АБУМАТРАН
ABUMATRAN

Workshop on statistical machine translation for curious translators

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering, S.L.

DCU prompsit Universitat d'Alacant Universidad de Alicante IEA ILSP

Outline

1) Introduction to machine translation

2) The Abu-MaTran project

3)Acquisition of parallel data from the web

- How a web crawler works
- Hands-on session: Bicrawler

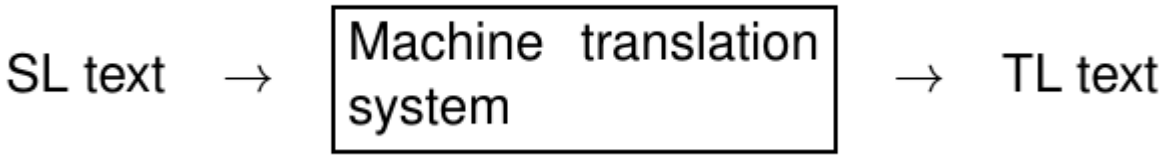
4) Statistical machine translation (SMT)

- Introduction to SMT
- Hands-on session: MTradumàtica

Introduction to machine translation

Machine translation

- Translation, by means of a computing system (computer+software) of texts in digital form from one natural language (source language; SL) to another (target language; TL)



- No human intervention whatsoever

Applications of machine translation

- Machine translation and professional translation, even if closely related in purpose, are not interchangeable products (Sager,1994)
- A machine translation, is really a translation?
  - It cannot be used as a professional product would
  - This does not mean machine translation is useless!

Applications of machine translation

- **Gisting (assimilation)**: ephemeral translation, ideally instantaneous, used to get a rough idea of a text when you do not speak the language or you speak it badly
  - Internet surfing, informal communication, etc.

Irish→English (Google)

Is ar an Oileán Fada a bhí mé féin agus m'fhear céile ag fanacht nuair a rinne muid an turas sin. Dúradh linn fanacht cois farraige ag am díthrá agus thuirling an muireitleán anuas chun sinn a thabhairt ar bord.	On the Long Island I was alone and my husband waiting when we made the trip. We told seaside stay at low tide and landed the Flying down to us on board.

Applications of machine translation

- **Post-editing (dissemination)**: permanent translation, ideally with few errors, for its publication after correction
  - Production of drafts for post-editing

French→English (Google)

Au cours de ses 43 années d'aventure européenne, Londres a souvent été perçu comme réticent à tout nouvel approfondissement de l'Union européenne et à une intégration plus avancée. Volontairement en dehors de la zone euro et de l'espace Schenghen, le pays a régulièrement critiqué les	During its 43 years of European Adventure, London has often been seen as reluctant to any further deepening of the European Union and further integration. Voluntarily outside the euro area and Schenghen space, the country has regularly criticized the European institutions and undermined its contribution to the EU bud-

Applications of machine translation

	Necessary	Unnecessary
Gisting	Understandability Fast translation	Syntactic *correctness* Lexical *correctness* Predictable errors ☺ Happy translators
Post-editing	Accurate syntax Predictable errors High accuracy (WER $\leq$ 20%) ☺ Happy translators	Understandability Fast translation

Applications of machine translation

- Gisting:
  - English (MT): *Match very difficult but fans unconditional support players very motivated
  - English (Cor.): ~~Match~~The game was very difficult but ~~fans~~ the unconditional support of fans made the players to be very motivated
  - Spanish (SL): El partido ha sido muy difícil pero el apoyo incondicional de la afición hizo que los jugadores estuvieran muy motivados

Applications of machine translation

- Post-editing (dissemination):
  - English (MT): *I eat you were not coming we left
  - English (Cor.): ~~I eat~~As you were not coming we left
  - Spanish (SL): Como no venías, nos fuimos

Rule-based machine translation

- Uses explicit representations of linguistic information: dictionaries, rules, etc.

