Nanyang Technological University, Singapore
August 31, 2007

Statistical Machine Translation: State of the Art and Future Challenges

Rafael E. Banchs

Universitat Politècnica de Catalunya – Barcelona Media Centre d'Innovació

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

Evolution of the machine translation problem

- IV century: Saint Jeronimus translates the Bible into Latin.
- XVII century: First attempts to develop universal languages.

- XX century: Digital computers come into scene
    - 40s: Translation as a cryptographic problem.
    - 70s: A work for Artificial Intelligence.
    - 90s: Statistical machine translation becomes feasible.
- XXI century: Practical applications of machine translation systems ???

Machine translation as a research topic

Search on *www.google.es*	*Results*
“ machine translation ”	***559.000***
“ machine translation ” + research	***196.000***
“ machine translation ” + research + university	***131.000***
“ machine translation ” + research – university	***63.800***
“ machine translation ” + conference	***123.000***
“ machine translation ” + journal	***98.100***

Machine translation around the world

- GALE: Global Autonomous Language Environments
  http://ciir.cs.umass.edu/research/nightingale.html
- TC-STAR: Technology and Corpora for Speech to Speech Translation
  http://www.tc-star.org
- MANOS: Multilingual Application Network for Olympic Services
  http://nlpr-web.ia.ac.cn/english/cip/english/project.htm
- NIST: National Institute of Standards and Technology
  http://www.nist.gov/speech/tests/mt/
- IWSLT: International Workshop on Spoken Language Translation
  http://www.is.cs.cmu.edu/iwslt2005/index.html

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

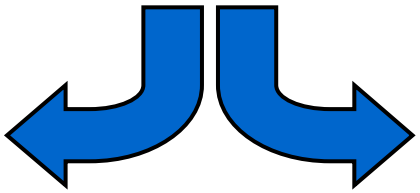
5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

