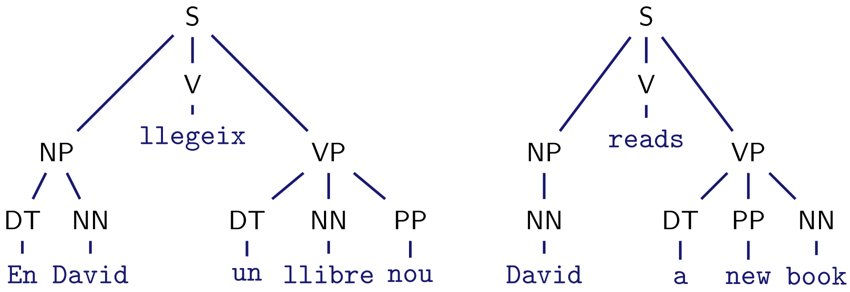
Syntactic models ease reordering. An intuitive example:

`En David llegeix un llibre nou`



Syntactic models ease reordering. An intuitive example:

`En David llegeix un llibre nou`



`David reads a new book`

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.

Beyond standard SMT: keep in mind

- Factored models include linguistic information in phrase-based models and are suitable for morphologically rich languages.
- Syntactic models consider somehow syntaxis 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

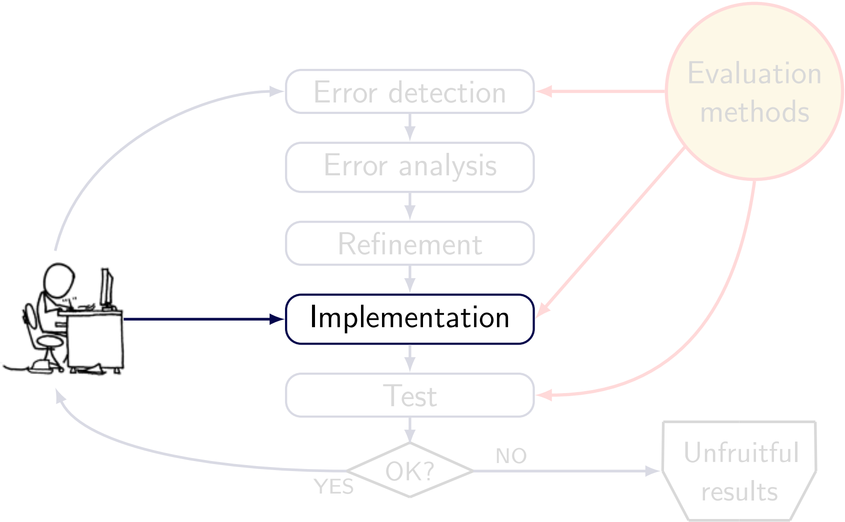
MT Evaluation

Outline

Outline

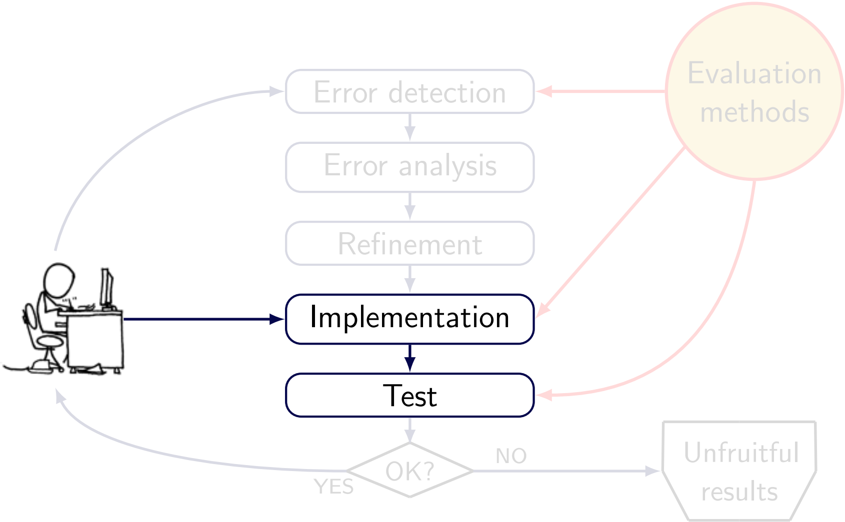
MT Evaluation

Importance for system development



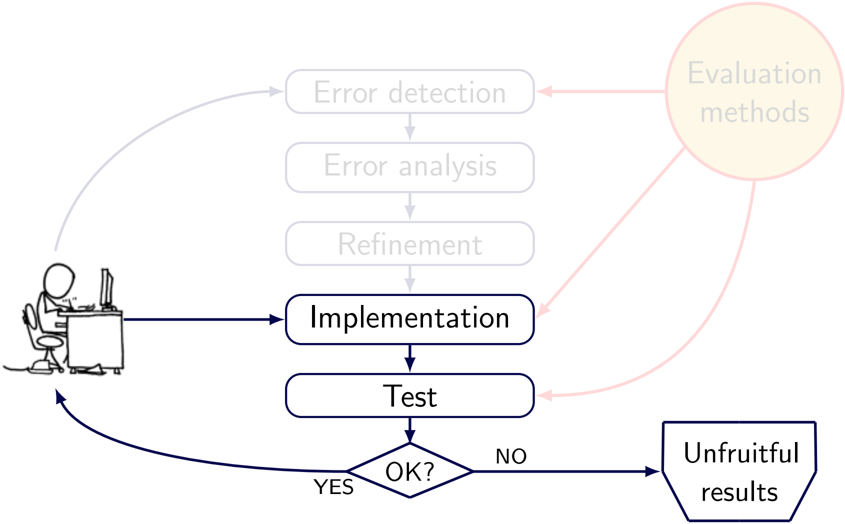
MT Evaluation

Importance for system development



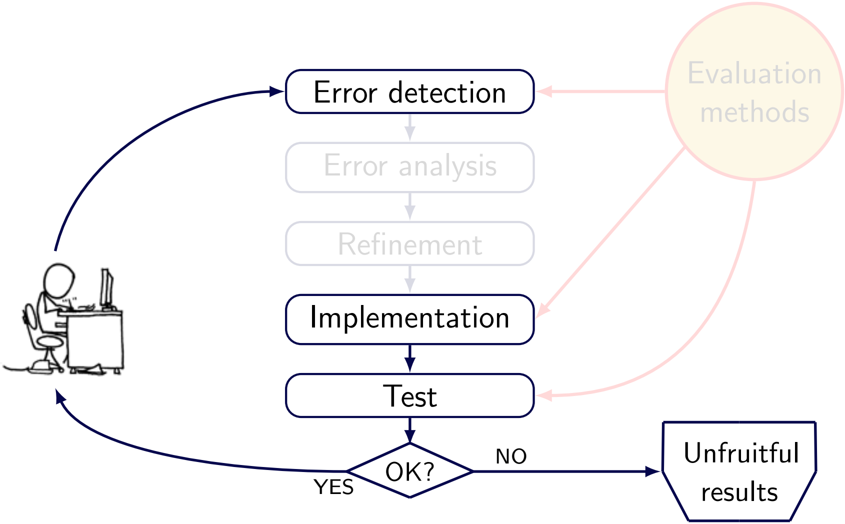
MT Evaluation

Importance for system development



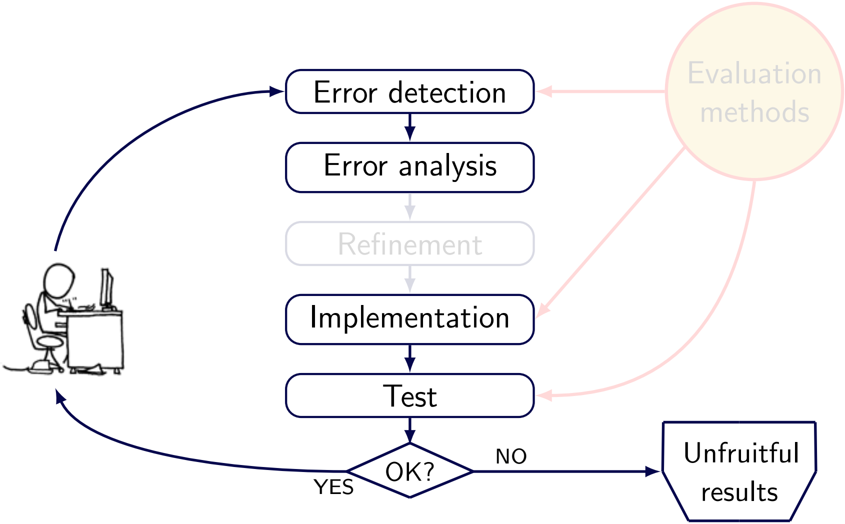
MT Evaluation

Importance for system development



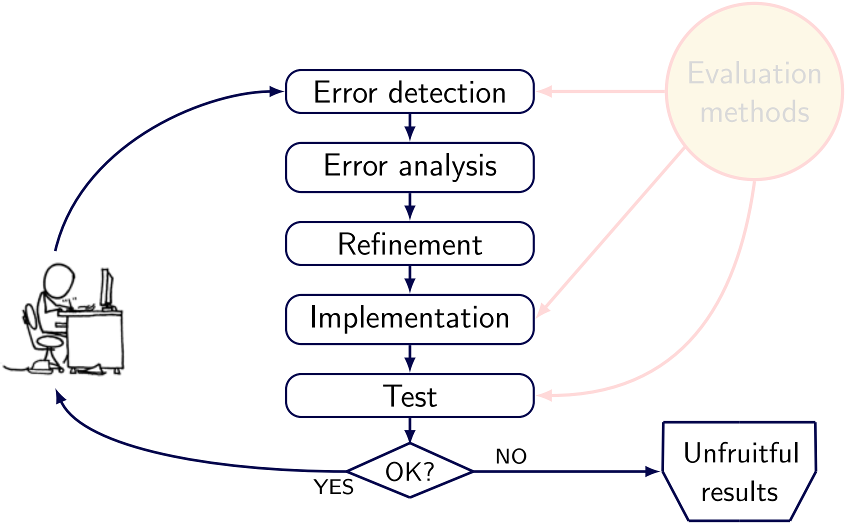
MT Evaluation

Importance for system development



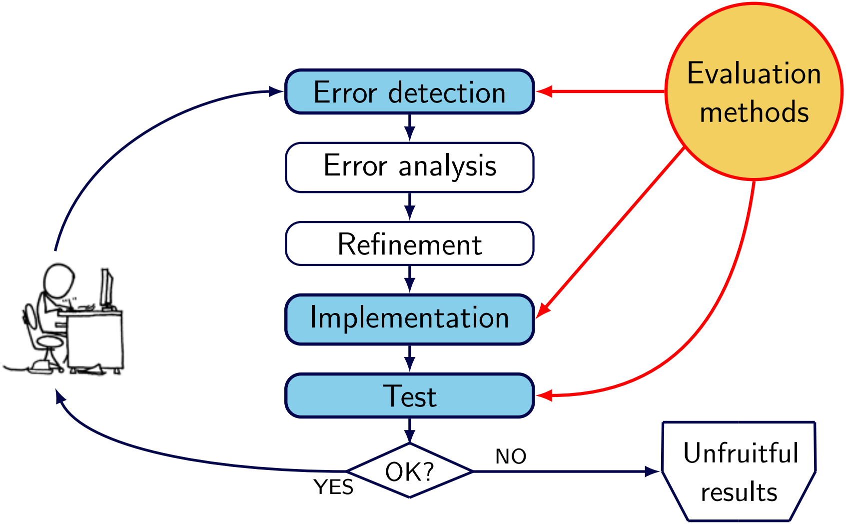
MT Evaluation

Importance for system development



MT Evaluation

Importance for system development



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)

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)

Risks of Automatic Evaluation

- **System overtuning**: when system parameters are adjusted towards a given metric
- **Blind system development**: when metrics are unable to capture actual system improvements
- **Unfair system comparisons**: when metrics are unable to reflect difference in quality between MT systems

Machine Translation is an open NLP task

- The correct translation is not unique
- The set of valid translations is not small
- Translation correctness is not black and white
