

# Statistical Machine Translation

Lucia Specia

`l.specia@sheffield.ac.uk`

LxMLS 2015

18 July 2015



The  
University  
Of  
Sheffield.

# Outline

- 1 Introduction
- 2 SMT
- 3 Evaluation
- 4 Success stories
- 5 Conclusions

Some slides from Wilker Aziz, Kevin Knight, Philipp Koehn, Adam Lopez

# Outline

- 1 Introduction
- 2 SMT
- 3 Evaluation
- 4 Success stories
- 5 Conclusions

# Introduction

- MT has been around since the early **1950s**

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec
- Availability of large collections of **data**, e.g. Europarl, TAUS data

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec
- Availability of large collections of **data**, e.g. Europarl, TAUS data
- More processing power



# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec
- Availability of large collections of **data**, e.g. Europarl, TAUS data
- More processing power
- Increasing **demand** for (cheap) translations - Google: 1 billion translations requests/day for 200 million users

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec
- Availability of large collections of **data**, e.g. Europarl, TAUS data
- More processing power
- Increasing **demand** for (cheap) translations - Google: 1 billion translations requests/day for 200 million users
- **Funding** for research worldwide

# Introduction

- MT has been around since the early **1950s**
- Increasingly popular since 1990: **statistical approaches**
- Software **toolkits** to build translation systems from data, e.g. Moses, cdec
- Availability of large collections of **data**, e.g. Europarl, TAUS data
- More processing power
- Increasing **demand** for (cheap) translations - Google: 1 billion translations requests/day for 200 million users
- **Funding** for research worldwide
- Exciting time for MT!

# The task of Machine Translation (MT)

The boy ate an apple

O menino comeu uma maçã

# The task of Machine Translation (MT)

The boy ate an apple

O menino comeu uma maçã

BUT

He said that the bottle floated into the cave

? Dijo que la botella entro a la cueva flotando

# The task of Machine Translation (MT)

The boy ate an apple

O menino comeu uma maçã

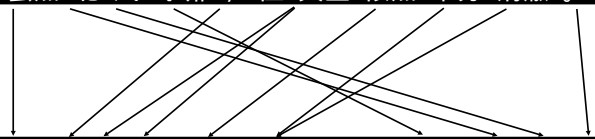
BUT

He said that the bottle floated into the cave

? Dijo que la botella entro a la cueva flotando

虽然 北 风 呼 啸 ， 但 天 空 依 然 十 分 清 澈 。

However, the sky remained clear under the strong north wind.



# Challenges in MT

- Lexical ambiguity
- Syntactic ambiguity
- Pronoun resolution
- Structural divergences

# Challenges in MT

- Lexical ambiguity
- Syntactic ambiguity
- Pronoun resolution
- Structural divergences
- Idioms

e.g. He finally **kicked the bucket** at the hospital  
→ Ele finalmente **bateu as botas** no hospital

- Multi-word expressions

e.g. Do **take** the long waiting list for organ donation in this country **into account**  
→ **Considerare** a longa lista de espera para doação de órgãos neste país



# Outline

- 1 Introduction
- 2 SMT**
- 3 Evaluation
- 4 Success stories
- 5 Conclusions

# Statistical Machine Translation

**Statistical Machine Translation** (SMT): “learn” how to generate translations from data

- Formalised early 1990s by IBM, but idea is much older:

Warren Weaver (1949)

When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

# Statistical Machine Translation

**Statistical Machine Translation** (SMT): “learn” how to generate translations from data

- Formalised early 1990s by IBM, but idea is much older:

Warren Weaver (1949)

When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

- Inspired by WWII code-breaking, and Shannon’s Information Theory
- Approach was not feasible with early computers

# Noisy Channel Model

## Noisy Channel Model

- Idea developed to model communication (Shannon)
  - e.g. communication over an imperfect phone line



- want to recover original message (here **word**) on basis of distorted signal received (here **noisy word**)

# Noisy Channel Model & SMT

- Output depends **probabilistically** on input
- To translate French ( $F$ ) into English ( $E$ ):

Given a French sentence  $F$ , search for English sentence  $E^*$  that maximises  $P(E|F)$

# Noisy Channel Model & SMT

- Find English sentence that maximizes  $P(E|F)$ , i.e.

# Noisy Channel Model & SMT

- Find English sentence that maximizes  $P(E|F)$ , i.e.

$$E^* = \operatorname{argmax}_E P(E|F)$$

# Noisy Channel Model & SMT

- Find English sentence that maximizes  $P(E|F)$ , i.e.

$$\begin{aligned} E^* &= \operatorname{argmax}_E P(E|F) \\ &= \operatorname{argmax}_E \frac{P(E) \cdot P(F|E)}{P(F)} \end{aligned} \quad \text{Bayes Rule}$$



# Noisy Channel Model & SMT

- Find English sentence that maximizes  $P(E|F)$ , i.e.

$$\begin{aligned} E^* &= \operatorname{argmax}_E P(E|F) \\ &= \operatorname{argmax}_E \frac{P(E) \cdot P(F|E)}{P(F)} \end{aligned} \quad \text{Bayes Rule}$$

- $P(F)$  constant across different  $E$ , so:

# Noisy Channel Model & SMT

- Find English sentence that maximizes  $P(E|F)$ , i.e.

$$\begin{aligned} E^* &= \operatorname{argmax}_E P(E|F) \\ &= \operatorname{argmax}_E \frac{P(E) \cdot P(F|E)}{P(F)} \quad \text{Bayes Rule} \end{aligned}$$

- $P(F)$  constant across different  $E$ , so:

$$E^* = \operatorname{argmax}_E P(E) \cdot P(F|E) \quad \text{drop } P(F)$$

# Noisy Channel Model & SMT

$$E^* = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E P(E) \cdot P(F|E)$$

- Decomposition of  $P(E|F)$  to  $P(E) \cdot P(F|E)$  breaks the problem in two parts:
  - $P(F|E)$  worries about picking E words that were likely used to generate F — **faithfulness**
  - $P(E)$  worries about picking E words that are likely to be said in English and that fit together — **fluency**

# Noisy Channel Model & SMT

$$E^* = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E P(E) \cdot P(F|E)$$