EN

```
house   -> house-N.sg
houses  -> house-N.pl
blue    -> blue-ADJ
...        ...
```

ES

```
casa    -> casa-N.f.sg
casas   -> casa-N.f.pl
azul    -> azul-ADJ.mf.sg
...        ...
```

EN-ES

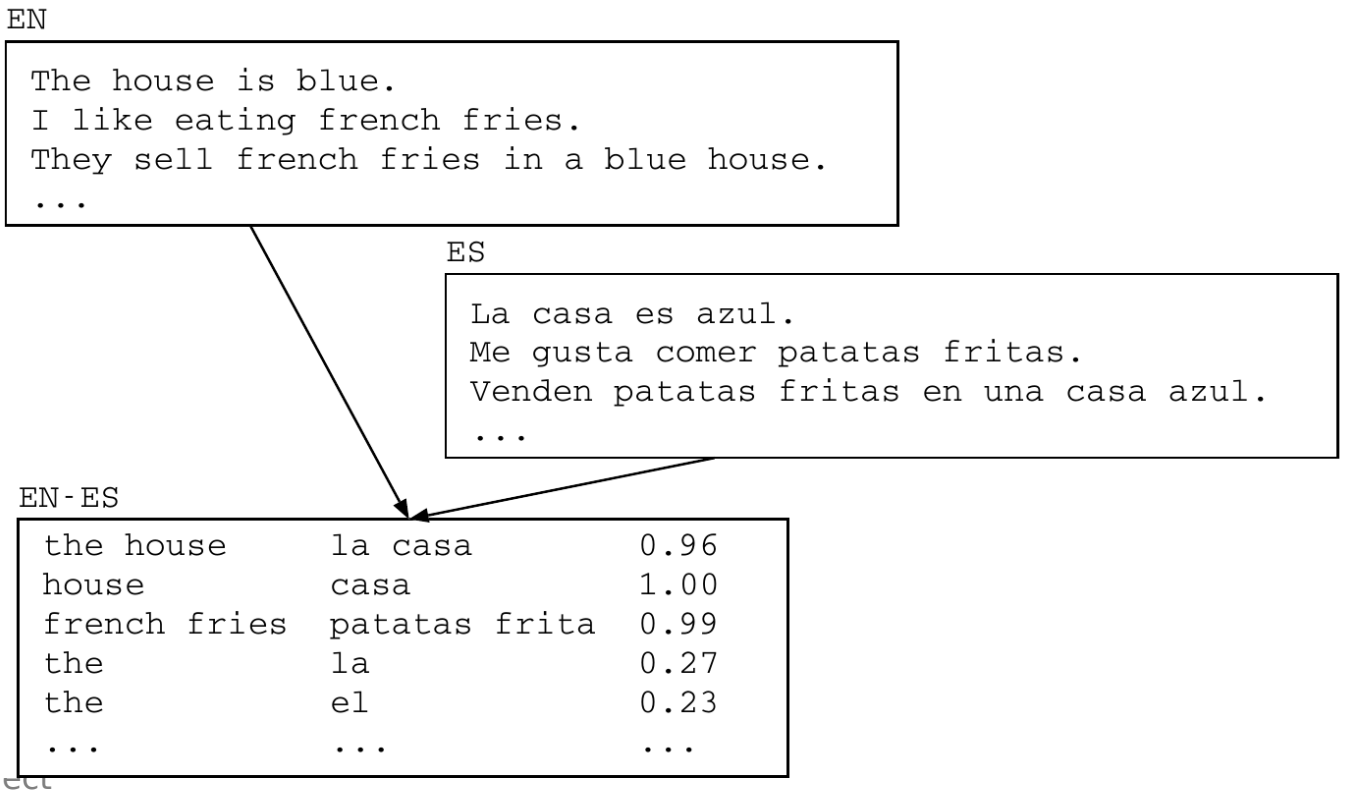
```
house-N   -> casa-N
blue-ADJ  -> azul-ADJ
...          ...
```

EN-ES

```
DT NP1 GS NP2  -> "el"-DT NP2 "de"-PR DT NP1
DT ADJ N       -> DT N ADJ
...               ...
```

Corpus-based machine translation

- Learns to translate from large amounts of existing translations (bitexts = parallel corpora)
- **Statistical machine translation** (SMT) is corpus-based