Two paradigms

Machine Translation Methods



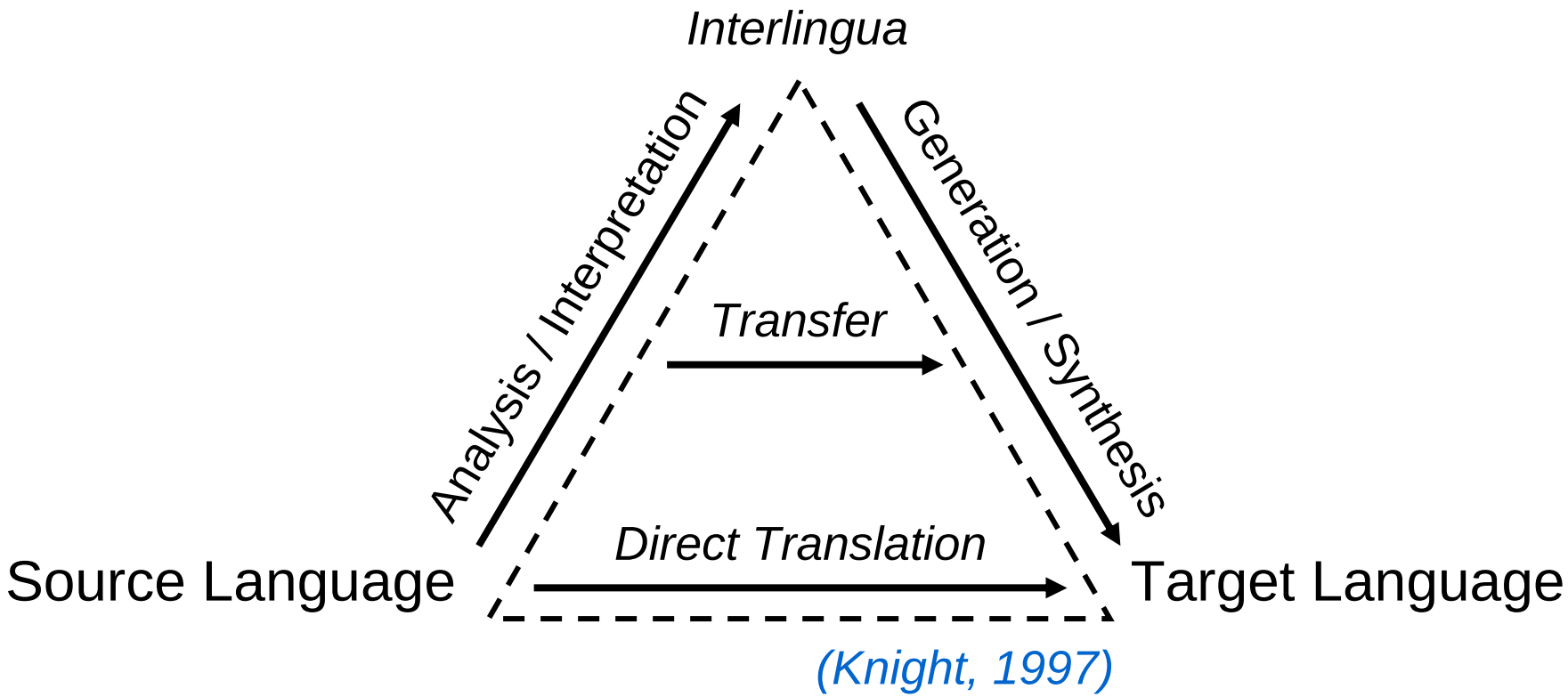
Knowledge-based methods

Based on using rules and linguistic knowledge

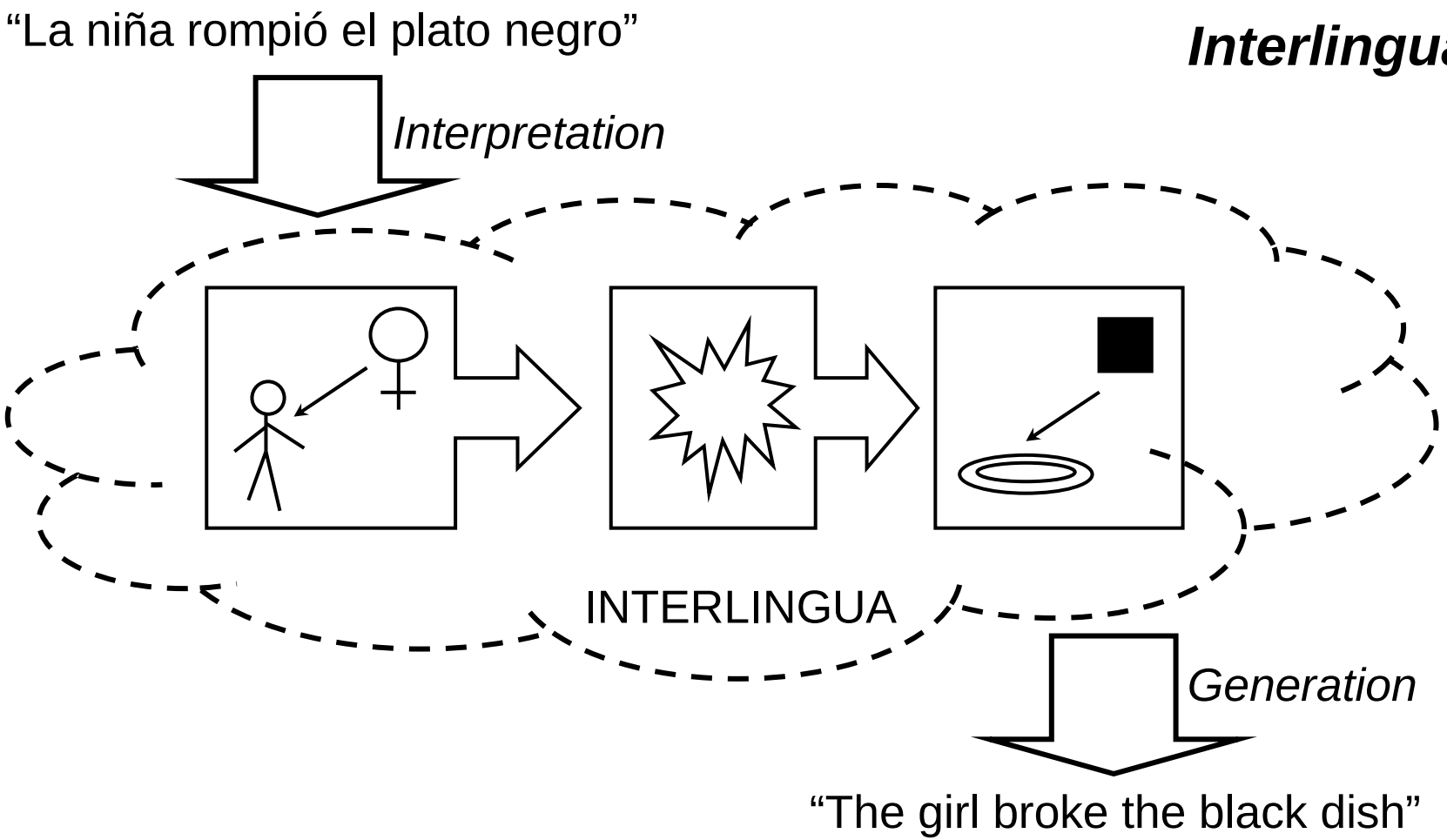
Data-based methods

Based on using huge volumes of bilingual data

Knowledge-based methods

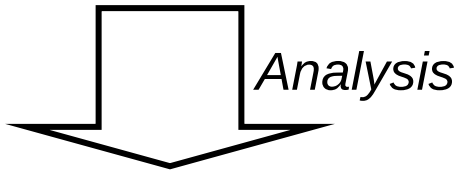


(Knight, 1997)



Transfer models

"La niña rompió el plato negro"

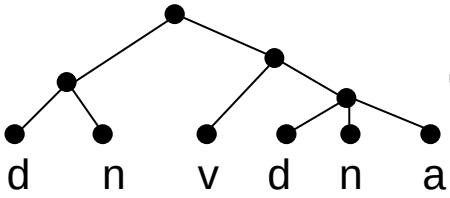


el_fs, niño_fs, romper_p3s, el_ms, plato_ms, negro_ms

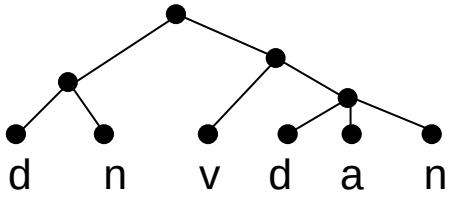
Lexical

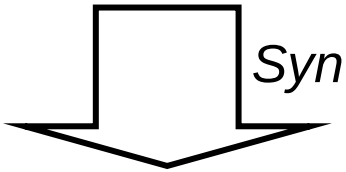
the, child_fs, break_p3s, the, dish_s, black_s

TRANSFER



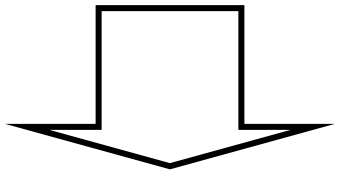
Syntactic





"The girl broke the black dish"

“La niña rompió el plato negro”

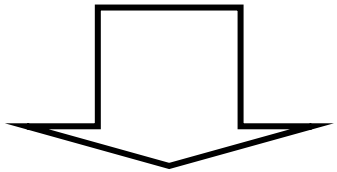


Direct translation

STEP #1: Content words are translated
(niña -> girl, rompió -> broke, plato -> dish, negro -> black)

STEP #2: Target word are reordered
(girl broke black <-> dish)

STEP #3: Final edition: (<the> girl broke <the> black dish)

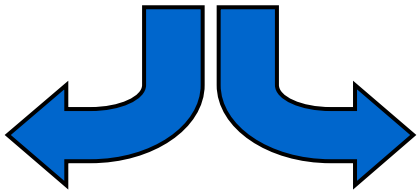


“The girl broke the black dish”

Data-based methods

Huge volumes of bilingual data (parallel corpora) are used as the main source of information for inferring new translations.

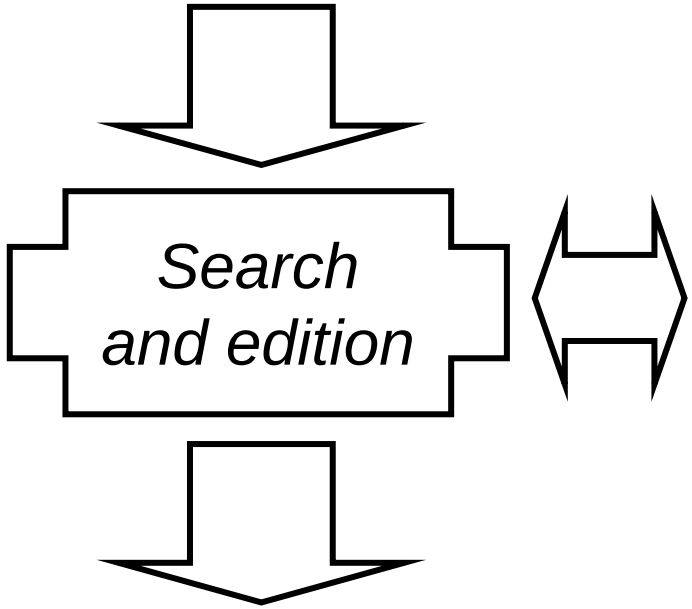
Example-based Machine Translation (EBMT)



Statistical Machine Translation (SMT)

Example-based machine translation

"La niña rompió el plato negro"



"The girl broke the black dish"

... so **the girl**, who was playing in the garden ...
... y **la niña** que estaba jugando en el parque ...