- Decomposition of  $P(E|F)$  to  $P(E) \cdot P(F|E)$  breaks the problem in two parts:
  - $P(F|E)$  worries about picking E words that were likely used to generate F — **faithfulness**
  - $P(E)$  worries about picking E words that are likely to be said in English and that fit together — **fluency**
- $P(E)$  and  $P(F|E)$  can be **trained independently**: **more reliable** model

# Main Components for Translation ( $F \mapsto E$ )

- **Translation model** (TM):  $P(F|E)$ 
  - **Faithfulness**: TMs created from (large) **parallel texts**

# Main Components for Translation ( $F \mapsto E$ )

- **Translation model** (TM):  $P(F|E)$ 
  - **Faithfulness**: TMs created from (large) **parallel texts**
- **Language model** (LM):  $P(E)$ 
  - **Fluency**: LMs created from large (fluent) **target language** texts

# Main Components for Translation ( $F \mapsto E$ )

- **Translation model** (TM):  $P(F|E)$ 
  - **Faithfulness**: TMs created from (large) **parallel texts**
- **Language model** (LM):  $P(E)$ 
  - **Fluency**: LMs created from large (fluent) **target language** texts
- A **Decoder**: ( $\text{argmax}$ )
  - Search algorithm to find  $E^*$

# Learning Translation Models - $P(F|E)$

- Requires a **sentence-aligned** bilingual corpus
  - e.g. European/Canadian/Hong Kong parliaments, subtitles, Bible, Web Crawl

here is advice for proceeding : gently excise this page and make it your bookmark .

Mr. Englund is a Swedish historian and journalist .

he is also the new permanent Secretary of the Swedish Academy , which awards the Nobel Prize in literature .

what he has written here is an unusual book , one he describes , not inaccurately , as " a work of anti @-@ history . "

it contains few big names , major Treaties or famous battles ; there are almost no ambassadors , dashing journalists or discussions of tactics and materiel .

it ' s not so much a book about what happened , he explains , as " a book about what it was like . " it ' s about " feelings , impressions , experiences and moods . "

" the beauty and the sorrow " threads together the wartime experiences of 20 more or less unremarkable men and women , on both sides of the war , from schoolgirls and botanists to mountain climbers , doctors , ambulance drivers and clerks .

a few of these people will become heroes .

a few will become prisoners of war , or lose limbs , go mad or die .

Aquí es consejos para proceder: Impuestos especiales amablemente esta página y dejar su bookmark.

Señor Englund es un historiador y periodista sueco.

También está el nuevo secretario permanente de la Academia Sueca, que otorga el premio Nobel de literatura.

Lo que él ha escrito aquí es una inusual, un libro que describe, no de forma poco precisa, como " un trabajo de anti-historia " .

Contiene algunas grandes nombres, principales Tratados o famoso batallas; no hay casi embajadores, corriendo periodistas o debates de tácticas y material.

No es tanto un libro sobre lo ocurrido, explica, como " un libro sobre lo que era como. " » sobre la " sentimientos, impresiones, experiencias y los estados de ánimo " .

" la belleza y el dolor " cabos, los tiempos de guerra experiencias de 20 más o menos brille hombres y mujeres, a ambos lados de la guerra, de las estudiantes y botanists de montaña climbers, los médicos, los conductores de ambulancias y burócratas.

Algunas de estas personas se convertirá en héroes.

Pocos serán los prisioneros de guerra, o perder extremidades, loca o morir.



# Learning Translation Models - $P(F|E)$

- Requires a **sentence-aligned** bilingual corpus
  - e.g. European/Canadian/Hong Kong parliaments, subtitles, Bible, Web Crawl

here is advice for proceeding : gently excise this page and make it your bookmark .

Mr. Englund is a Swedish historian and journalist .

he is also the new permanent Secretary of the Swedish Academy , which awards the Nobel Prize in literature .

what he has written here is an unusual book , one he describes , not inaccurately , as " a work of anti @-@ history . "

it contains few big names , major Treaties or famous battles ; there are almost no ambassadors , dashing journalists or discussions of tactics and materiel .

it ' s not so much a book about what happened , he explains , as " a book about what it was like . " it ' s about " feelings , impressions , experiences and moods . "

" the beauty and the sorrow " threads together the wartime experiences of 20 more or less unremarkable men and women , on both sides of the war , from schoolgirls and botanists to mountain climbers , doctors , ambulance drivers and clerks .

a few of these people will become heroes .

a few will become prisoners of war , or lose limbs , go mad or die .

Aquí es consejos para proceder: Impuestos especiales amablemente esta página y dejar su bookmark.

Señor Englund es un historiador y periodista sueco.

También está el nuevo secretario permanente de la Academia Sueca, que otorga el premio Nobel de literatura.

Lo que él ha escrito aquí es una inusual, un libro que describe, no de forma poco precisa, como " un trabajo de anti-historia " .

Contiene algunas grandes nombres, principales Tratados o famoso batallas; no hay casi embajadores, corriendo periodistas o debates de tácticas y material.

No es tanto un libro sobre lo ocurrido, explica, como " un libro sobre lo que era como. " » sobre la " sentimientos, impresiones, experiencias y los estados de ánimo " .

" la belleza y el dolor " cabos, los tiempos de guerra experiencias de 20 más o menos brille hombres y mujeres, a ambos lados de la guerra, de las estudiantes y botanists de montaña climbers, los médicos, los conductores de ambulancias y burócratas.

Algunas de estas personas se convertirá en héroes.

Pocos serán los prisioneros de guerra, o perder extremidades, loca o morir.

- Can we estimate  $P(F|E)$  from entire sentences?

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

## Zenish

Uh useh

Uh jeje

Yiguo useh

## English

A home

A garden

I arrived home

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words** are a good starting point
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

## Zenish

Uh **useh**

Uh jeije

Yiguo **useh**

- The same word happens in source 1 and 3

## English

A home

A garden

I arrived home

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

## Zenish

Uh **useh**

Uh jeije

Yiguo **useh**

- Could we expect the same in the target side?