Approaches to machine translation

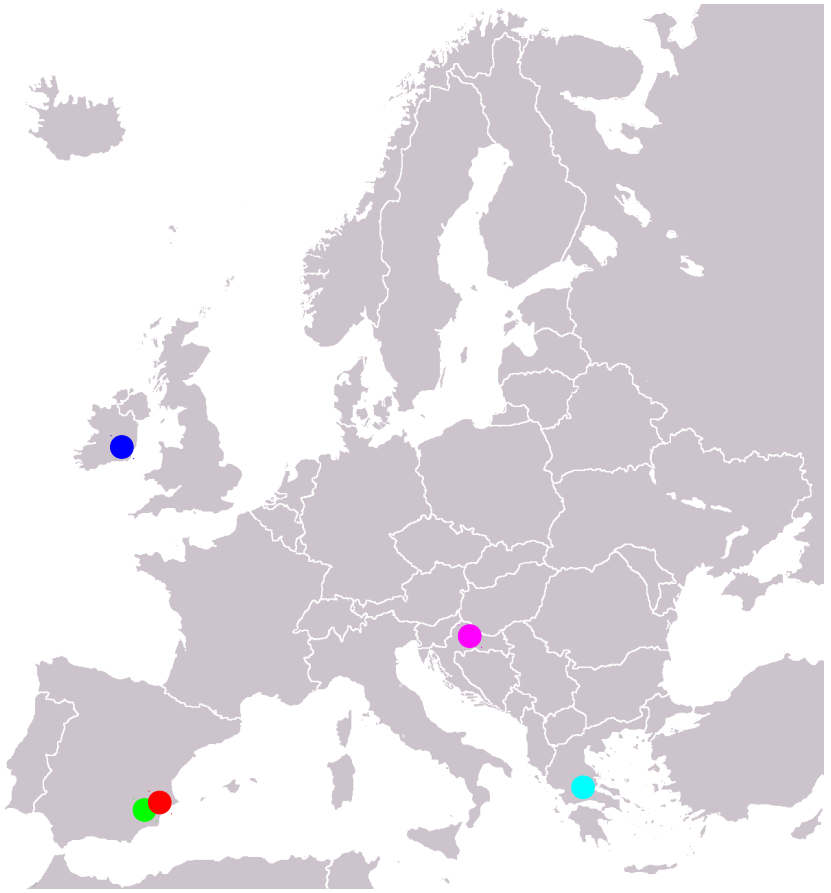
- Corpus-based MT works best when . . .
  - You have a big bitext of pre-translated and aligned sentences
  - The languages involved are not morphologically complex
  - The texts to be translated are in the same domain as those used to learn
- Rule-based MT works best when . . .
  - You do not have bitexts, or they are of low quality
  - The languages involved are typologically similar (e.g. es–ca, es–pt, es–fr)
  - You are translating formal language

The Abu-MaTran project

Abu-MaTran in a nutshell

- Marie Curie IAPP (Industry-Academia Partnerships and Pathways)
  - core activity: transfer of knowledge
  - by means of secondments: put in contact academic and industrial partners
- Duration: 48 months (from January 2013): it is about to end

Partners



- Dublin City University (Ireland)
- **Prompsit Language Engineering (Spain)**
- **University of Alicante (Spain)**
- University of Zagreb (Croatia)
- Institute for Language and Speech Processing (Greece)

Abu-MaTran in a nutshell

- Enhance industry-academia cooperation to tackle multilinguality
- Increase low industrial adoption of machine translation
- Transfer back to academia the know-how of industry to make research products more robust
- Resources produced to be released as free/open-source software
- Focus on Croatian: language of new EU member state
- Emphasis on dissemination

Some results (I)

- Multiple open-source tools released:
  - Web crawlers, rule inference toolkits for rule-based machine translation, etc.
- Corpora released:
  - General-domain monolingual corpora for Croatian, Serbian, Bosnian, Catalan and Finnish
  - General-domain parallel corpora for English-to Croatian, Serbian, Bosnian and Finnish
  - Tourism domain parallel corpora for English-Croatian
  - …
- Machine translation systems created:
  - Rule-based: Serbian-Croatian
  - Statistical:English-Croatian (general domain and tourism domain), English-Greek (tourism domain)