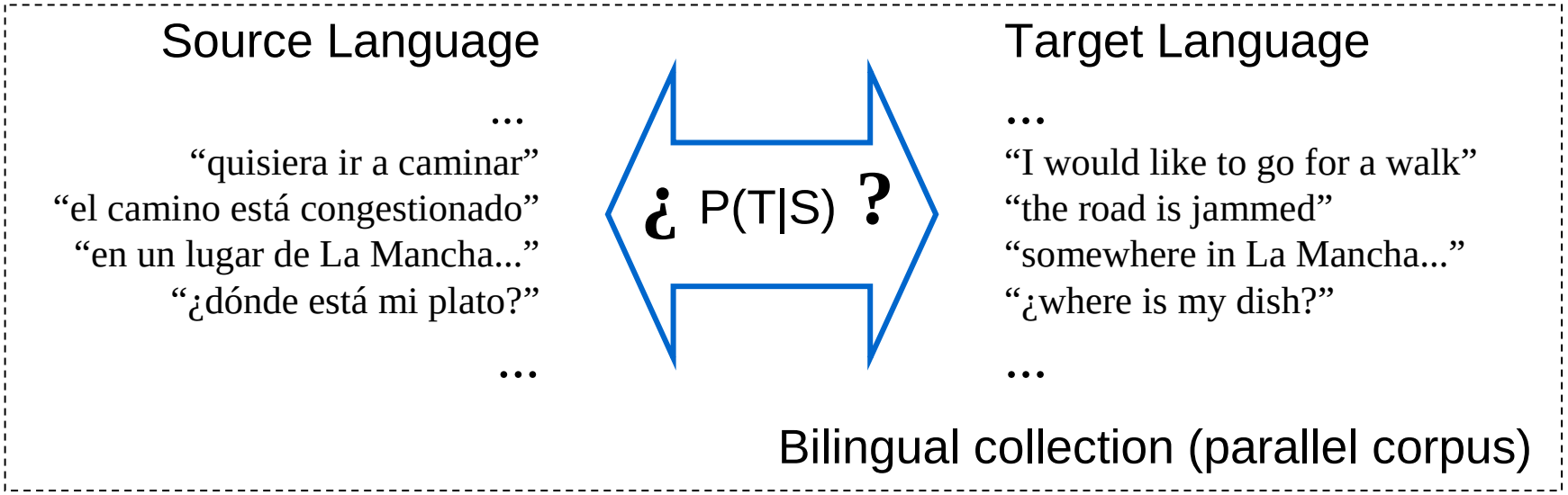
... one of the thieves **broke** his arm when he was ...
... el delincuente se **rompió** el brazo al tratar de ...

... it was actually a **black dish** with golden dots in ...
... se trataba de un **plato negro** con puntos dorados...

... and the waitress handed **the dish** out to the ...
... y el camarero le paso **el plato** al cliente sin ...

Bilingual collection
(parallel corpus)

Statistical machine translation



P(T|S): conditional probability of target sentence *T* given the source sentence *S*

$$\hat{T}_x = \underset{T}{\textbf{argmax}}\ P(T|S_x)$$

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

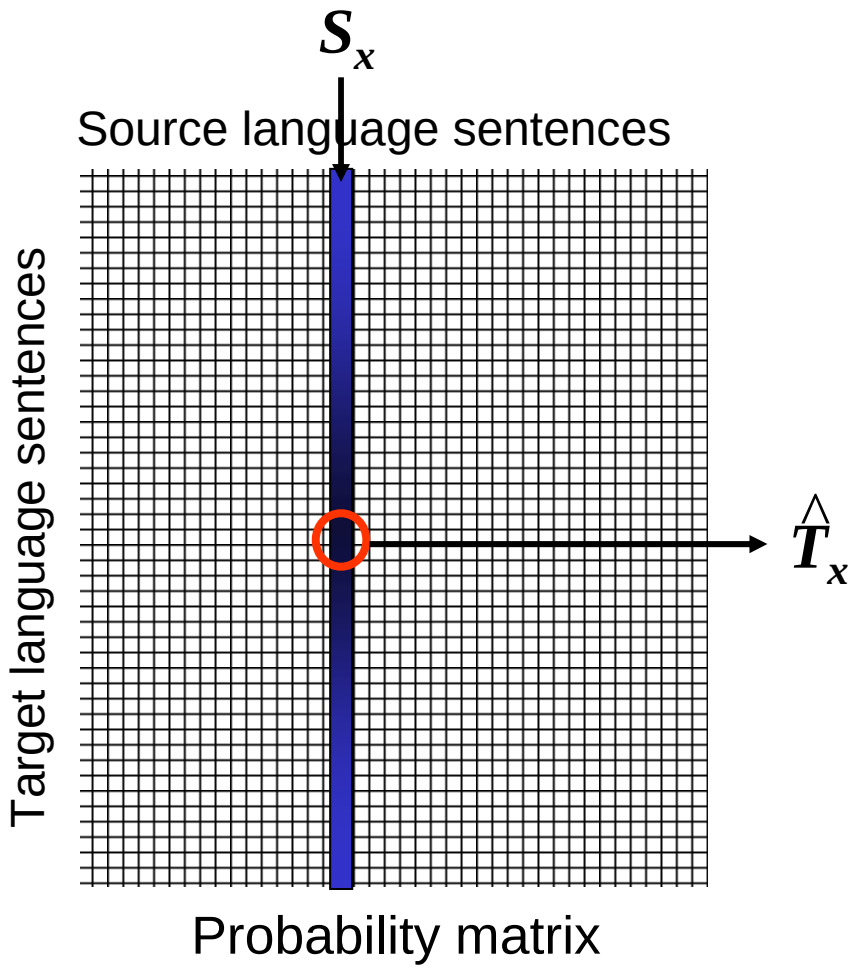
3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach



Theoretical foundation

$$\hat{T}_x = \underset{T}{\operatorname{argmax}}\, P(T|S_x)$$

The best translation will be provided by the target language sentence T_x which maximizes the conditional probability $P(T|S_x)$