- Quality aspects are heterogeneous

Adequacy (or Fidelity) Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

Fluency (or Intelligibility) Is the output fluent? This involves both grammatical correctness and idiomatic word choices.

Post–edition effort Time required to *repair* the translation, number of key strokes, etc.

Outline

Likert scales – TAUS recommendation

Adequacy How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?

4	Everything
3	Most
2	Little
1	None

Fluency To what extent is a target side translation grammatically well informed, without spelling errors and experienced as using natural/intuitive language by a native speaker?

4	Flawless
3	Good
2	Disfluent
1	Incomprehensible

https://www.taus.net/think-tank/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

Likert scales – NIST example

Adequacy I How much of the meaning expressed in the Reference translation is also expressed in the System translation?

7-point scale ranging from 1 (None) to 7 (All)

Adequacy II Does the Machine translation mean essentially the same as the Reference translation?

Yes/No, Adequacy I > 4
No, Adequacy II ≤ 4

http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/results/correlationResults.html

Ranking – Pair-wise comparison

Annotators chose the best system, given the source and target sentence, and 2 anonymised random systems.

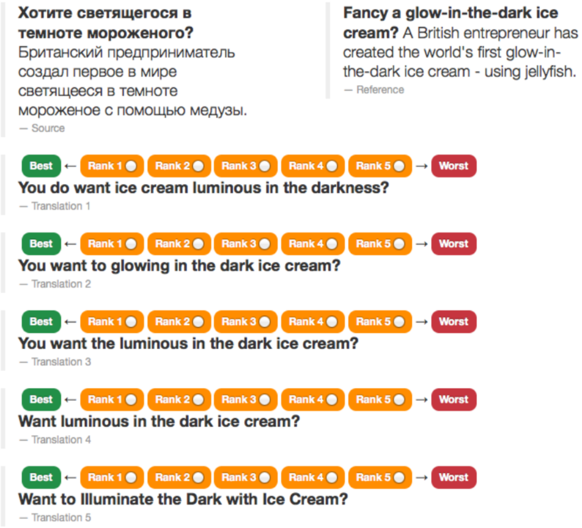
Ranking

Annotators rank *n* anonymised systems, randomly selected and randomly ordered.

Manual Evaluation

Appraise

Appraise
(Federmann 2012)



"**Appraise** is an open-source tool for manual evaluation of Machine Translation output."

Appraise allows to collect **human judgments** on translation output, implementing annotation tasks such as

- translation quality checking;
- ranking of translations;
- error classification;
- manual post-editing.