Some results (II)

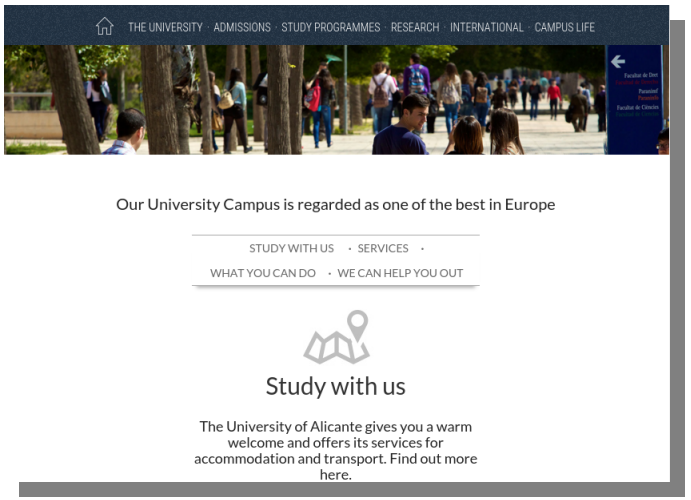
- Organization of Spanish Linguistics Olympiad 2014-2015-2016
- Workshop organization:
  - 2014, Dublin: Software management for researchers
  - 2014-2015, Zagreb: data creation for Croatian RBMT
  - 2014, Reykjavik: free/open-source RBMT linguistic resources
  - 2016, Dublin: Hybrid machine translation
  - 2016, Dublin: Tools for linguists
  - 2016, UA: Statistical machine translation

Acquisition of parallel data from the web

1)Web crawling
2)Hands-on session: Bicrawler

Web crawling

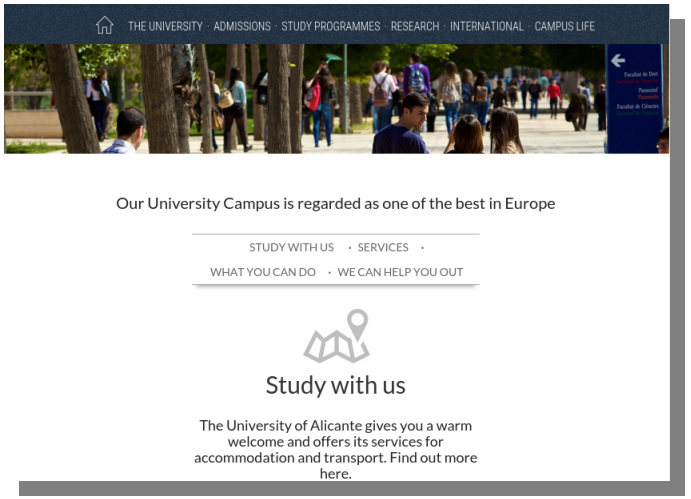
- We can find many multilingual websites on the Internet



- Parallel corpora are essential to build SMT systems
- We can **automatically** obtain a parallel corpus from a multilingual website with a **web crawler**

How a web crawler works

- How can we turn a multilingual website ...



- … into a parallel corpus ready for SMT?

Our University Campus is regarded as one the best in Europe	La Universidad puede presumir de tener uno de los mejores campus europeos
Study with us	¿Vienes?

How a web crawler works

1) Download web pages (documents)
2) Extract text and remove HTML tags
3) Detect language of documents
4) Identify documents that are mutual translation **(most difficult part)**
5) Extract parallel sentences from each document pair

How a web crawler works

1)Download web pages (documents)

- The most time-consuming part: downloading a big website can take days and even **weeks**!
- From the main page (e.g. www.ua.es), hyperlinks are followed in order to get new documents
- From new documents, hyperlinks are followed in order to get more documents, and so on…

How a web crawler works

2)Extract text and remove HTML tags

- HTML tags need to be stored: they are needed in subsequent steps
- Text is split into **paragraphs**

```
<div class="row">
<div class="col-md-12">
<h2 class="subSeccionIcono"
id="vienes"><img
src="https://web.ua.es/secciones-
ua/images/acceso/estudia/vida-
universitaria/icono1.jpg" /> Study with
us</h2>
<h3 class="subtituloIcono">The University
of Alicante gives you a warm welcome and
offers its services for accommodation and
transport. Find out more here.</h3>
```

Study with us

The University of Alicante gives you a warm welcome and offers its services for accommodation and transport. Find out more here.