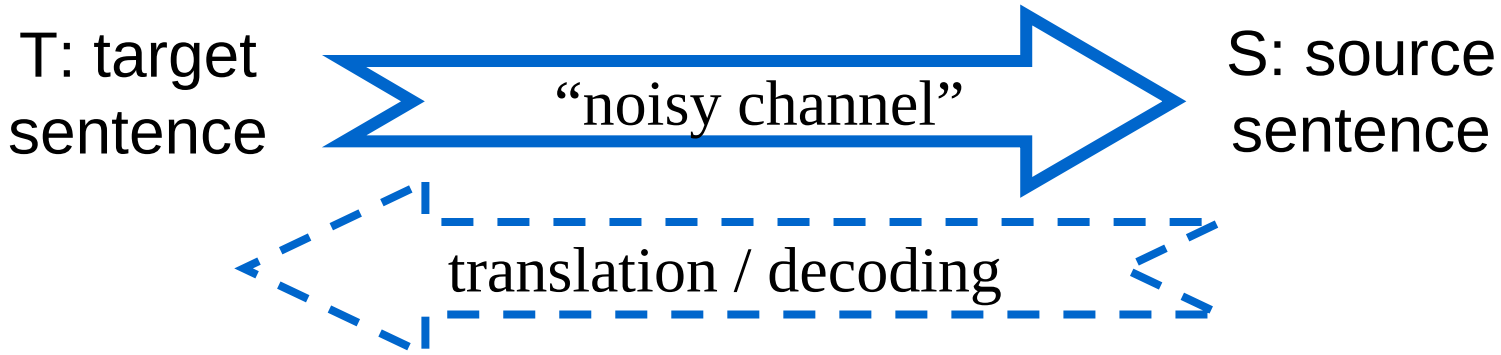
Two practical problems

1.- It is impossible to compute the values of $P(T|S)$ for a given pair of source and target languages; so, we have to restrict ourselves to approximate as much as possible such probability values.

2.- Independently from the previous issue, the search space is so huge that we should restrict ourselves to search within specific sub-regions of the search space.

The first statistical machine translation model

(Brown et al. 1993)



$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|S) = \underset{T}{\operatorname{argmax}}\, P(S|T)\, P(T)$$

Target Language Model

Translation Model

Model implications

The search for the best translation *P(T|S)* becomes the simultaneous optimization of two features:

1.- *Adequacy*: which represents the suitability of target sentences according to the translation model *P(T|S)*

2.- *Fluency*: which represents the suitability of target sentences according to the target language model *P(T)*

Target language model: P(T)

Given a sentence $T: t_1\ t_2\ t_3 \ldots t_k$

It's probability of occurrence can be computed by:

$$p(T) = p(t_1, t_2, t_3 \ldots t_k)$$

$$p(T) = p(t_1)\ p(t_2 \mid t_1)\ p(t_3 \mid t_2, t_1) \ldots p(t_k \mid t_{k-1} \ldots t_2, t_1)$$

n-gram approximation (generally n=3):

$$p(t_j \mid t_{j-1} \ldots t_2, t_1) \approx p(t_j \mid t_{j-1} \ldots t_{j-n+1})$$

n-gram model training

n-gram probabilities are easy to estimate from data

$$p(t_j \mid t_{j-1}, t_{j-2}) \approx \frac{\textit{Number of occurrences: } t_j, t_{j-1}, t_{j-2}}{\textit{Number of occurrences : } t_{j-1}, t_{j-2}}$$

n-gram models are generally interpolated and smoothed

Translation Model: P(S|T)

The first statistical translation models were the ones proposed by *Brown et al. (1993)*:

1.- they are word-based translation models: $p(s_i|t_j)$

2.- a total of 5 models of growing complexity:

IBM1 ⇨ IBM2 ⇨ IBM3 ⇨ IBM4 ⇨ IBM5

3.- they require a word-to-word aligned bilingual corpus

Word-to-word alignments

It refers to a given set of links defined between the words of a source sentence and the words of it corresponding target sentence.

For a bilingual pair of sentences *S* and *T*, containing *M* and *N* words, respectively, 2^{MN} different possible alignments can be defined!!!

Example of word-to-word alignments M=N=2

Alignment restrictions in the IBM models

In the IBM model formulation, alignments are constrained to:

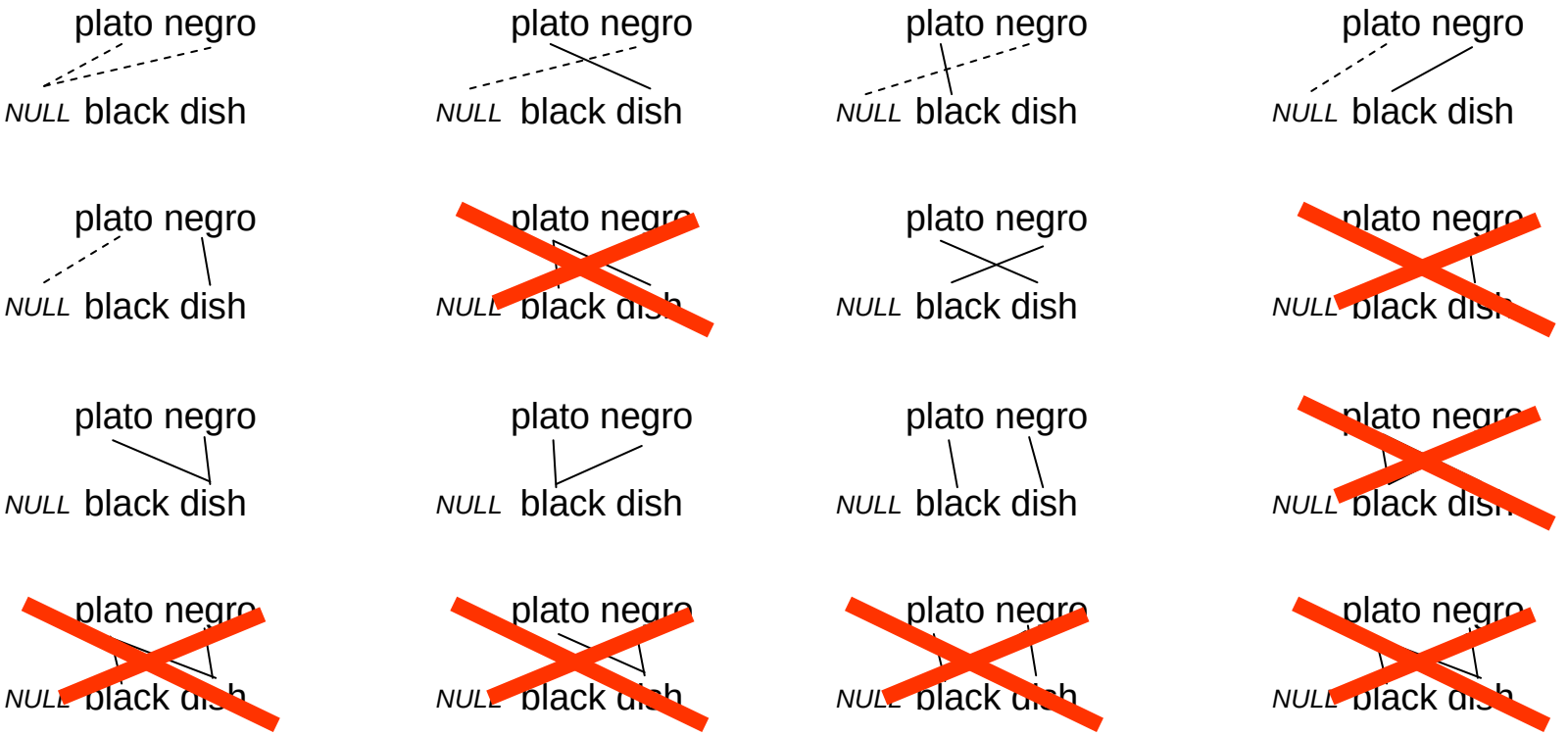
1.- all words in the source sentence *S* must have one, and only one, link associated,

2.- a *NULL* token is included in the target sentence *T* in order to allow linking those words in *S* which do not have any correspondence to any word in *T*.