Manual Evaluation
Pros & Cons

- Likert scales have to be defined
- 4-, 5-, 7, 10-point likert scales have been used
- The concept of ranking is easy
- Ranks provide less information
- Agreement among annotators (common!)

Cohen's kappa coefficient, κ (Cohen, 1960)

$$\kappa = \frac{Pr(\text{agreement}) - Pr(\text{expected})}{1 - Pr(\text{expected})}$$

Kappa **interpretation** (Landis & Kogh, 1977)

0.0–0.2	slight
0.2–0.4	fair
0.4–0.6	moderate
0.6–0.8	substantial
0.8–1.0	almost perfect

Workshop on statistical machine translation, **WMT13**

- Inter-κ only slight or fair
- Even Intra-κ only fair or moderate

	Inter-κ	Intra-κ
CZ–EN	0.244	0.479
EN–CZ	0.168	0.290
DE–EN	0.299	0.535
EN–DE	0.267	0.498
ES–EN	0.277	0.575
EN–ES	0.206	0.492
FR–EN	0.275	0.578
EN–FR	0.231	0.495
RU–EN	0.278	0.450
EN–RU	0.243	0.513

Human-targeted Translation Error Rate, HTER

Annotator Post-edition of the candidate translation to have the same meaning as a reference translation with as few edits as possible

Evaluation TER with the candidate translation and the post-edited reference

$$HTER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}$$

Outline

Setting Compute **similarity** between system's output and one or several reference translations

Challenge The similarity measure should be able to discriminate whether the two sentences convey the same meaning (**semantic equivalence**)

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

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.

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
IBM Research Division

"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."

Automatic 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.`

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.`

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.`

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.
```

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: Prec. $= \frac{1+}{7}$

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.`

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{2+}{7}$

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.`

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{3+}{7}$

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.`

Modified n-gram precision (1-gram)

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

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.`

Modified n-gram precision (1-gram)

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

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.`

Modified n-gram precision (1-gram)

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

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.`

Modified n-gram precision (1-gram)

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

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.`

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 minimun 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.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \rightarrow \mathrm{The}$
   $\#w_{i,R1} = 2$
   $\#w_{i,R2} = 1$
2. $\mathrm{Max}_{(1)}$=2, $\#w_{i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

Modified n-gram precision (1-gram)

Modified 1-gram precision: $\mathrm{P}_1 =$

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. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}$=2, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 = \frac{2}{}$

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. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}$=2, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

Modified n-gram precision (1-gram)

Modified 1-gram precision: $\mathsf{P}_1 = \frac{2}{7}$

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. $w_i \to \text{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\text{Max}_{(1)}=2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

Modified n-gram precision

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

$$\mathrm{P}_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} Count_{\text{clipped}}(n\text{gram})}{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} Count(n\text{gram})}$$

low n	high n
adequacy	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`

$P_1 = 2/2$, $P_2 = 1/1$

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

$$\mathrm{BP} = \begin{cases} 1 & \text{if } \; c > r \\ e^{1-r/c} & \text{if } \; c \leq r \end{cases}$$

c candidate length, r reference length

- Multiplicative factor
- At sentence level, huge punishment for short sentences
- Estimated at document level

BiLingual Evaluation Understudy, BLEU

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log \mathrm{P}_n\right)$$

- Geometric average of P_n (empirical suggestion)
- w_n positive weights summing to one
- Brevity penalty

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/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.

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

Beyond 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.

Extending the reference material
METEOR, Banerjee and Lavie (2005)

Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P***recision** and *R***ecall**
weighted harmonic mean

$$Pen = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}}\right)^\beta$$

Penalty factor, penalises
non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P***recision** and *R***ecall** weighted harmonic mean

$$Pen = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta$$

Penalty factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

Limits of lexical similarity
Beyond 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.

Limits of lexical similarity

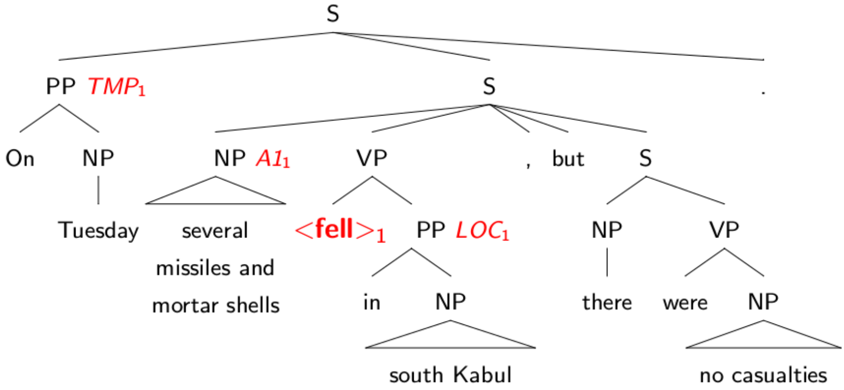
Comparing other linguistic features than words

Candidate:

`On Tuesday several missiles and mortar shells fell in south Kabul, but there were no casualties.`

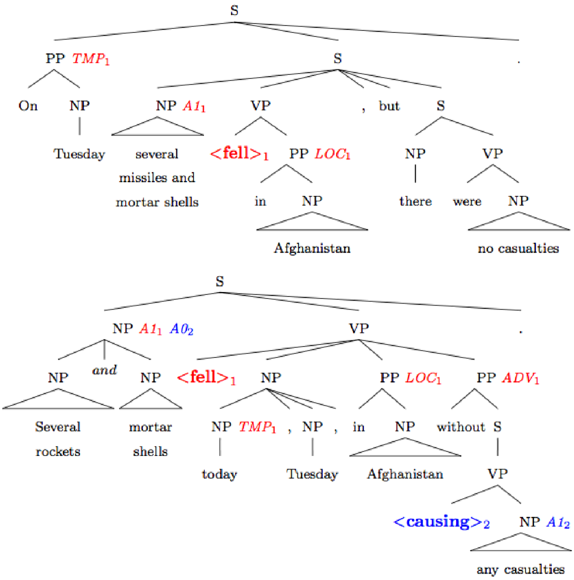
Reference:

`Several rockets and mortar shells fell today, Tuesday, in south Kabul without causing any casualties.`


S
PP TMP1
S
.
On
NP
NP A1₁
VP
,
but
S
Tuesday
several missiles and mortar shells
<fell>1
PP LOC1
NP
VP
in
NP
there
were
NP
south Kabul
no casualties

Limits of lexical similarity

Comparing other linguistic features than words



Overlap

Generic similarity measure among Linguistic Elements.
Inspired by the Jaccard similarity coefficient.

Linguistic element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.

- For instance: POS tags, word lemmas, NPs, syntactic phrases
- A sentence can be seen as a bag (or a sequence) of LEs of a certain type
- LEs may embed

Overlap

Generic similarity measure among Linguistic Elements.
Inspired by the Jaccard similarity coefficient.

Linguistic element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.

- For instance: POS tags, word lemmas, NPs, syntactic phrases
- A sentence can be seen as a bag (or a sequence) of LEs of a certain type
- LEs may embed

Limits of lexical similarity

Comparing other linguistic features than words

$$\mathrm{O}(t) = \frac{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{cand})\ \cap\ \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{cand}}(i, t)}{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{cand})\ \cup\ \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{cand}}(i, t), \mathrm{count}_{\mathrm{ref}}(i, t))}$$

t is the LE type

'cand': candidate translation

'ref': reference translation

$\mathrm{items}_t(s)$: set of items occurring inside LEs of type t

$\mathrm{count}_s(i, t)$: occurrences of item i in s inside a LE of type t

Limits of lexical similarity

Comparing other linguistic features than words

Coarser variant: **micro-averaged overlap over all types**

$$\mathrm{O}(\star) = \frac{\displaystyle\sum_{t \in T} \sum_{i \in (\mathrm{items}_t(\mathrm{cand}) \,\cap\, \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{cand}}(i, t)}{\displaystyle\sum_{t \in T} \sum_{i \in (\mathrm{items}_t(\mathrm{cand}) \,\cup\, \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{cand}}(i, t), \mathrm{count}_{\mathrm{ref}}(i, t))}$$

T: set of all LE types associated to the given LE class

Extend the reference material:

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

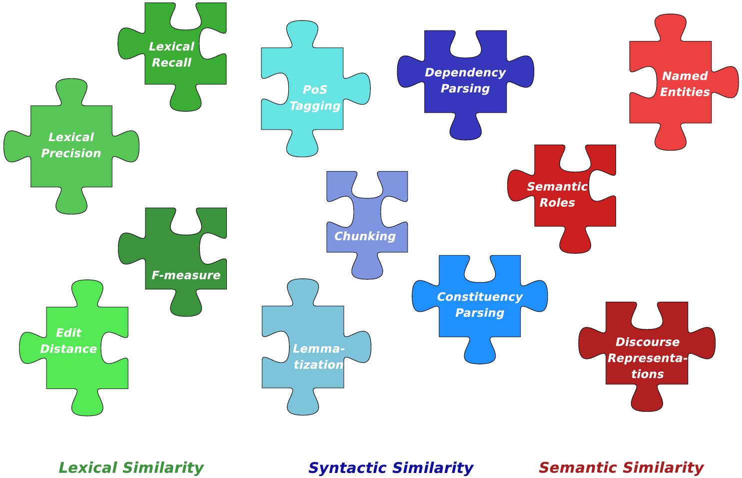
Compare other linguistic features than words:

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

Combination of the existing metrics.

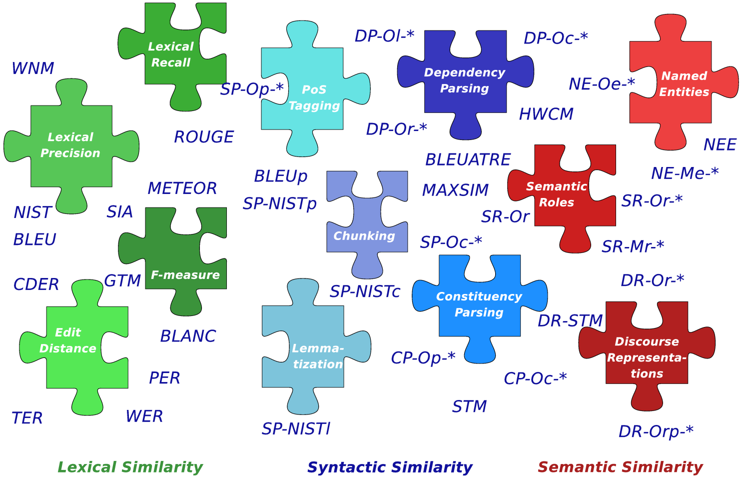
Limits of lexical similarity

Combination of the existing metrics



Limits of lexical similarity

Combination of the existing metrics



Limits of lexical similarity

Combination of the existing metrics

- Different measures capture **different aspects** of similarity suitable for combination
- The most simple approach: **ULC**

Uniformly averaged **linear combination** of measures (ULC):

$$\mathrm{ULC}_M(\mathrm{cand}, \mathrm{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\mathrm{cand}, \mathrm{ref})$$

- Different measures capture **different aspects** of similarity suitable for combination
- The most simple approach: **ULC**

Uniformly averaged **linear combination** of measures (ULC):

$$\mathrm{ULC}_M(\mathrm{cand}, \mathrm{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\mathrm{cand}, \mathrm{ref})$$

MT Evaluation

MT Evaluation: keep in mind

- Evaluation is important in the system development cycle. Automatic evaluation accelerates significatively the process.
- Manual evaluation is still necessary but shows low agreements among annotators
- 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.

Outline

Evaluate your translations

1. With BLEU scoring tool. Available as a `Moses` script or from NIST:
   ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl
2. With `Asiya` package:
   http://nlp.lsi.upc.edu/asiya/

Asiya

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/

Tools

In practice

1. With BLEU scoring tool in `Moses`:

```
moses/scripts/generic/multi-bleu.perl references.en <
testset.translated.en
```

Tools

In practice

2. 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
#-------------------------------------------------------
```

2. 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
#-------------------------------------------------------
```

Tools

In practice

```
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-Oc-* CP-Op-* CP-STM-9 DP-HWC-c-4
DP-HWC-r-4 DP-HWC-w-4 DP-Oc-* DP-Ol-* 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-Oc-* SP-Op-* SP-cNIST-5
SP-iobNIST-5 SP-lNIST-5 SP-pNIST-5 SR-Mr-* SR-Mrv-* SR-Or SR-Or-* SR-Orv
```

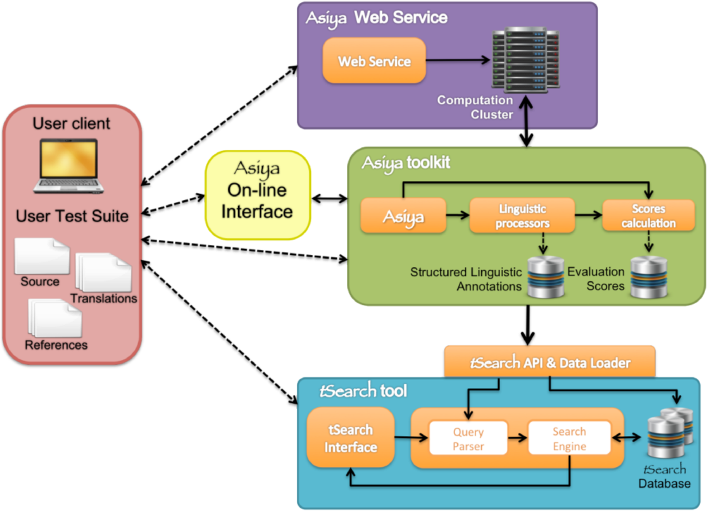
Tools

In practice

```
------------------------------------------------------------------------------------------------------------------------
METRIC NAMES
------------------------------------------------------------------------------------------------------------------------
668 metrics are available for language 'en'