How a web crawler works

3)Detect language of documents

Study with us

The University of Alicante gives you a warm welcome and offers its services for accommodation and transport. Find out more here.

English

¿Vienes?

La Universidad de Alicante te acoge con toda clase de facilidades para el alojamiento o el transporte. Conócelas aquí.

Spanish

How a web crawler works

4) Identify documents that are mutual translation

- The most difficult part
- Clues that help us to identify pairs of documents:
  - URL: e.g. https://web.ua.es/en/university-life.html and https://web.ua.es/es/university-life.html
  - Images
  - Numbers
  - Named entities
  - HTML structure/layout
  - Links
  - Similarity after being translated with some bilingual resource: finding parallel resources is difficult for some language pairs!

How a web crawler works

5) Extract parallel sentences from each document pair

- Split sentences from each paragraph

Study with us	¿Vienes?
The University of Alicante gives you a warm welcome and offers its services for accommodation and transport.	La Universidad de Alicante te acoge con toda clase de facilidades para el alojamiento o el transporte.
Find out more here.	Conócelas aquí.

Linguistic resources for web crawling

- Bilingual dictionaries are an essential resource for *Bitextor*, one of the web crawling tools developed in Abu-MaTran
  - They are used for identifying documents that are mutual translation
  - Can be automatically obtained from parallel corpora
  - If we are crawling data for a resource-poor language pair, we may need to create them by hand

Bicrawler

- Web-based service for extracting parallel corpora from multilingual websites
- Makes acquisition of parallel data available to everyone
- Developed by Prompsit Language Engineering
- Built upon the web crawlers released by Abu-MaTran
- Added an additional cleaning layer to remove possible errors introduced by the crawling tools
- Free use, but limited in terms of crawling time
- Unlimited (premium) version will be available soon

Hands-on session

Download instructions from http://abumatran.eu/ua-dec-2016-guide.pdf

Statistical machine translation (SMT)

1)Introduction to SMT
2)Hands-on session: MTradumàtica

Statistical machine translation

- Statistical machine translation is a corpus-based machine translation approach
- It is the most popular one in translation industry
- It allows us to automatically build an MT system from existing translations (bitexts)
  - The texts must be **segmented** into sentences
  - Sentences must be **aligned**, i.e. sentences which are translation of each other must be identified

Phrase-based statistical machine translation

- **Translation**: TL sentence with highest probability according to a combination of statistical models
- Translation hypotheses are built by splitting the SL sentence in segments and concatenating (not necessarily in the same order) their translations according to a **phrase table**
- Example: *the small houses*

the	*el*	0.5
the	*las*	0.2
the small	*el*	0.05
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
houses	*hogar*	0.3

el casas pequeñas	0.35
el hogar	0.015
las casas pequeñas	0.14
el medianas hogares	0.015

Why do we need more models?

- *el* and *casas pequeñas* are correct translations **in some particular contexts**
- We need a tool that tells us whether the chosen phrase translations **match** and produce a fluent sentence in the TL

Example: *the small houses*

the	*el*	0.5
the	*las*	0.2
the small	*el*	0.05
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
houses	*hogar*	0.3

T

el casas pequeñas	0.35
el hogar	0.015
las casas pequeñas	0.14
el medianas hogares	0.015

SMT models

- Phrase translation model in both directions
- **Language model** of the target language (TL)
- Word penalty
- Phrase penalty
- Reordering model
- ...