## English

A home

A garden

I arrived home

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

## Zenish

Uh **useh**

Uh jeije

Yiguo **useh**

## English

A **home**

A garden

I arrived **home**

- useh = home

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

Zenish

Uh **useh**

Uh jeije

**Yiguo** **useh**

- What about the contexts?

English

A **home**

A garden

I arrived **home**

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

Zenish

Uh useh

Uh jeije

Yiguo useh

English

A home

A garden

I arrived home

- We can align them: Yiguo = I; Yiguo = arrived; Uh = A



# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

Zenish

Uh useh

Uh jeije

Yiguo useh

English

A home

A garden

I arrived home

- Reuse this knowledge to align more sentences: Uh = A

# Learning Translation Models - Word-based SMT

- Break sentences into smaller units: **words**
- Learn translation probabilities by **word aligning** a sentence-aligned corpus:

Zenish

Uh **useh**

Uh jeje

Yiguo **useh**

English

A **home**

A garden

I **arrived home**

- And the context again: jeje = garden

# Learning Translation Models - Word-based SMT

## Word-alignment:

- Identify **correspondences** between two languages at the **word level**
- Basis for word-based SMT, first step for other approaches
- Alignment learned via **Expectation Maximization** (EM)
  - Start with all alternative word alignments as equally likely
  - Observe across sentences that Zenish **useh** often links to English **home**
    - Increase probability of this word pair aligning
    - Knock-on effect: update alignment of other words
  - Iteratively redistribute probabilities, until they identify most likely links for each word (convergence)

# Learning Translation Models - Word-based SMT

- Word alignment commonly done using **IBM Models 1-5**

# Learning Translation Models - Word-based SMT

- Word alignment commonly done using **IBM Models 1-5**
- **IBM 1** is a straightforward application of EM, including the alignment to *null token* (deletion)
  - Finds **translation probabilities** for **words in isolation**, regardless of their position in parallel sentence

# Learning Translation Models - Word-based SMT

- Word alignment commonly done using **IBM Models 1-5**
- **IBM 1** is a straightforward application of EM, including the alignment to *null token* (deletion)
  - Finds **translation probabilities** for **words in isolation**, regardless of their position in parallel sentence
- **IBM 2-5** improve these distributions by considering:
  - **Position** of words in target sentence are related to position of words in source sentence (**distortion** model)
  - Some source words may be translated into **multiple** target words (**fertility** of the words)
  - Position of a target word may be related to **position of neighbouring words** (**relative distortion** model)

# Learning Translation Models - Word-based SMT

- IBM1 produces a probabilistic **dictionary** based on entire parallel corpus:

uh	a	0.90
uh	home	0.05
uh	garden	0.05
useh	a	0.03
useh	home	0.95
useh	I arrived	0.02
jejje	a	0.30
jejje	garden	0.70
yiguo	I arrived	0.80
yiguo	home	0.20

- Higher models estimate other probabilities: fertility, position, etc.

# Learning Translation Models - Word-based SMT

At translation (decoding) time

For a new sentence to translate, take the set of translations that jointly maximise the whole translation probability

$$E^* = \operatorname{argmax}_E P(F|E)$$

What about the fluency in the target language?



# Learning Language Models - Word-based SMT

**Language model:**  $P(E)$

$$E^* = \operatorname{argmax}_E P(F|E) \cdot P(E)$$

- Different translation options and different word orders are possible, some are more likely to happen in E

# Learning Language Models - Word-based SMT

**Language model:**  $P(E)$

$$E^* = \operatorname{argmax}_E P(F|E) \cdot P(E)$$

- Different translation options and different word orders are possible, some are more likely to happen in  $E$
- $P(E)$  = probability of strings  $E$  based on **relative frequencies** in a large corpus of language  $E$

# Learning Language Models - Word-based SMT

E.g.:

- Given the **new** sentence: “Yiguo la ta jeje”
- Assume new dictionary entries: la = at; ta = the
- Translation model could generate many possible translations, e.g.:

---

I arrived at the a  
I arrived at the garden  
home at the a  
home at the garden  
the a at I arrived

---

...

- Score each of them according to  $P(E)$

# Learning Language Models - Word-based SMT

E.g.:

- Given the **new** sentence: “Yiguo la ta jeje”
- Assume new dictionary entries: la = at; ta = the
- Translation model could generate many possible translations, e.g.:

---

I arrived at the a  
I arrived at the garden  
home at the a  
home at the garden  
the a at I arrived  
...

---

- Score each of them according to  $P(E)$

# Learning Language Models - Word-based SMT

E.g.:

- Given the **new** sentence: “Yiguo la ta jeje”
- Assume new dictionary entries: la = at; ta = the
- Translation model could generate many possible translations, e.g.:

---

I arrived at the a  
I arrived at the garden  
home at the a  
home at the garden  
the a at I arrived  
...

---

- Score each of them according to  $P(E)$

# Learning Language Models - Word-based SMT

- $P(E) = P(e_1, e_2, e_3, \dots, e_n)$   
 $= P(e_1)P(e_2|e_1) \cdot P(e_3|e_1, e_2) \cdots P(e_n|e_1, \dots, e_{n-1})$
- Difficult to have reliable estimates for whole sentences  $\rightarrow$  break it down into smaller sequences: **n-grams**
  - **Markov assumption:** *only the previous  $n-1$  words matter for predicting a word. For trigram models,  $n = 3$*

$$\simeq P(e_1) \cdot P(e_2|e_1) \cdot P(e_3|e_1, e_2) \cdot \\ P(e_4|e_2, e_3) \cdots P(e_n|e_{n-2}, e_{n-1})$$

# Learning Language Models - Word-based SMT

- Relative frequencies to compute these probabilities. E.g. trigrams:

$$P(e_3|e_1, e_2) = \frac{\text{count}(e_1 e_2 e_3)}{\text{count}(e_1 e_2)}$$

$$P(\textit{garden}|\textit{at}, \textit{the}) = \frac{\text{count}(\textit{at the garden})}{\text{count}(\textit{at the})}$$