METRICS = { -PER, -TER, -TERbase, -TERp, -TERp-A, -WER, BLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, CP-Oc(*), CP-Oc(ADJP), CP-Oc(ADVP), CP-Oc(CONJP), CP-Oc(FRA
G), CP-Oc(INTJ), CP-Oc(LST), CP-Oc(NAC), CP-Oc(NP), CP-Oc(NX), CP-Oc(O), CP-Oc(PP), CP-Oc(PRN), CP-Oc(PRT), CP-Oc(QP), CP-Oc(RRC), CP-Oc(S), CP-Oc(SBAR), CP-Oc(SINV), CP-Oc(SQ), CP
-Oc(UCP), CP-Oc(VP), CP-Oc(WHADJP), CP-Oc(WHADVP), CP-Oc(WHNP), CP-Oc(WHPP), CP-Oc(X), CP-Op(#), CP-Op($), CP-Op(''), CP-Op((), CP-Op()), CP-Op(*), CP-Op(,), CP-Op(.), CP-Op(:), CP
-Op(CC), CP-Op(CD), CP-Op(DT), CP-Op(EX), CP-Op(F), CP-Op(FW), CP-Op(IN), CP-Op(J), CP-Op(JJ), CP-Op(JJR), CP-Op(JJS), CP-Op(LS), CP-Op(MD), CP-Op(N), CP-Op(NN), CP-Op(NNP), CP-Op(
NNPS), CP-Op(NNS), CP-Op(P), CP-Op(PDT), CP-Op(POS), CP-Op(PRP$), CP-Op(PRP), CP-Op(R), CP-Op(RB), CP-Op(RBR), CP-Op(RBS), CP-Op(RP), CP-Op(SYM), CP-Op(TO), CP-Op(UH), CP-Op(V), CP
-Op(VB), CP-Op(VBD), CP-Op(VBG), CP-Op(VBN), CP-Op(VBP), CP-Op(VBZ), CP-Op(W), CP-Op(WDT), CP-Op(WP$), CP-Op(WP), CP-Op(WRB), CP-Op(``), CP-STM-1, CP-STM-2, CP-STM-3, CP-STM-4, CP-
STM-5, CP-STM-6, CP-STM-7, CP-STM-8, CP-STM-9, CP-STMi-2, CP-STMi-3, CP-STMi-4, CP-STMi-5, CP-STMi-6, CP-STMi-7, CP-STMi-8, CP-STMi-9, DP-HWCM_c-1, DP-HWCM_c-2, DP-HWCM_c-3, DP-HWC
M_c-4, DP-HWCM_r-1, DP-HWCM_r-2, DP-HWCM_r-3, DP-HWCM_r-4, DP-HWCM_w-1, DP-HWCM_w-2, DP-HWCM_w-3, DP-HWCM_w-4, DP-HWCMi_c-2, DP-HWCMi_c-3, DP-HWCMi_c-4, DP-HWCMi_r-2, DP-HWCMi_r-3,
 DP-HWCMi_r-4, DP-HWCMi_w-2, DP-HWCMi_w-3, DP-HWCMi_w-4, DP-Oc(*), DP-Oc(a), DP-Oc(as), DP-Oc(aux), DP-Oc(be), DP-Oc(c), DP-Oc(comp), DP-Oc(det), DP-Oc(have), DP-Oc(n), DP-Oc(postd
et), DP-Oc(ppspec), DP-Oc(predet), DP-Oc(prep), DP-Oc(saidx), DP-Oc(sentadjunct), DP-Oc(subj), DP-Oc(that), DP-Oc(u), DP-Oc(v), DP-Oc(vbe), DP-Oc(xsaid), DP-Ol(*), DP-Ol(1), DP-Ol(
2), DP-Ol(3), DP-Ol(4), DP-Ol(5), DP-Ol(6), DP-Ol(7), DP-Ol(8), DP-Ol(9), DP-Or(*), DP-Or(amod), DP-Or(amount-value), DP-Or(appo), DP-Or(appo-mod), DP-Or(as-arg), DP-Or(as1), DP-Or
(as2), DP-Or(aux), DP-Or(be), DP-Or(being), DP-Or(by-subj), DP-Or(c), DP-Or(cn), DP-Or(comp1), DP-Or(conj), DP-Or(desc), DP-Or(dest), DP-Or(det), DP-Or(else), DP-Or(fc), DP-Or(gen)
, DP-Or(guest), DP-Or(have), DP-Or(head), DP-Or(i), DP-Or(inv-aux), DP-Or(inv-have), DP-Or(lex-dep), DP-Or(lex-mod), DP-Or(mod), DP-Or(mod-before), DP-Or(neg), DP-Or(nn), DP-Or(num
), DP-Or(num-mod), DP-Or(obj), DP-Or(obj1), DP-Or(obj2), DP-Or(p), DP-Or(p-spec), DP-Or(pcomp-c), DP-Or(pcomp-n), DP-Or(person), DP-Or(pnmod), DP-Or(poss), DP-Or(post), DP-Or(pre),
 DP-Or(pred), DP-Or(punc), DP-Or(rel), DP-Or(s), DP-Or(sc), DP-Or(subcat), DP-Or(subclass), DP-Or(subj), DP-Or(title), DP-Or(vrel), DP-Or(wha), DP-Or(whn), DP-Or(whp), DPm-HWCM_c-1
, DPm-HWCM_c-2, DPm-HWCM_c-3, DPm-HWCM_c-4, DPm-HWCM_r-1, DPm-HWCM_r-2, DPm-HWCM_r-3, DPm-HWCM_r-4, DPm-HWCM_w-1, DPm-HWCM_w-2, DPm-HWCM_w-3, DPm-HWCM_w-4, DPm-HWCMi_c-2, DPm-HWCMi
_c-3, DPm-HWCMi_c-4, DPm-HWCMi_r-2, DPm-HWCMi_r-3, DPm-HWCMi_r-4, DPm-HWCMi_w-2, DPm-HWCMi_w-3, DPm-HWCMi_w-4, DPm-Oc(*), DPm-Oc(......), DPm-Ol(*), DPm-Ol(1), DPm-Ol(2), DPm-Ol(3)