In this way, the possible number of alignments is reduced from 2^{MN} to $(N+1)^{M}$.

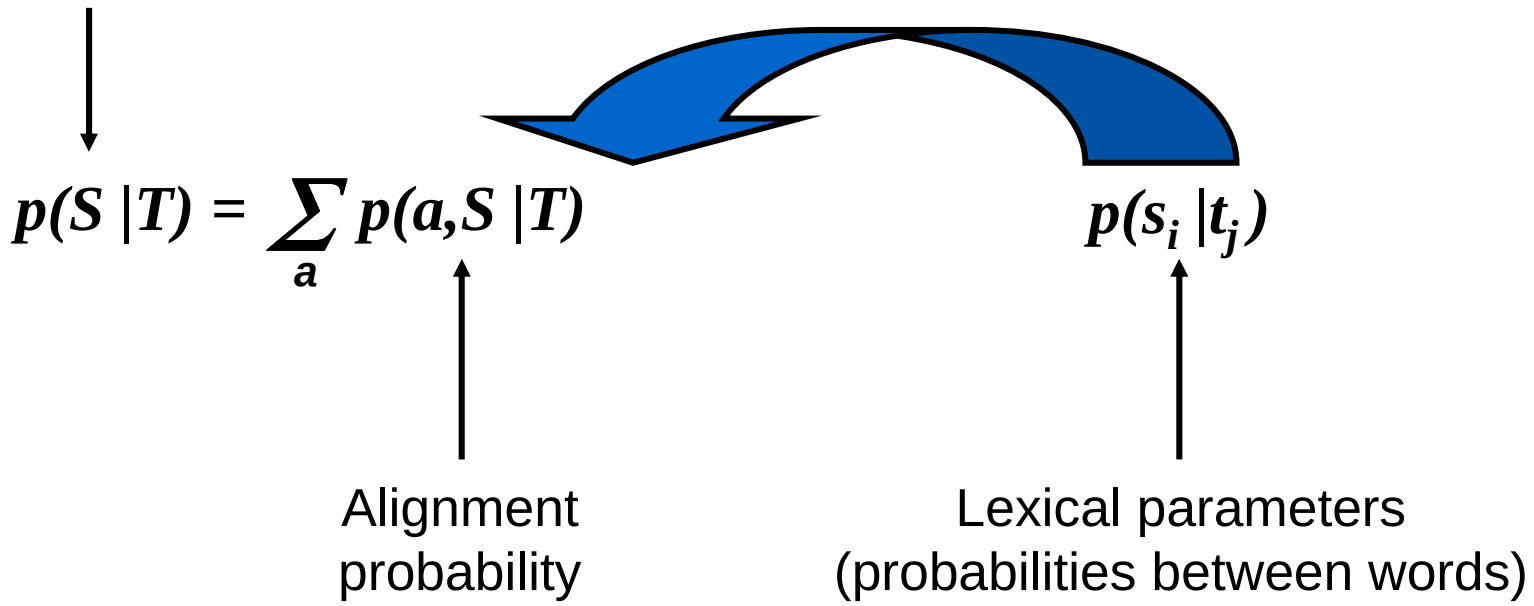
Allowed IBM model alignments (S: English)

Allowed IBM model alignments (S: Spanish)

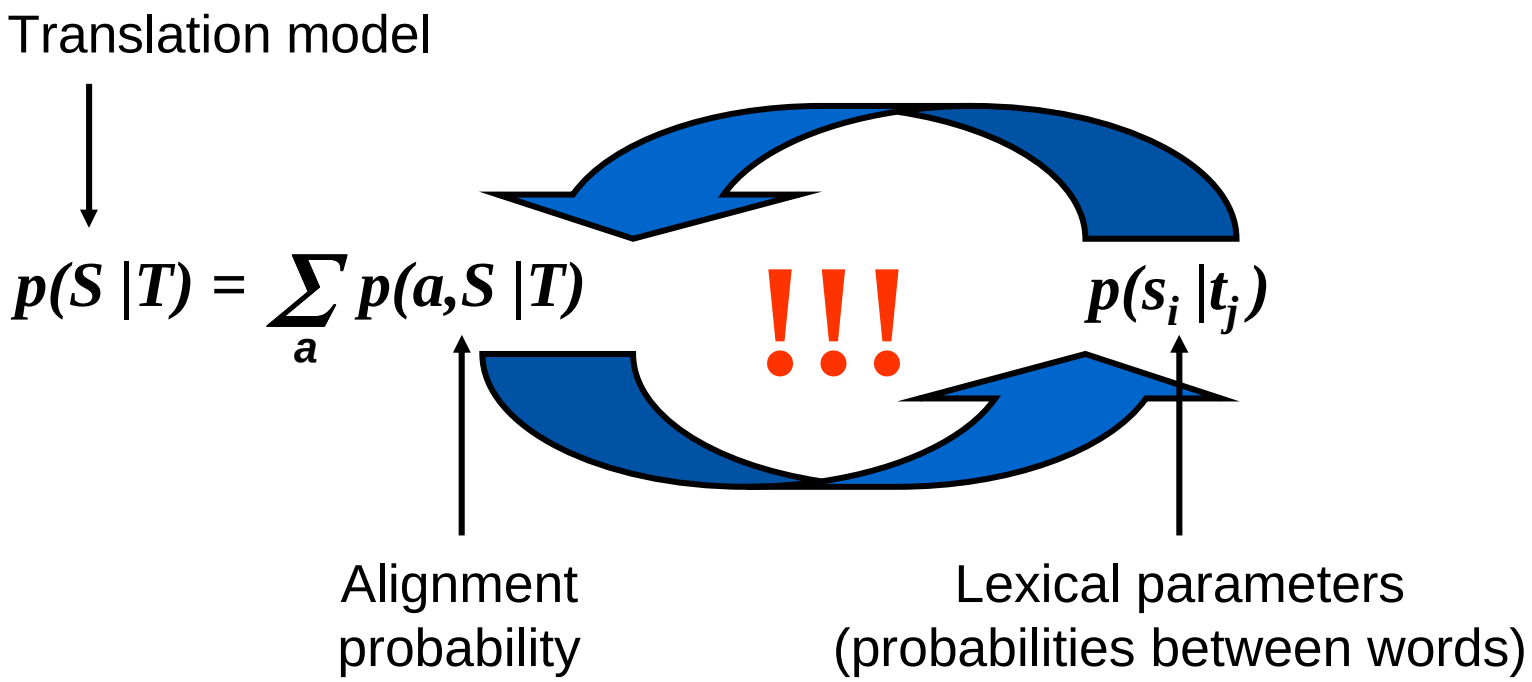


Parameter computation

Translation model



Parameter computation



The EM algorithm

Model parameters can be estimated in a recursive fashion by using the EM algorithm proposed by *Baum (1972)*:

1.- start from some initial lexical parameter values

2.- use lexical parameters to estimate alignment probabilities

3.- use alignment probabilities to update lexical parameters

4.- use lexical parameters to update alignment probabilities

5.- and so on...