# Learning Language Models - Word-based SMT

- Relative frequencies to compute these probabilities. E.g. trigrams:

$$P(e_3|e_1, e_2) = \frac{\text{count}(e_1 e_2 e_3)}{\text{count}(e_1 e_2)}$$

$$P(\text{garden}|\text{at}, \text{the}) = \frac{\text{count}(\text{at the garden})}{\text{count}(\text{at the})}$$

- For candidate: **I arrived at the garden**  
 $P(I|Start) \cdot P(arrived|Start, I) \cdot P(at|I, arrived) \cdot$   
 $P(the|arrived, at) \cdot P(garden|at, the) \cdot P(End|garden, the)$



# Learning Language Models - Word-based SMT

- Relative frequencies to compute these probabilities. E.g. trigrams:

$$P(e_3|e_1, e_2) = \frac{\text{count}(e_1 e_2 e_3)}{\text{count}(e_1 e_2)}$$

$$P(\textit{garden}|\textit{at}, \textit{the}) = \frac{\text{count}(\textit{at the garden})}{\text{count}(\textit{at the})}$$

- For candidate: **I arrived at the garden**  
 $P(I|Start) \cdot P(arrived|Start, I) \cdot P(at|I, arrived) \cdot$   
 $P(the|arrived, at) \cdot P(garden|at, the) \cdot P(End|garden, the)$
- Smoothing, back-off models, etc. to improve over relative counts

# Word-based SMT – Limitations

- Difficult to word-align, and hence learn a TM, for languages with **different words orders**
  - Considering *all* possible word orders – too costly, too noisy
  - Poor reordering model

# Word-based SMT – Limitations

- Difficult to word-align, and hence learn a TM, for languages with **different words orders**
  - Considering *all* possible word orders – too costly, too noisy
  - Poor reordering model
- **Fertility/n-m alignments**: Some languages may have **different notions of what counts as a word**

**Donaydampfshiffahrtsgesellschaftskapitaenskajuetenschluesseloch**

The keyhole of the door of the cabin of the captain of a steamship company operating on the Danube

# Phrase-based SMT

Most popular approach since early 2000s

No voy<sub>1</sub> a la<sub>2</sub> casa<sub>3</sub> → I am not going<sub>1</sub> to the<sub>2</sub> house<sub>3</sub>

# Phrase-based SMT

Most popular approach since early 2000s

No voy<sub>1</sub> a la<sub>2</sub> casa<sub>3</sub> → I am not going<sub>1</sub> to the<sub>2</sub> house<sub>3</sub>  
it seems to<sub>1</sub> me<sub>2</sub> → me<sub>1</sub> parece<sub>2</sub>

# Phrase-based SMT

Most popular approach since early 2000s

No voy<sub>1</sub> a la<sub>2</sub> casa<sub>3</sub> → I am not going<sub>1</sub> to the<sub>2</sub> house<sub>3</sub>

it seems to<sub>1</sub> me<sub>2</sub> → me<sub>1</sub> parece<sub>2</sub>

Je<sub>1</sub> ne vais pas<sub>2</sub> à la<sub>3</sub> maison<sub>4</sub> → I<sub>1</sub> am not going<sub>2</sub> to the<sub>3</sub> house<sub>4</sub>

# Phrase-based SMT

Most popular approach since early 2000s

No voy<sub>1</sub> a la<sub>2</sub> casa<sub>3</sub> → I am not going<sub>1</sub> to the<sub>2</sub> house<sub>3</sub>

it seems to<sub>1</sub> me<sub>2</sub> → me<sub>1</sub> parece<sub>2</sub>

Je<sub>1</sub> ne vais pas<sub>2</sub> à la<sub>3</sub> maison<sub>4</sub> → I<sub>1</sub> am not going<sub>2</sub> to the<sub>3</sub> house<sub>4</sub>

Eu<sub>1</sub> sinto saudade de você<sub>2</sub> → I<sub>1</sub> miss you<sub>2</sub>

# Phrase-based SMT

Most popular approach since early 2000s

No<sub>1</sub> voy<sub>1</sub> a<sub>2</sub> la<sub>2</sub> casa<sub>3</sub> → I<sub>1</sub> am not going<sub>1</sub> to<sub>2</sub> the<sub>2</sub> house<sub>3</sub>

it seems to<sub>1</sub> me<sub>2</sub> → me<sub>1</sub> parece<sub>2</sub>

Je<sub>1</sub> ne vais pas<sub>2</sub> à la<sub>3</sub> maison<sub>4</sub> → I<sub>1</sub> am not going<sub>2</sub> to<sub>2</sub> the<sub>3</sub> house<sub>4</sub>

Eu<sub>1</sub> sinto saudade de você<sub>2</sub> → I<sub>1</sub> miss you<sub>2</sub>

I<sub>1</sub> miss you<sub>2</sub> → Eu<sub>1</sub> sinto sua falta<sub>2</sub>



# Phrase-based SMT

Most popular approach since early 2000s

No<sub>1</sub> voy<sub>1</sub> a<sub>2</sub> la<sub>2</sub> casa<sub>3</sub> → I am not going<sub>1</sub> to the<sub>2</sub> house<sub>3</sub>

it seems to<sub>1</sub> me<sub>2</sub> → me<sub>1</sub> parece<sub>2</sub>

Je<sub>1</sub> ne vais pas<sub>2</sub> à la<sub>3</sub> maison<sub>4</sub> → I<sub>1</sub> am not going<sub>2</sub> to the<sub>3</sub> house<sub>4</sub>

Eu<sub>1</sub> sinto saudade de você<sub>2</sub> → I<sub>1</sub> miss you<sub>2</sub>

I<sub>1</sub> miss you<sub>2</sub> → Eu<sub>1</sub> sinto sua falta<sub>2</sub>

natuerlich<sub>1</sub> hat<sub>2</sub> John<sub>3</sub> spass am<sub>4</sub> spiel<sub>5</sub> → of course<sub>1</sub> John<sub>2</sub> has<sub>3</sub> fun  
with the<sub>4</sub> game<sub>5</sub>

- More intuitive and reliable alignments
  - Account for reordering within the phrases
  - Phrases can still be reordered