, DPm-Ol(4), DPm-Ol(5), DPm-Ol(6), DPm-Ol(7), DPm-Ol(8), DPm-Ol(9), DPm-Or(*), DPm-Or(.......), DR-Fr(*), DR-Frp(*), DR-Ol, DR-Or(*), DR-Or(*)_b, DR-Or(*)_i, DR-Or(alfa), DR-Or(car
d), DR-Or(drs), DR-Or(eq), DR-Or(imp), DR-Or(merge), DR-Or(named), DR-Or(not), DR-Or(or), DR-Or(pred), DR-Or(prop), DR-Or(rel), DR-Or(smerge), DR-Or(timex), DR-Or(whq), DR-Or-(dr),
 DR-Orp(*), DR-Orp(*)_b, DR-Orp(*)_i, DR-Orp(alfa), DR-Orp(card), DR-Orp(dr), DR-Orp(drs), DR-Orp(eq), DR-Orp(imp), DR-Orp(merge), DR-Orp(named), DR-Orp(not), DR-Orp(or), DR-Orp(pr
ed), DR-Orp(prop), DR-Orp(rel), DR-Orp(smerge), DR-Orp(timex), DR-Orp(whq), DR-Pr(*), DR-Prp(*), DR-Rr(*), DR-Rrp(*), DR-STM-1, DR-STM-2, DR-STM-3, DR-STM-4, DR-STM-4_b, DR-STM-4_i
, DR-STM-5, DR-STM-6, DR-STM-7, DR-STM-8, DR-STM-9, DR-STMi-2, DR-STMi-3, DR-STMi-4, DR-STMi-5, DR-STMi-6, DR-STMi-7, DR-STMi-8, DR-STMi-9, DRdoc-Ol, DRdoc-Or(*), DRdoc-Or(*)_b, DR
doc-Or(*)_i, DRdoc-Or(alfa), DRdoc-Or(card), DRdoc-Or(dr), DRdoc-Or(drs), DRdoc-Or(eq), DRdoc-Or(imp), DRdoc-Or(merge), DRdoc-Or(named), DRdoc-Or(not), DRdoc-Or(or), DRdoc-Or(pred)
, DRdoc-Or(prop), DRdoc-Or(rel), DRdoc-Or(smerge), DRdoc-Or(timex), DRdoc-Or(whq), DRdoc-Orp(*), DRdoc-Orp(*)_b, DRdoc-Orp(*)_i, DRdoc-Orp(alfa), DRdoc-Orp(card), DRdoc-Orp(dr), DR
doc-Orp(drs), DRdoc-Orp(eq), DRdoc-Orp(imp), DRdoc-Orp(merge), DRdoc-Orp(named), DRdoc-Orp(not), DRdoc-Orp(or), DRdoc-Orp(pred), DRdoc-Orp(prop), DRdoc-Orp(rel), DRdoc-Orp(smerge),
 DRdoc-Orp(timex), DRdoc-Orp(whq), DRdoc-STM-1, DRdoc-STM-2, DRdoc-STM-3, DRdoc-STM-4, DRdoc-STM-4_b, DRdoc-STM-4_i, DRdoc-STM-5, DRdoc-STM-6, DRdoc-STM-7, DRdoc-STM-8, DRdoc-STM-9
, DRdoc-STMi-2, DRdoc-STMi-3, DRdoc-STMi-4, DRdoc-STMi-5, DRdoc-STMi-6, DRdoc-STMi-7, DRdoc-STMi-8, DRdoc-STMi-9, Fl, GTM-1, GTM-2, GTM-3, METEOR-ex, METEOR-pa, METEOR-st, METEOR-s
y, NE-Me(*), NE-Me(ANGLE_QUANTITY), NE-Me(DATE), NE-Me(DISTANCE_QUANTITY), NE-Me(LANGUAGE), NE-Me(LOC), NE-Me(MEASURE), NE-Me(METHOD), NE-Me(MISC), NE-Me(MONEY), NE-Me(NUM), NE-Me(
ORG), NE-Me(PER), NE-Me(PERCENT), NE-Me(PROJECT), NE-Me(SIZE_QUANTITY), NE-Me(SPEED_QUANTITY), NE-Me(SYSTEM), NE-Me(TEMPERATURE_QUANTITY), NE-Me(TIME), NE-Me(WEIGHT_QUANTITY), NE-O
e(*), NE-Oe(**), NE-Oe(ANGLE_QUANTITY), NE-Oe(DATE), NE-Oe(DISTANCE_QUANTITY), NE-Oe(LANGUAGE), NE-Oe(LOC), NE-Oe(MEASURE), NE-Oe(METHOD), NE-Oe(MISC), NE-Oe(MONEY), NE-Oe(NUM), NE
-Oe(O), NE-Oe(ORG), NE-Oe(PER), NE-Oe(PERCENT), NE-Oe(PROJECT), NE-Oe(SIZE_QUANTITY), NE-Oe(SPEED_QUANTITY), NE-Oe(SYSTEM), NE-Oe(TEMPERATURE_QUANTITY), NE-Oe(TIME), NE-Oe(WEIGHT_Q
UANTITY), NIST, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5, Ol, Pl, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W, Rl, S
P-Oc(*), SP-Oc(ADJP), SP-Oc(ADVP), SP-Oc(CONJP), SP-Oc(INTJ), SP-Oc(LST), SP-Oc(NP), SP-Oc(O), SP-Oc(PP), SP-Oc(PRT), SP-Oc(SBAR), SP-Oc(UCP), SP-Oc(VP), SP-Op(#), SP-Op($), SP-Op(
''), SP-Op((), SP-Op()), SP-Op(*), SP-Op(,), SP-Op(.), SP-Op(:), SP-Op(CC), SP-Op(CD), SP-Op(DT), SP-Op(EX), SP-Op(F), SP-Op(FW), SP-Op(IN), SP-Op(J), SP-Op(JJ), SP-Op(JJR), SP-Op(