The EM algorithm

$$p(S|T) = \sum_a p(a,S|T)$$

Initial Guess

$$p(a,S|T) = \prod_k t(s_k|t_{ak})$$

$$p(a|S,T) = \frac{p(a,S|T)}{\sum_a p(a,S|T)}$$

$$c(s|t) = \sum_n \sum_a p(a|S_n,T_n)\ \delta(a,s,t)$$

$$t(s|t) = \frac{c(s|t)}{\sum_s c(s|t)}$$

Other IBM model parameters

In addition to lexical parameters and alignment probabilities, the IBM models consider other important parameters:

1.- Fertility: represents the probabilities for target words being associated to *0, 1, 2* or more alignment links.

2.- Distortion: represents the probabilities for the lengths of alignment links.

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

Two main changes in one decade

Noisy Channel approach → Log-linear combination of models

(Och y Ney, 2002)

Word-based translation models → Phrase-based translation models

(Zens et al. 2002, Koehn et al. 2003)

Log-linear combination of models

It constitutes a more general approach, which is based on maximum entropy principles *(Berger et al. 1996)*

$$\hat{T} = \underset{T}{argmax}\ p(T|S) \approx \underset{T}{argmax} \prod_i p_i(S,T)^{\lambda_i}$$

Noisy Channel ⇨ particular case:

$$p_1(S,T) = p(S|T),\ p_2(S,T) = p(T),\ y\ \lambda_1 = \lambda_2 = 1$$

Weight optimization

For a given combination of *m* models, the values of λ_1 , λ_2 ... λ_m are computed by using an optimization process.

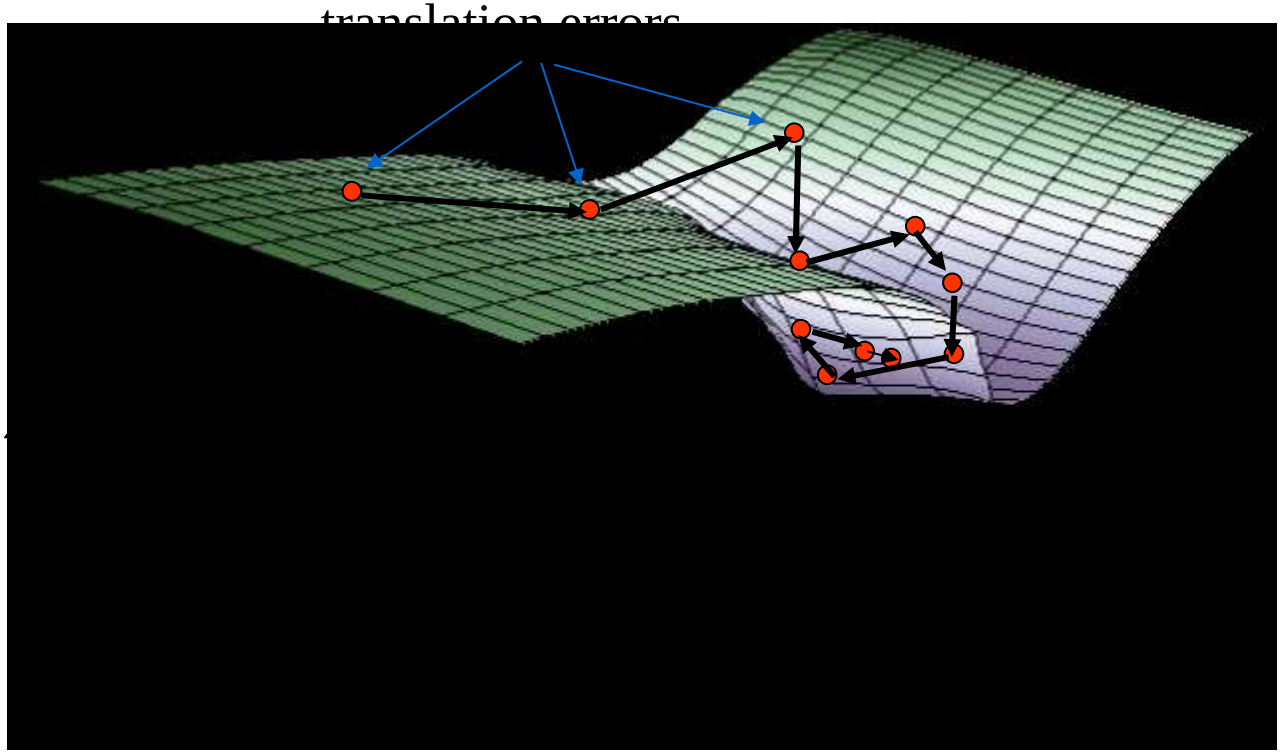
1.- a “development” corpus is used,

2.- translation quality is automatically evaluated,

3.- the weights are adjusted in an iterative fashion until reaching the maximum translation quality over the development corpus.

Optimization algorithms

- *Simplex*
- *SPSA*



Error

error = f (λ_1 ,

Phrase-based translation models

Phrases are bilingual units that can contain one or more source and target words.