# Phrase-based SMT - Phrases from word alignments

Extract **phrase pairs** that are **consistent** with WA

- **Phrase**: sequence of tokens, not linguistically motivated
- WA produced by IBM Models, like before, only once

	reanudación	del	período	de	sesiones
resumption					
of					
the					
session					

# Phrase-based SMT - Phrases from word alignments

Extract **phrase pairs** that are **consistent** with WA

- **Phrase**: sequence of tokens, not linguistically motivated
- WA produced by IBM Models, like before, only once

	reanudación	del	período	de	sesiones
resumption					
of					
the					
session					

- 1 resumption  $\leftrightarrow$  reanudación
- 2 of the  $\leftrightarrow$  del
- 3 session  $\leftrightarrow$  período de sesiones

# Phrase-based SMT - Phrases from word alignments

Extract **phrase pairs** that are **consistent** with WA

- **Phrase**: sequence of tokens, not linguistically motivated
- WA produced by IBM Models, like before, only once

	reanudación	del	período	de	sesiones
resumption					
of					
the					
session					

# Phrase-based SMT - Phrases from word alignments

Extract **phrase pairs** that are **consistent** with WA

- **Phrase**: sequence of tokens, not linguistically motivated
- WA produced by IBM Models, like before, only once

	reanudación	del	período	de	sesiones
resumption					
of					
the					
session					

- 1 **resumption of the** ↔ **reanudación del**
- 2 **of the session** ↔ **del período de sesiones**
- 3 **resumption of the session** ↔ **reanudación del período de sesiones**

# Phrase-based SMT - Phrases from word alignments

Extract **phrase pairs** that are **consistent** with WA

- **Phrase**: sequence of tokens, not linguistically motivated
- WA produced by IBM Models, like before, only once

	reanudación	del	período	de	sesiones
resumption					
of					
the					
session					

- 1 **resumption of the** ↔ **reanudación del**
- 2 **of the session** ↔ **del período de sesiones**
- 3 **resumption of the session** ↔ **reanudación del período de sesiones**



# Phrase-based SMT - Phrase probabilities

- 1 Extract phrase pairs from word aligned parallel corpus ✓
- 2 Extract **counts** of those phrases from large parallel corpus (MLE):

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\text{count}(\bar{e})}$$

- 3 Store phrases and their probabilities in a **phrase table**
  - Probabilistic dictionary of **phrases**

# Phrase-based SMT - Weighing components

$$E^* = \operatorname{argmax}_E P(F|E) \cdot P(E)$$

- Rewriting equation for phrases:

ps. LM ( $P(E)$ ) remains the same: computed for n-grams

$$\mathbf{e}^* = \operatorname{argmax}_{\mathbf{e}} \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i) \cdot \prod_{i=1}^{|\mathbf{e}|} P(e_i | e_1 \cdots e_{i-1})$$

- Which component is **more important**?
  - $P(F|E)$  or  $P(E)$  ?

# Phrase-based SMT - Weighing components

$$E^* = \operatorname{argmax}_E P(F|E) \cdot P(E)$$

- Rewriting equation for phrases:  
ps. LM ( $P(E)$ ) remains the same: computed for n-grams

$$\mathbf{e}^* = \operatorname{argmax}_{\mathbf{e}} \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i) \cdot \prod_{i=1}^{|\mathbf{e}|} P(e_i | e_1 \cdots e_{i-1})$$

- Which component is **more important**?
  - $P(F|E)$  or  $P(E)$  ?
- Depends on size/quality of corpus, language-pair, etc.
- In generative model: components **equally important**

# Phrase-based SMT - Linear model

Weigh components for a given task (parallel corpus):

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \left[ \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i)^{\lambda_\phi} \cdot \prod_{i=1}^{|\mathbf{e}|} P(e_i | e_1 \cdots e_{i-1})^{\lambda_{LM}} \right]$$

Applying *log* (simpler to compute):

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \left[ \lambda_\phi \sum_{i=1}^I \log \phi(\bar{f}_i | \bar{e}_i) + \lambda_{LM} \sum_{i=1}^{|\mathbf{e}|} \log P(e_i | e_1 \cdots e_{i-1}) \right]$$

# Phrase-based SMT - Linear model

Model

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

# Phrase-based SMT - Linear model

## Model

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

## Components

1  $P_{LM}$

2  $\phi$

# Phrase-based SMT - Linear model

## Model

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

## Components

①  $P_{LM}$

②  $\phi$

## Weights

①  $\lambda_{LM}$

②  $\lambda_{\phi}$

# Phrase-based SMT - Linear model

## Model

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

### Components

①  $P_{LM}$

②  $\phi$

### Weights

①  $\lambda_{LM}$

②  $\lambda_{\phi}$

### Feature Functions

①  $h_1 = \log P_{LM}$

②  $h_2 = \log \phi$



# Phrase-based SMT - Linear model

## Model

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \exp \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

### Components

①  $P_{LM}$

②  $\phi$

### Weights

①  $\lambda_{LM}$

②  $\lambda_{\phi}$

### Feature Functions

①  $h_1 = \log P_{LM}$

②  $h_2 = \log \phi$

## Benefits

① Extensible

② Weights can be tuned, i.e., learned from examples

# Phrase-based SMT - Linear model

Common additional components  $h(\mathbf{f}, \mathbf{e})$ :

- **Direct** phrase translation probabilities:  $\phi(\bar{e}|\bar{f})$  extracted just like  $\phi(\bar{f}|\bar{e})$
- Distance-based phrase **reordering**:  
 $d(\text{start}_i - \text{end}_{i-1} - 1)$ , for every phrase  $\phi(\bar{f}_i|\bar{e}_i)$ 
  - Exponential decaying cost function  $d(x) = \alpha^{|x|}$
  - $x = \text{start}_i - \text{end}_{i-1} - 1$ : is there a **gap** in the translation between two source phrases?
- **Phrase penalty**: constant  $\rho$  for each phrase produced;  
 $\rho < 1$  to favour fewer, but longer phrases (more fluent)

# Phrase-based SMT - Linear model

Common additional components  $h(\mathbf{f}, \mathbf{e})$ :