JJS), SP-Op(LS), SP-Op(MD), SP-Op(N), SP-Op(NN), SP-Op(NNP), SP-Op(NNPS), SP-Op(NNS), SP-Op(P), SP-Op(PDT), SP-Op(POS), SP-Op(PRP$), SP-Op(PRP), SP-Op(R), SP-Op(RB), SP-Op(RBR), SP
-Op(RBS), SP-Op(RP), SP-Op(SYM), SP-Op(TO), SP-Op(UH), SP-Op(V), SP-Op(VB), SP-Op(VBD), SP-Op(VBG), SP-Op(VBN), SP-Op(VBP), SP-Op(VBZ), SP-Op(W), SP-Op(WDT), SP-Op(WP$), SP-Op(WP),
 SP-Op(WRB), SP-Op(``), SP-cNIST, SP-cNIST-1, SP-cNIST-2, SP-cNIST-3, SP-cNIST-4, SP-cNIST-5, SP-cNISTi-2, SP-cNISTi-3, SP-cNISTi-4, SP-cNISTi-5, SP-iobNIST, SP-iobNIST-1, SP-iobNI
ST-2, SP-iobNIST-3, SP-iobNIST-4, SP-iobNIST-5, SP-iobNISTi-2, SP-iobNISTi-3, SP-iobNISTi-4, SP-iobNISTi-5, SP-lNIST, SP-lNIST-1, SP-lNIST-2, SP-lNIST-3, SP-lNIST-4, SP-lNIST-5, SP
-lNISTi-2, SP-lNISTi-3, SP-lNISTi-4, SP-lNISTi-5, SP-pNIST, SP-pNIST-1, SP-pNIST-2, SP-pNIST-3, SP-pNIST-4, SP-pNIST-5, SP-pNISTi-2, SP-pNISTi-3, SP-pNISTi-4, SP-pNISTi-5, SR-Fr(*)
, SR-MFr(*), SR-MPr(*), SR-MRr(*), SR-Mr(*), SR-Mr(*)_b, SR-Mr(*)_i, SR-Mr(A0), SR-Mr(A1), SR-Mr(A2), SR-Mr(A3), SR-Mr(A4), SR-Mr(A5), SR-Mr(AA), SR-Mr(AM-ADV), SR-Mr(AM-CAU), SR-M
r(AM-DIR), SR-Mr(AM-DIS), SR-Mr(AM-EXT), SR-Mr(AM-LOC), SR-Mr(AM-MNR), SR-Mr(AM-MOD), SR-Mr(AM-NEG), SR-Mr(AM-PNC), SR-Mr(AM-PRD), SR-Mr(AM-REC), SR-Mr(AM-TMP), SR-Mrv(*), SR-Mrv(*
)_b, SR-Mrv(*)_i, SR-Mrv(A0), SR-Mrv(A1), SR-Mrv(A2), SR-Mrv(A3), SR-Mrv(A4), SR-Mrv(A5), SR-Mrv(AA), SR-Mrv(AM-ADV), SR-Mrv(AM-CAU), SR-Mrv(AM-DIR), SR-Mrv(AM-DIS), SR-Mrv(AM-EXT)
, SR-Mrv(AM-LOC), SR-Mrv(AM-MNR), SR-Mrv(AM-MOD), SR-Mrv(AM-NEG), SR-Mrv(AM-PNC), SR-Mrv(AM-PRD), SR-Mrv(AM-REC), SR-Mrv(AM-TMP), SR-Nv, SR-Ol, SR-Or, SR-Or(*), SR-Or(*)_b, SR-Or(*
)_i, SR-Or(A0), SR-Or(A1), SR-Or(A2), SR-Or(A3), SR-Or(A4), SR-Or(A5), SR-Or(AA), SR-Or(AM-ADV), SR-Or(AM-CAU), SR-Or(AM-DIR), SR-Or(AM-DIS), SR-Or(AM-EXT), SR-Or(AM-LOC), SR-Or(AM
-MNR), SR-Or(AM-MOD), SR-Or(AM-NEG), SR-Or(AM-PNC), SR-Or(AM-PRD), SR-Or(AM-REC), SR-Or(AM-TMP), SR-Or_b, SR-Or_i, SR-Orv, SR-Orv(*), SR-Orv(*)_b, SR-Orv(*)_i, SR-Orv(A0), SR-Orv(A
1), SR-Orv(A2), SR-Orv(A3), SR-Orv(A4), SR-Orv(A5), SR-Orv(AA), SR-Orv(AM-ADV), SR-Orv(AM-CAU), SR-Orv(AM-DIR), SR-Orv(AM-DIS), SR-Orv(AM-EXT), SR-Orv(AM-LOC), SR-Orv(AM-MNR), SR-O
rv(AM-MOD), SR-Orv(AM-NEG), SR-Orv(AM-PNC), SR-Orv(AM-PRD), SR-Orv(AM-REC), SR-Orv(AM-TMP), SR-Orv_b, SR-Orv_i, SR-Ov, SR-Pr(*), SR-Rr(*) }
```

Asiya interfaces



Evaluate the results on-line

1. Asiya Interface

   http://asiya.lsi.upc.edu/demo/asiya_online.php

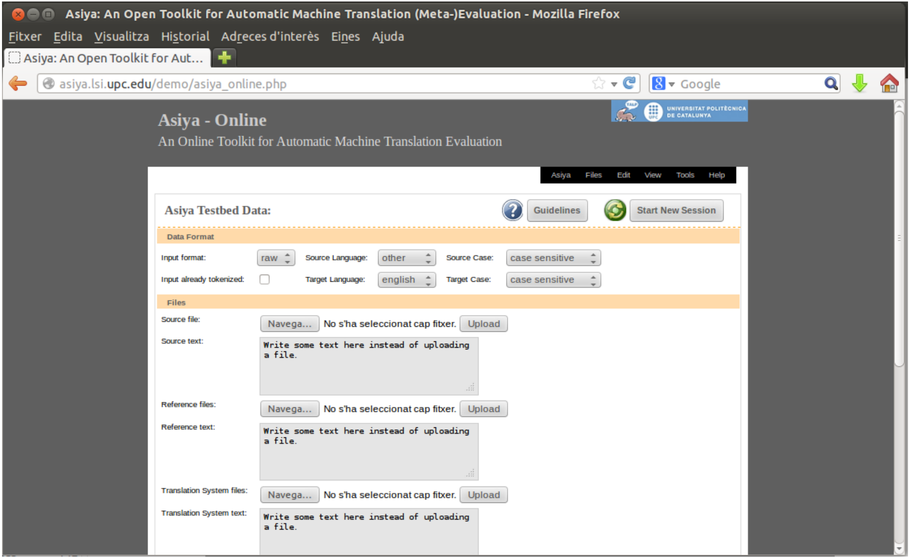
Analise the results on-line

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 III

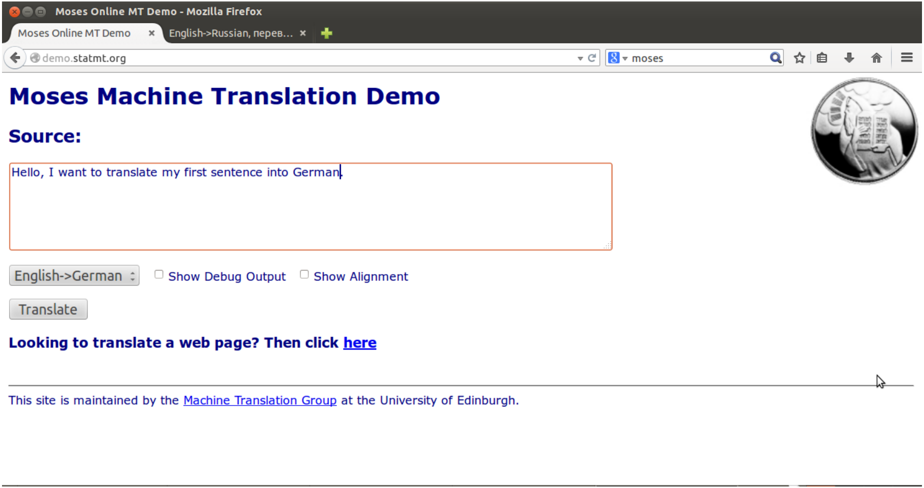
SMT experiments

Outline Part III

10. Translation system
    - Demos
    - Software
    - Steps

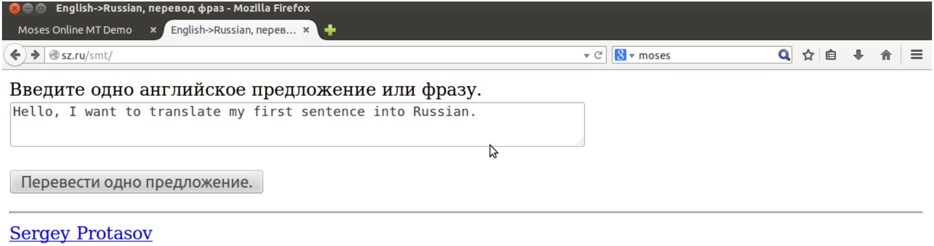
SMT system

Demo: http://demo.statmt.org/



SMT system

Demo: http://sz.ru/smt/



Build your own SMT system

1. Language model with `SRILM`.
   http://www-speech.sri.com/projects/srilm/download.html

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

1. 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)