Phrase translation model

- Phrase table
  - Multi-word probabilistic bilingual dictionary (in both directions) with variable-length segments

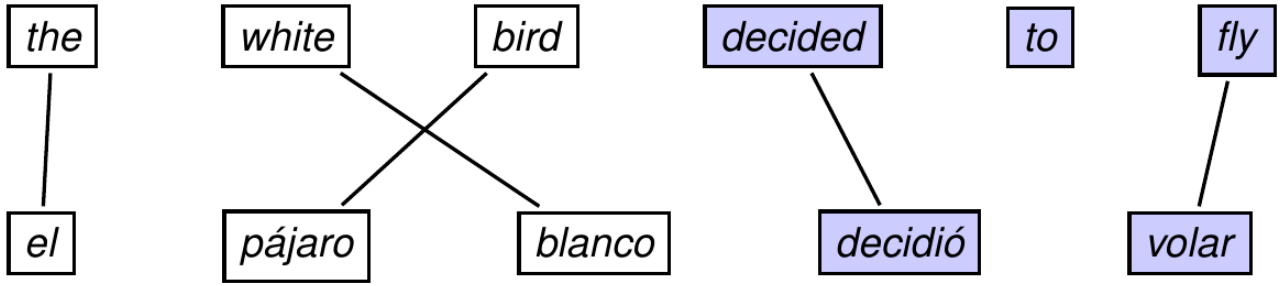
Source (s)	Target (t)	$p(s\|t)$	$p(t\|s)$
...	...	...	...
here are the dates for	voilà les dates de	1.00	1.00
here are the dates	voilà les dates	1.00	1.00
here are the	voici donc les	0.33	0.50
here are the	voilà les	0.04	0.50
...	...	...	...

Phrase translation model

Obtained from a **parallel corpus**

1) Compute word alignments

2) Extract bilingual phrases from the word alignments



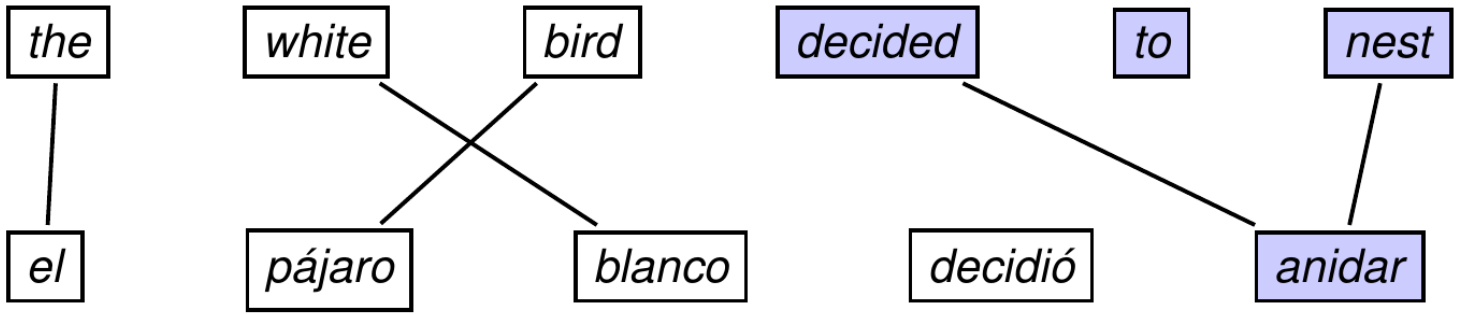
3) Compute translation probabilities

$$p(s|t) = \frac{\text{count}(s \leftrightarrow t)}{\text{count}(t)}; \quad p(t|s) = \frac{\text{count}(s \leftrightarrow t)}{\text{count}(s)}$$

Phrase translation model

Is corpus size important?