- **Direct** phrase translation probabilities:  $\phi(\bar{e}|\bar{f})$  extracted just like  $\phi(\bar{f}|\bar{e})$
- Distance-based phrase **reordering**:  
 $d(\text{start}_i - \text{end}_{i-1} - 1)$ , for every phrase  $\phi(\bar{f}_i|\bar{e}_i)$ 
  - Exponential decaying cost function  $d(x) = \alpha^{|x|}$
  - $x = \text{start}_i - \text{end}_{i-1} - 1$ : is there a **gap** in the translation between two source phrases?
- **Phrase penalty**: constant  $\rho$  for each phrase produced;  $\rho < 1$  to favour fewer, but longer phrases (more fluent)
- Etc:  $\sim 15$  popular components/features

# Phrase-based SMT - Linear model

- Decoder remains similar, now with **weights** associated to components
- **Discriminative** model: learn  $\lambda$  weights such as to **minimise error** in small corpus

# Phrase-based SMT - Decoding

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{i=1}^n \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

- **Search problem:** finding the best scoring translation according to the model
  - Translation is build in sequence (left to right)
  - Input words may be covered **out of sequence** (allow for reordering)

# Phrase-based SMT - Decoding

- All phrases matching source words selected from phrase table. E.g.:

J'	ai	les	yeux	noirs	.
I	have	the	eyes	black	.
me	has	them	eye	dark	,
I have		eyes		espresso	!
I am		the eyes		somber	.
I did		some	black eyes		.
I had		black eyes			.
I have		black eyes			.
black eyes		I have			.

- Decoder selects **phrases** whose combination (in a given **order**) yields the **highest score** acc to the linear model

# Phrase-based SMT - Decoding

Incrementally construct translation **hypotheses** by trying out several possibilities:

- Generating target words in sequence, from left to right
- Computing the (so far) overall **log-linear model score** for each hypothesis
- Pruning search space via heuristics. e.g:
  - **Distortion limit** - at most 4 positions different from source order
  - Keep only partial hypothesis that are **promising**, e.g. model score is close to that of the best partial hypothesis so far

# Phrase-based SMT - Decoding

Incrementally construct translation **hypotheses** by trying out several possibilities:

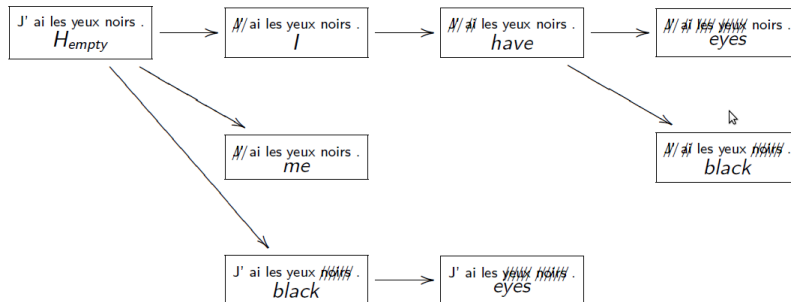
- Generating target words in sequence, from left to right
- Computing the (so far) overall **log-linear model score** for each hypothesis
- Pruning search space via heuristics. e.g:
  - **Distortion limit** - at most 4 positions different from source order
  - Keep only partial hypothesis that are **promising**, e.g. model score is close to that of the best partial hypothesis so far

Approximate search, e.g. stack-based **beam** search



# Phrase-based SMT - Decoding

## Search space



## Hypothesis

- Covered source words
- Target (output) words
- Model score

# Phrase-based SMT - Tuning Parameters

- Apply decoder with **uniform** weights to produce an **n-best** list of translations (e.g. top 1,000 translations)

**er** geht **ja** **nicht** **nach** **hause** → **he** **does** **not** **go** **home**

Rank	Translation	$\phi(\bar{e} \bar{f})$	$\phi(\bar{f} \bar{e})$	$lex(\bar{e} \bar{f}, a)$	$lex(\bar{f} \bar{e}, a)$	WP	PLM	Error
1	it is not under house	-9.93	-19.00	-5.08	-8.22	-5	-32.22	0.8
2	he is not under house	-7.40	-16.33	-5.01	-8.15	-5	-34.50	0.6
3	it is not a home	-12.74	-19.29	-5.08	-8.42	-5	-28.49	0.6
4	it is not to go home	-10.34	-20.87	-4.38	-13.11	-6	-32.53	0.8
5	it is not for house	-17.25	-20.43	-4.90	-6.90	-5	-31.75	0.8
6	he is not to go home	-10.95	-18.20	-4.85	-13.04	-6	-35.79	0.6
<b>7</b>	<b>he does not home</b>	<b>-11.84</b>	<b>-16.98</b>	<b>-3.67</b>	<b>-8.76</b>	<b>-4</b>	<b>-32.64</b>	<b>0.2</b>
8	it is not packing	-10.63	-17.65	-5.08	-9.89	-4	-32.26	0.8
9	he is not packing	-8.10	-14.98	-5.01	-9.82	-4	-34.55	0.6
10	he is not for home	-13.52	-17.09	-6.22	-7.82	-5	-36.70	0.4

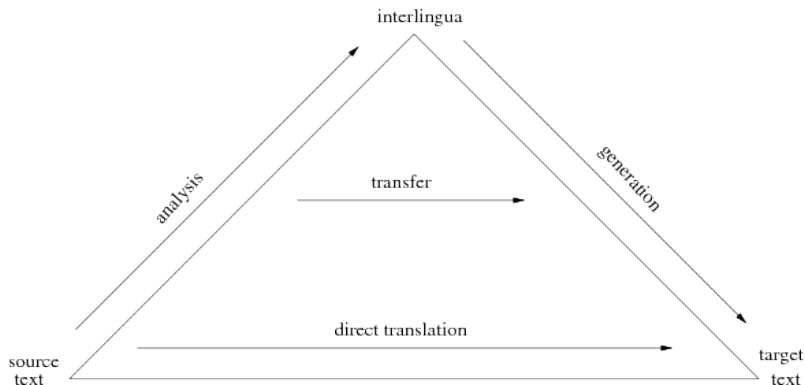
- Iteratively adapt weights to re-rank n-best translations, e.g. to make **7** appear at the top
- Error** acc. to evaluation metric (BLEU) against ref translation
- MERT, PRO, MIRA...