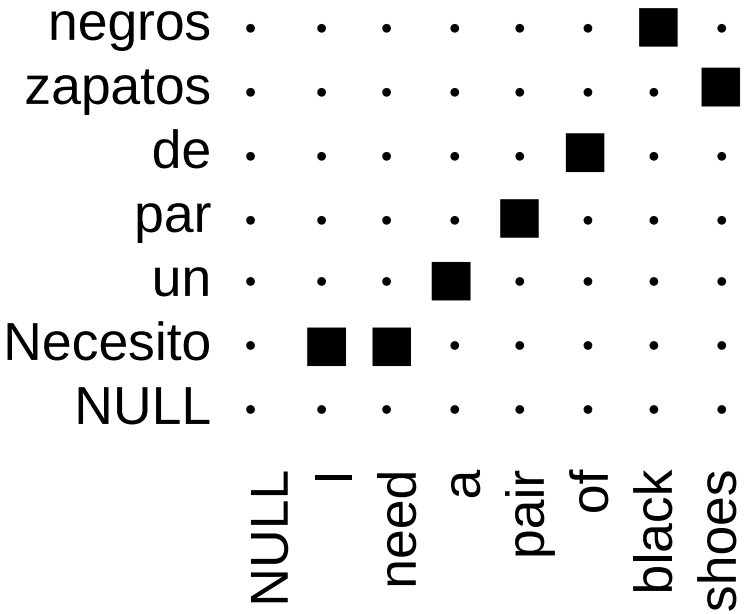
1.- all words inside a phrase can only be aligned to words within the same phrase,

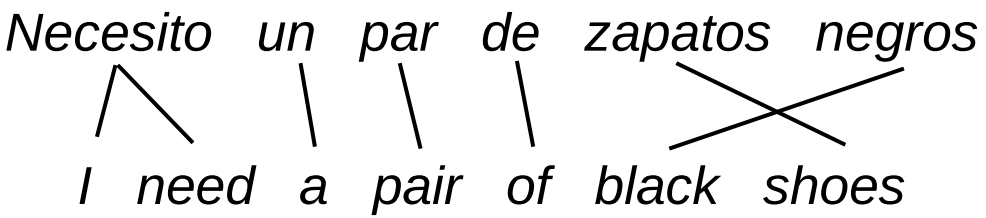
2.- their length (number of words) is usually restricted to a maximum value,

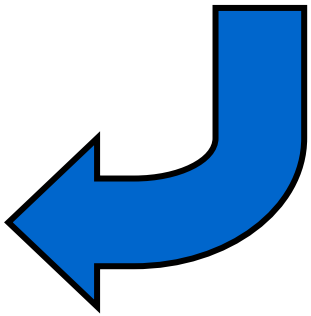
3.- their probabilities are estimated by counting occurrences within the training corpus: $p(s|t) = N(s,t) / N(t)$