- Words not found in the SL side of the phrase table are not translated; just copied to the output
- Infrequent words in the corpus are likely to be wrongly aligned:



- **The bigger, the better!**

Target language model

- It allows us to measure how likely (fluent) a TL sentence is, how “good” it is that sentence in the TL
- Like when you use Google to solve translation doubts:
  - *el casas pequeñas*: (21.000) vs *las casas pequeñas*: (276.000) results
- Instead of Google, we use **large** TL monolingual texts
- Since we may not found the full hypotheses in the text, we use an statistical model based on segments of n words (n-grams):

$$
\begin{aligned}
&p(\text{The potential of machine translation is clear}) = \\
&\quad p(\text{The}) \times p(\text{potential}|\text{The}) \times p(\text{of}|\text{The potential}) \times \\
&\quad p(\text{machine}|\text{potential of}) \times p(\text{translation}|\text{of machine}) \times \\
&\quad p(\text{is}|\text{machine translation}) \times p(\text{clear}|\text{translation is})
\end{aligned}
$$

Target language model

- Probabilities obtained as:

$$p(\text{house}|\text{the red}) = \frac{\text{count}(\text{the red house})}{\text{count}(\text{the red } *)}$$

- Why **large TL monolingual texts**?

$$p(\text{las casas pequeñas}) =$$
$$p(\text{las}) \times p(\text{casas}|\text{las}) \times p(\text{pequeñas}|\text{las casas})$$

$$p(\text{casas}|\text{las}) = \frac{\text{count}(\text{las casas})}{\text{count}(\text{las } *)}$$

  - What happens if *casas* is not in the monolingual corpus?

Target language model

If the language model help us to combine the translation of each SL segment, why do we need multi-word segments?

Example: *estación de esquí → *ski season*

Source (s)	Target (t)	p(t|s)
estación	season	0.4
estación	station	0.4
estación	resort	0.2
de esquí	ski	1.0

Phrase table: *ski season (0.4), ski station (0.4), ski resort (0.2)*

Language model: *ski season (0.5), ski station (0.1), ski resort (0.5)*

Multi-word segments allow us to take into account context in the SL

Other models

- Word penalty: number of words in the target translation
  - The language model likes short sentences (less n-grams to score)
  - Used to avoid producing very short translations
- Phrase penalty: number of bilingual phrases used to produce the target
  - Used to promote the use of long phrases (fewer phrases)
- Reordering model: how likely is to change the order of a phrase when assembling the translation hypothesis.

Parameter tuning

- Not all models are equally important
- Probability of a translation hypothesis:

$$p(\text{target}|\text{source}) \propto \lambda_1 h_1(\cdot) + \lambda_2 h_2(\cdot) + \cdots + \lambda_{14} h_{14}(\cdot)$$

- $h_i(.)$: prob of hypothesis according to model; λ_i : weight of model h_i
- Tuning: starting with random values for the weights λ_i, find the set of values that maximises translation quality
  - From a (small) development parallel corpus
  - Its SL side is translated, compared to the TL side and weights are updated to obtain a more accurate translation
  - The process is repeated iteratively

Parameter tuning

- Why do we need to give weights to models?

Source (s)	Target (t)	p(t|s)
We managed to	conseguimos	1.0
stem	raíz	0.5
stem	tallo	0.4
stem	detener	0.1
the bleeding	la hemorragia	1.0

Source: *we managed to stem the bleeding*

Hyp 1: *conseguimos raíz la hemorragia* PT=0.5; LM=0.1; sum=0.6

Hyp 2: *conseguimos tallo la hemorragia* PT=0.4; LM=0.25; sum=0.75

Hyp 3: *conseguimos detener la hemorragia* PT=0.1; LM=0.4; sum=0.5

Mtradumàtica (I)

- Web interface for Moses
- Developed by Prompsit Language Engineering for Universitat Autònoma de Barcelona
- It will be released by Universitat Autònoma de Barcelona soon
- Allows you to easily experiment with SMT:
  - Manage files and corpora
  - Train LMs and SMT systems
  - Tune systems
  - Translate text
  - Inspect phrase table and language model

Mtradumàtica (II)

- Currently you cannot:
  - Apply domain adaptation methods
  - Evaluate systems with automatic metrics

- Useful tool for understanding how SMT works

Hands-on session

Download instructions from http://abumatran.eu/ua-dec-2016-guide.pdf

АБУМАТРАН
ABUMATRAN

Thank you for your attention

The Abu-MaTran project

* Part of the presentation was created by **Felipe Sánchez Martínez**


Universitat d'Alacant
Universidad de Alicante