# Hierarchical and Syntax-based SMT

- PBSMT has trouble with **long-distance reorderings**
- Alternative approaches to bring structure and linguistic knowledge into the transfer rules of phrase table

# Hierarchical and Syntax-based SMT

- PBSMT has trouble with **long-distance reorderings**
- Alternative approaches to bring structure and linguistic knowledge into the transfer rules of phrase table



# Hierarchical SMT - Motivation

Introduce **structure** into phrase-based SMT models to deal with long-distance reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

# Hierarchical SMT - Motivation

Introduce **structure** into phrase-based SMT models to deal with long-distance reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

- How can we get a phrase for **shall be passing on?**

# Hierarchical SMT - Motivation

Introduce **structure** into phrase-based SMT models to deal with long-distance reordering

	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to			X				
you			X				
some				X			
comments						X	

- How can we get a phrase for **shall be passing on**?
- We cannot, unless we get **to you some comments** along

# Hierarchical SMT - Motivation

Introduce **structure** into phrase-based SMT models to deal with long-distance reordering

	Ich	werde					aushändigen
I							
shall							
be							
passing							
on							

- How can we get a phrase for **shall be passing on**?
- We cannot, unless we get **to you some comments** along
- Unless we replace all those words by a variable



# Hierarchical SMT - Motivation

shall be passing on to you some comments



werde Ihnen die entsprechenden Anmerkungen  
aushändigen

shall be passing on to you some comments



werde Ihnen die entsprechenden Anmerkungen  
aushändigen

shall be passing on X



werde X aushändigen

# Hierarchical SMT - basics

Learnt from word-aligned parallel corpora in the same way as before

# Hierarchical SMT - basics

Learnt from word-aligned parallel corpora in the same way as before

- Based on the fact that language has **recursive** structures
- Phrases within other phrases treated as nonterminals:  
replaced by **X**
- No linguistic constraints added - yet, some **structure**

# Hierarchical SMT - basics

shall be passing on to you some comments



werde Ihnen die entsprechenden Anmerkungen  
aushändigen

---

shall be passing on  $X_1$  some comments



werde  $X_1$  die entsprechenden Anmerkungen  
aushändigen

---

shall be passing on  $X_1$   $X_2$



werde  $X_1$   $X_2$  aushändigen

# Hierarchical SMT - phrase-table

$[X] \rightarrow$  shall be passing on  $X_1$   $X_2$  | werde  $X_1$   $X_2$  aushändigen

$[X] \rightarrow$  shall be passing on  $X_3$  | werde  $X_3$  aushändigen

$[X] \rightarrow$  to you | Ihnen

$[X] \rightarrow$  some comments | die entsprechenden Anmerkungen

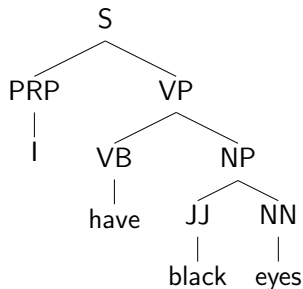
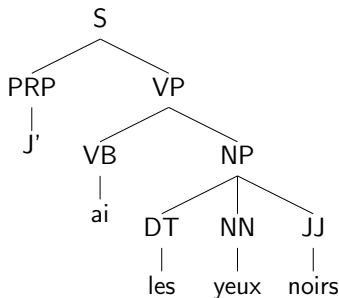
$[X] \rightarrow$  to you some comments | Ihnen die entsprechenden Anmerkungen

Learn a bilingual (**synchronous**) set of context-free rules on how to translate  $F \rightarrow E$

# Syntax-based SMT

- Hierarchical models are not very informative (**Xs**) and suffer from an exponential number of rules
- **Syntax-based** SMT: uses linguistic categorise to label nodes
- Syntactic parser at pre-processing time for at least one language (the other could have Xs)
- Learn a bilingual (**synchronous**) set of **linguistically** informed context-free rules on how to translate  $F \rightarrow E$
- Rules extracted acc to word-alignment, **constrained by heuristics for syntax**

# Syntax-based SMT



Standard **constraints** on rule construction:

- Single nonterminal on the left
- Consistent with word-alignment
- Nonterminals on the right must align one-to-one

	J'	ai	les	yeux	noirs
I					
have					
black					
eyes					

# Syntax-based SMT

## Grammar

$\text{PRP} \rightarrow \text{J}' \mid \text{I}$

$\text{JJ} \rightarrow \text{noirs} \mid \text{black}$

$\text{NP} \rightarrow \text{les yeux JJ} \mid \text{JJ eyes}$

$\text{VP} \rightarrow \text{ai NP} \mid \text{have NP}$

$\text{S} \rightarrow \text{PRP VP} \mid \text{PRP VP}$

**Decoding** becomes a (probabilistic) parsing problem!



# Outline

- 1 Introduction
- 2 SMT
- 3 Evaluation**
- 4 Success stories
- 5 Conclusions

# MT evaluation metrics

For developers/researchers:

- Measure progress over time
- Compare MT systems
- Tune model parameters

Quality =

Close to human translation

- **N-gram matching** between system output and one or more **reference** (human) translations

# MT evaluation metrics

## BLEU: BiLingual Evaluation Understudy

- Most widely used metric
- Matching of n-grams between MT and Ref: rewards **same words** in **equal order**

# MT evaluation metrics

## BLEU: BiLingual Evaluation Understudy

- Most widely used metric
- Matching of n-grams between MT and Ref: rewards **same words** in **equal order**

**Ref:** the Iraqi **weapons** are to be handed over **to the army** within **two weeks**

**MT:** in **two weeks** Iraq's **weapons** will give **to the army**

- 1-gram precision:  $6/10$
- 2-gram precision:  $3/9$
- 3-gram precision:  $3/8$
- $\text{BLEU} = \left( \prod_{n=1}^3 p_n \right)^{\frac{1}{3}}$
- $\text{BLEU} = \left( \frac{6}{10} * \frac{3}{9} * \frac{2}{8} \right)^{\frac{1}{3}} = 0.368$