```
wmt10scripts/tokenizer.perl -l es < eurov4.es-en.NOTOK.es >
eurov4.es-en.TOK.es
wmt10scripts/tokenizer.perl -l en < eurov4.es-en.NOTOK.en >
eurov4.es-en.TOK.en
```

```
wmt10scripts/tokenizer.perl -l es < eurov4.es-en.NOTOK.dev.es >
eurov4.es-en.TOK.dev.es
wmt10scripts/tokenizer.perl -l en < eurov4.es-en.NOTOK.dev.en >
eurov4.es-en.TOK.dev.en
```

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)

```
wmt10scripts/lowercase.perl < eurov4.es-en.TOK.clean.es >
eurov4.es-en.es
wmt10scripts/lowercase.perl < eurov4.es-en.TOK.clean.en >
eurov4.es-en.en
```

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:  (--first-step to --last-step)
(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)

1. 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!)

```
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:

```
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 lowecase 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
   ```

Part IV

Appendix: Classical References

Classical 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]

Classical 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]

Classical References

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]

Classical References

Language model

- Kneser & Ney, 1995 [KN95]

MERT

- Och, 2003 [Och03]

Domain adaptation

- Bertoldi and Federico, 2009 [Och03]

Classical References

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]

Classical References

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]

Classical References

Manual Evaluation

- Cohen, 1960 [Coh60]
- Landis & Koch, 1977 [LK77]
- Federmann 2012 [Fed12]

Classical References

Automatic Evaluation

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

Classical References

Metrics I

- WER [NOLN00]
- PER [TVN+97]
- TER [SDS+06]

Classical References

Metrics II

- BLEU [PRWZ02]
- NIST [Dod02]
- METEOR [BL05]
- ROUGE [LO04]

Metrics III

- GTM [MGT03]
- BLANC [Dod02]
- CDER [LUN06]
- ULC [GA06]

Classical References

Surveys, theses and tutorials

- Knight, 1999

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

- Knight & Koehn, 2003

  http://people.csail.mit.edu/people/koehn/publications/tutorial2003.pdf

- Koehn, 2006

  http://www.iccs.informatics.ed.ac.uk/ pkoehn/publications/tutorial2006.pdf

- Way & Hassan, 2009

  http://www.medar.info/conference_all/2009/Tutorial_3.pdf

- Lopez, 2008 [Lop08]
- Giménez, 2009 [Gim08]

Classical References I

Automatic Language Processing Advisory Committee (ALPAC).
Language and Machines. Computers in Translation and Linguistics.
Technical Report Publication 1416, Division of Behavioural Sciences, National Academy of Sciences, National Research Council, Washington, D.C., 1966.

Phil Blunsom, Trevor Cohn, and Miles Osborne.
A discriminative latent variable model for statistical machine translation.
In *ACL-08: HLT. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 200–208, 2008.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin.
A statistical approach to machine translation.
Computational Linguistics, 16(2):79–85, 1990.

Nguyen Bach, Qin Gao, and Stephan Vogel.
Source-side dependency tree reordering models with subtree movements and constraints.
In *Proceedings of the Twelfth Machine Translation Summit (MTSummit-XII)*, Ottawa, Canada, August 2009. International Association for Machine Translation.

Classical References II

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak.
Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–159, 2007.

Satanjeev Banerjee and Alon Lavie.
METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.
In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer.
The mathematics of statistical machine translation: parameter estimation.
Computational Linguistics, 19(2):263–311, 1993.

Xavier Carreras and Michael Collins.
Non-projective parsing for statistical machine translation.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 200–209, Singapore, August 2009.

Classical References III

David Chiang.
A hierarchical phrase-based model for statistical machine translation.
In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

David Chiang, Kevin Knight, and Wei Wang.
11,001 new features for statistical machine translation.
In *NAACL '09: Human Language Technologies: the 2009 annual conference of the North American Chapter of the ACL*, pages 218–226. Association for Computational Linguistics, 2009.

Josep M^a Crego and José B. Mari no.
Improving smt by coupling reordering and decoding.
Machine Translation, 20(3):199–215, March 2006.

Jacob Cohen.
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46, 1960.

Classical References IV

Sander Canisius and Antal van den Bosch.
A constraint satisfaction approach to machine translation.
In Lluís Màrquez and Harold Somers, editors, *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 182–189, 2009.

Marine Carpuat and Dekai Wu.
Improving Statistical Machine Translation Using Word Sense Disambiguation.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–72, 2007.

Han-Bin Chen, Jian-Cheng Wu, and Jason S. Chang.
Learning bilingual linguistic reordering model for statistical machine translation.
In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 254–262, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Jinhua Du, Yanjun Ma, and Andy Way.
Source-side context-informed hypothesis alignment for combining outputs from Machine Translation systems.
In *Proceedings of the Machine Translation Summit XII*, pages 230–237, Ottawa, ON, Canada., 2009.

Classical References V

George Doddington.
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145, 2002.

Christian Federmann.
Appraise: An open-source toolkit for manual evaluation of machine translation output.
The Prague Bulletin of Mathematical Linguistics, 98:25–35, September 2012.

Andrew Finch and Eiichiro Sumita.
Bidirectional phrase-based statistical machine translation.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1132, Singapore, August 2009. Association for Computational Linguistics.

Jesús Giménez and Enrique Amigó.
IQMT: A Framework for Automatic Machine Translation Evaluation.
In *Proceedings of the 5th LREC*, pages 685–690, 2006.

Jeś Giménez.
Empirical Machine Translation and its Evaluation.
PhD thesis, Universitat Politècnica de Catalunya, July 2008.

Classical References VI

Jesús Giménez and Lluís Màrquez.
Discriminative Phrase Selection for SMT, pages 205–236.
NIPS Workshop Series. MIT Press, 2008.

Hany Hassan, Khalil Sima'an, and Andy Way.
A syntactified direct translation model with linear-time decoding.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1182–1191, Singapore, August 2009. Association for Computational Linguistics.

W. J. Hutchins.
Machine translation and machine-aided translation.
Journal of Documentation, 34(2):119–159, 1978.

Almut Silja Hildebrand and Stephan Vogel.
CMU system combination for WMT'09.
In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 47–50, Athens, Greece, March 2009. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang.
Factored Translation Models.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 868–876, 2007.

Classical References VII

R. Kneser and H. Ney.
Improved backing-off for m-gram language modeling.
icassp, 1:181–184, 1995.

Philipp Koehn, Franz Josef Och, and Daniel Marcu.
Statistical phrase-based translation.
In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1 2003.

Mu Li, Nan Duan, Dongdong Zhang, Chi-Ho Li, and Ming Zhou.
Collaborative decoding: Partial hypothesis re-ranking using translation consensus between decoders.
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 585–592, Suntec, Singapore, August 2009. Association for Computational Linguistics.

J. R. Landis and G. G. Koch.
The measurement of observer agreement for categorical data.
Biometrics, 33(1):159–174, 1977.

Classical References VIII

Chin-Yew Lin and Franz Josef Och.
Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics.
In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

Adam Lopez.
Statistical machine translation.
ACM Comput. Surv., 40(3), 2008.

Gregor Leusch, Nicola Ueffing, and Hermann Ney.
CDER: Efficient MT Evaluation Using Block Movements.
In *Proceedings of EACL*, pages 241–248, 2006.

I. Dan Melamed, Ryan Green, and Joseph P. Turian.
Precision and Recall of Machine Translation.
In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.

Classical References IX

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney.
An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research.
In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

Franz Josef Och.
Minimum error rate training in statistical machine translation.
In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003.

Franz Josef Och and Hermann Ney.
Discriminative Training and Maximum Entropy Models for Statistical Machine Translation.
In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, 2002.

Franz Josef Och and Hermann Ney.
The alignment template approach to statistical machine translation.
Computational Linguistics, 30(4):417–449, 2004.

Classical References X

Franz Josef Och, Christoph Tillmann, and Hermann Ney.
Improved alignment models for statistical machine translation.
In *Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June 1999.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the Association of Computational Linguistics*, pages 311–318, 2002.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, , and John Makhoul.
A Study of Translation Edit Rate with Targeted Human Annotation.
In *Proceedings of AMTA*, pages 223–231, 2006.

Jonathan Slocum.
A survey of machine translation: its history, current status, and future prospects.
Comput. Linguist., 11(1):1–17, 1985.

Classical References XI

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel.
Effective use of linguistic and contextual information for statistical machine translation.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 72–80, Singapore, August 2009. Association for Computational Linguistics.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf.
Accelerated DP based Search for Statistical Translation.
In *Proceedings of European Conference on Speech Communication and Technology*, 1997.

Warren Weaver.
Translation.
In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949/1955.
Reprinted from a memorandum written by Weaver in 1949.

Kenji Yamada and Kevin Knight.
A syntax-based statistical translation model.
In *Proceedings of the 39rd Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France, July 2001.