# Outline

- 1 Introduction
- 2 SMT
- 3 Evaluation
- 4 Success stories**
- 5 Conclusions

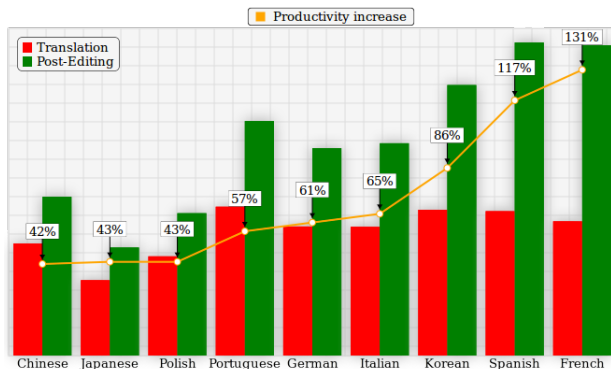
# Interest in MT beyond academia

- **Government:** U.S. (**intelligence** purposes); EU (societal/political/commercial purposes):
  - EU spends more than 300M€ on human translation / year: 23 official languages (253 language pairs) in 2011
- **End-users:** Google Translate, Bing Translator, etc.
- **Language industry:** considerable savings with MT
  - Customised SMT: KantanMT, AsiaOnline, etc.

# Commercial interest in MT

**Autodesk:** productivity test (**post-editing** of MT) [Aut11]

- 2-day translation vs post-editing, 37 participants
- In-house Moses (Autodesk data: software)
- **Time** spent on each segment



# Commercial interest in MT

**Intel** - User satisfaction, unedited MT

- Translation is good if customer can **solve problem**



# Commercial interest in MT

**Intel** - User satisfaction, unedited MT

- Translation is good if customer can **solve problem**
- MT for Customer Support websites [Int10]
  - Overall customer satisfaction: **75%** for English→Chinese

# Commercial interest in MT

## Intel - User satisfaction, unedited MT

- Translation is good if customer can **solve problem**
- MT for Customer Support websites [Int10]
  - Overall customer satisfaction: **75%** for English→Chinese
  - **95%** reduction in cost
  - Project cycle from **10 days** to **1 day**
  - Customers in China using MT texts were more satisfied with support than natives using original texts (**68%**)!

# Other uses of MT+PE

**WIPO**: MT for patent translation

- **PATENTSCOPE**: customised Moses
- <https://www3.wipo.int/patentscope/translate/translate.jsf>

# Other uses of MT+PE

**WIPO**: MT for patent translation

- **PATENTSCOPE**: customised Moses
- <https://www3.wipo.int/patentscope/translate/translate.jsf>

**United Nations** uses same workflow as WIPO

# Other uses of MT+PE

## WIPO: MT for patent translation

- **PATENTSCOPE**: customised Moses
- <https://www3.wipo.int/patentscope/translate/translate.jsf>

## United Nations uses same workflow as WIPO European Commission

- Customised Moses for all EU languages [EC-13]
- Technology free of charge to any PA in an EU country, or in an EU institution or agency
- [http://ec.europa.eu/dgs/translation/translationresources/machine\\_translation/index\\_en.htm](http://ec.europa.eu/dgs/translation/translationresources/machine_translation/index_en.htm)

# MT for communication

## Skype Translator

- Microsoft's speech to speech translation
- Pipeline:
  - Data cleaning
  - Pre-processing (named entity recognition)
  - Speech recognition
  - SMT
  - Speech generation
- Strong interaction component
- <https://www.microsoft.com/translator/skype.aspx>

# Outline

- 1 Introduction
- 2 SMT
- 3 Evaluation
- 4 Success stories
- 5 Conclusions**

# Conclusions

- (Cheap) translation is in high **demand**, which cannot be supplied by human translators
- MT **quality is good** for some languages, types of texts, applications



# Conclusions

- (Cheap) translation is in high **demand**, which cannot be supplied by human translators
- MT **quality is good** for some languages, types of texts, applications
- Popular approaches:
  - Most language pairs: **phrase-based** SMT is sufficient
  - Language pairs with long-distance reordering: **syntax-based** SMT does better

# Conclusions

- (Cheap) translation is in high **demand**, which cannot be supplied by human translators
- MT **quality is good** for some languages, types of texts, applications
- Popular approaches:
  - Most language pairs: **phrase-based** SMT is sufficient
  - Language pairs with long-distance reordering: **syntax-based** SMT does better
- Possible improvements from:
  - **More information**: linguistic (local and global), contextual
  - **Better methods**: fully discriminative, deciphering, DNNs, exact search

# Conclusions

- (Cheap) translation is in high **demand**, which cannot be supplied by human translators
- MT **quality is good** for some languages, types of texts, applications
- Popular approaches:
  - Most language pairs: **phrase-based** SMT is sufficient
  - Language pairs with long-distance reordering: **syntax-based** SMT does better
- Possible improvements from:
  - **More information**: linguistic (local and global), contextual
  - **Better methods**: fully discriminative, deciphering, DNNs, exact search
- SOA?

# Conclusions

- (Cheap) translation is in high **demand**, which cannot be supplied by human translators
- MT **quality is good** for some languages, types of texts, applications
- Popular approaches:
  - Most language pairs: **phrase-based** SMT is sufficient
  - Language pairs with long-distance reordering: **syntax-based** SMT does better
- Possible improvements from:
  - **More information**: linguistic (local and global), contextual
  - **Better methods**: fully discriminative, deciphering, DNNs, exact search
- SOA? **DNNs** are a promising direction to better model context, long-distance dependencies

# Statistical Machine Translation

Lucia Specia

`l.specia@sheffield.ac.uk`

LxMLS 2015

18 July 2015



The  
University  
Of  
Sheffield.

# References I



## Autodesk.

Translation and Post-Editing Productivity.

In <http://translate.autodesk.com/productivity.html>, 2011.



## Machine translation.

In Anabela Pereira, editor, *Languages and Translation*, [http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue\\_06\\_en.pdf](http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_06_en.pdf). Directorate-General for Translation, 2013.



## Intel.

Being Streetwise with Machine Translation in an Enterprise Neighborhood.

In [http://mtmarathon2010.info/JEC2010\\_Burgett\\_slides.pptx](http://mtmarathon2010.info/JEC2010_Burgett_slides.pptx), 2010.