Example of phrase extraction

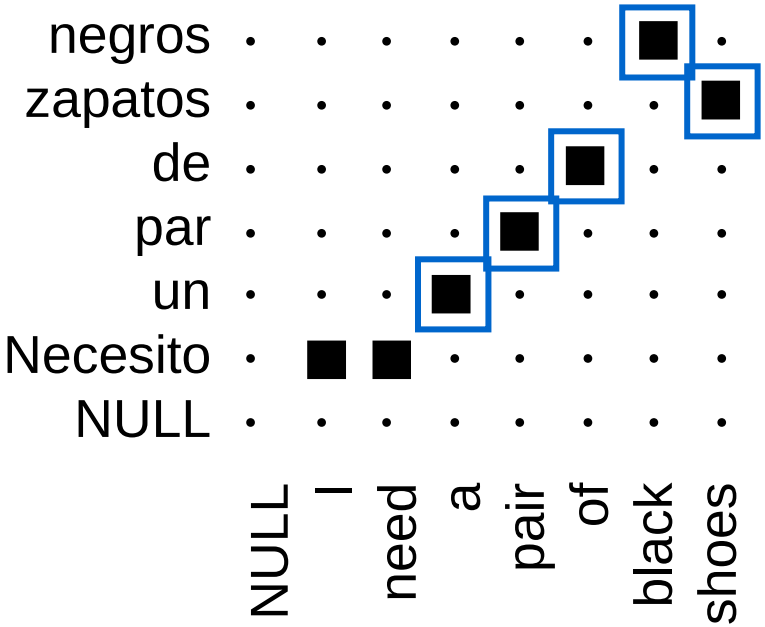
Matrix representation







Matrix representation

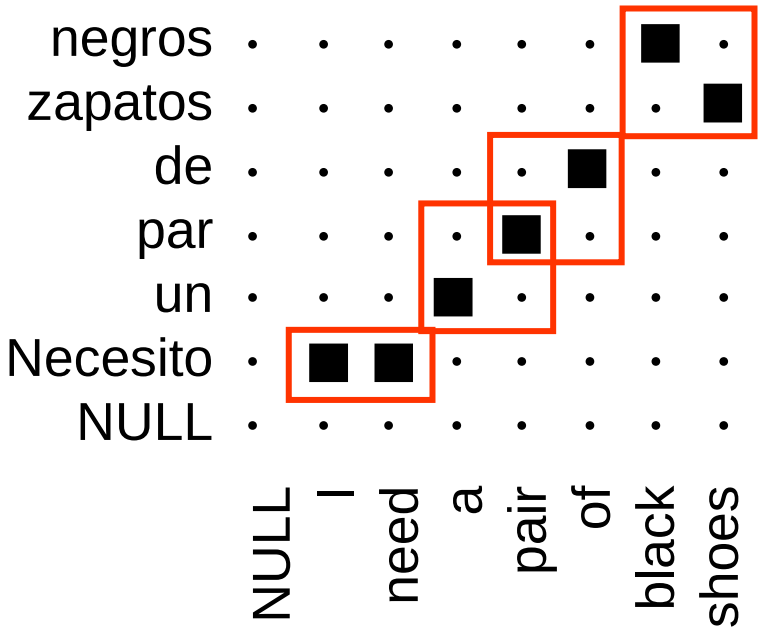


Example of phrase extraction

Phrases of maximum length = 1

un # a ; par # pair ; de # of
negros # black ; zapatos # shoes

Matrix representation



Example of phrase extraction

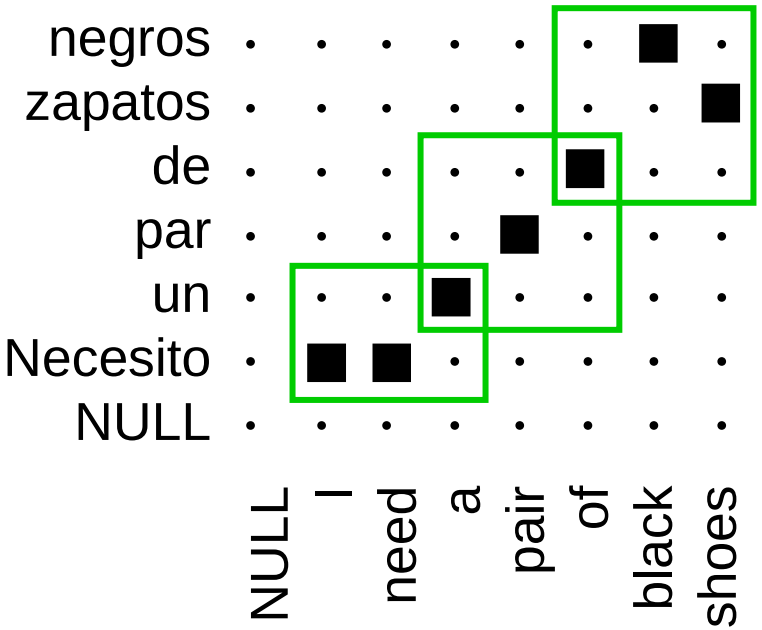
Phrases of maximum length = 1

un # a par # pair de # of
negros # black zapatos # shoes

Phrases of maximum length = 2

necesito # I need un par # a pair
par de # pair of
zapatos negros # black shoes

Matrix representation



Example of phrase extraction

Phrases of maximum length = 1

un # a par # pair de # of
negros # black zapatos # shoes

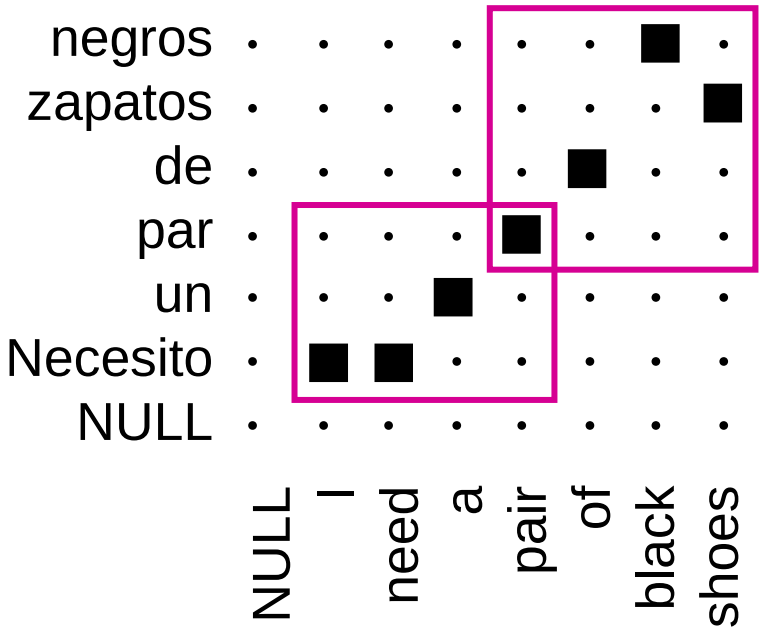
Phrases of maximum length = 2

necesito # I need un par # a pair
par de # pair of
zapatos negros # black shoes

Phrases of maximum length = 3

necesito un # I need a
de zapatos negros # of black shoes
un par de # a pair of

Matrix representation

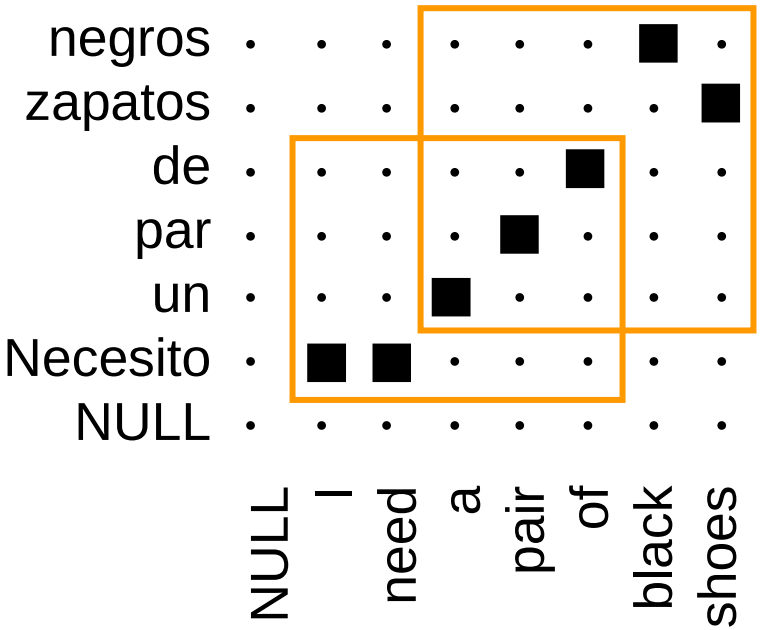


Example of phrase extraction

Phrases of maximum length = 4

necesito un par # I need a pair
par de zapatos negros # pair of black shoes

Matrix representation



Example of phrase extraction

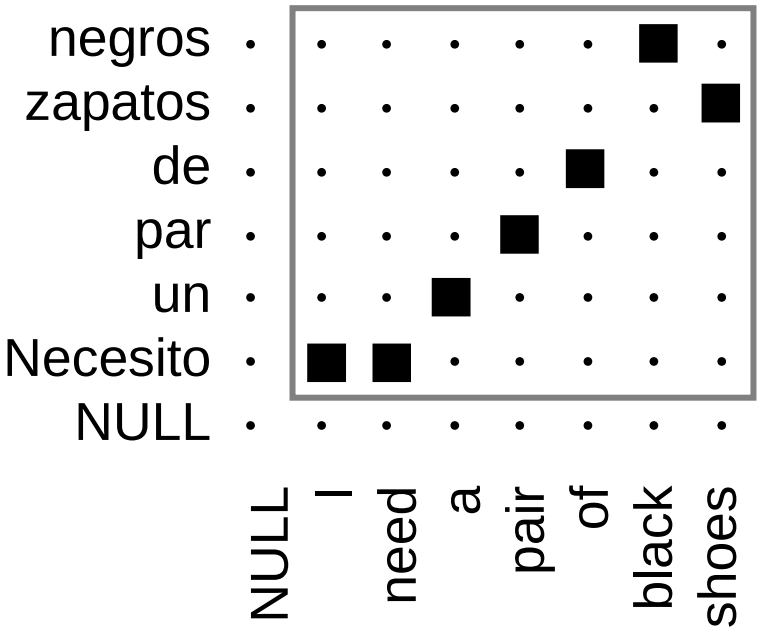
Phrases of maximum length = 4

necesito un par # I need a pair
par de zapatos negros # pair of black shoes

Phrases of maximum length = 5

necesito un par de # I need a pair of
un par de zapatos negros # a pair of black shoes

Matrix representation



Example of phrase extraction

Phrases of maximum length = 4

necesito un par # I need a pair
par de zapatos negros # pair of black shoes

Phrases of maximum length = 5

necesito un par de # I need a pair of
un par de zapatos negros # a pair of black shoes

Phrases of maximum length = 7

necesito un par de zapatos negros # I need a pair of black shoes

Feature functions for phrase probabilities

Typically, for phrase probability models are combined to implement the phrase-based translation model:

1.- Relative frequencies: $p(s,t) = N(s,t) / N(t)$

Generally computed in both the source-to-target and target-to-source directions.

2.- IBM1-Model: $p(s,t) = (I+1)^{-J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i \mid s_n^j)$

Also, generally computed in both directions.

More feature functions for phrases

But actually, any additional model can be implemented and included into the log-linear combination:

1.- Bonuses and/or penalties.

2.- Inspired on lingüistic information:

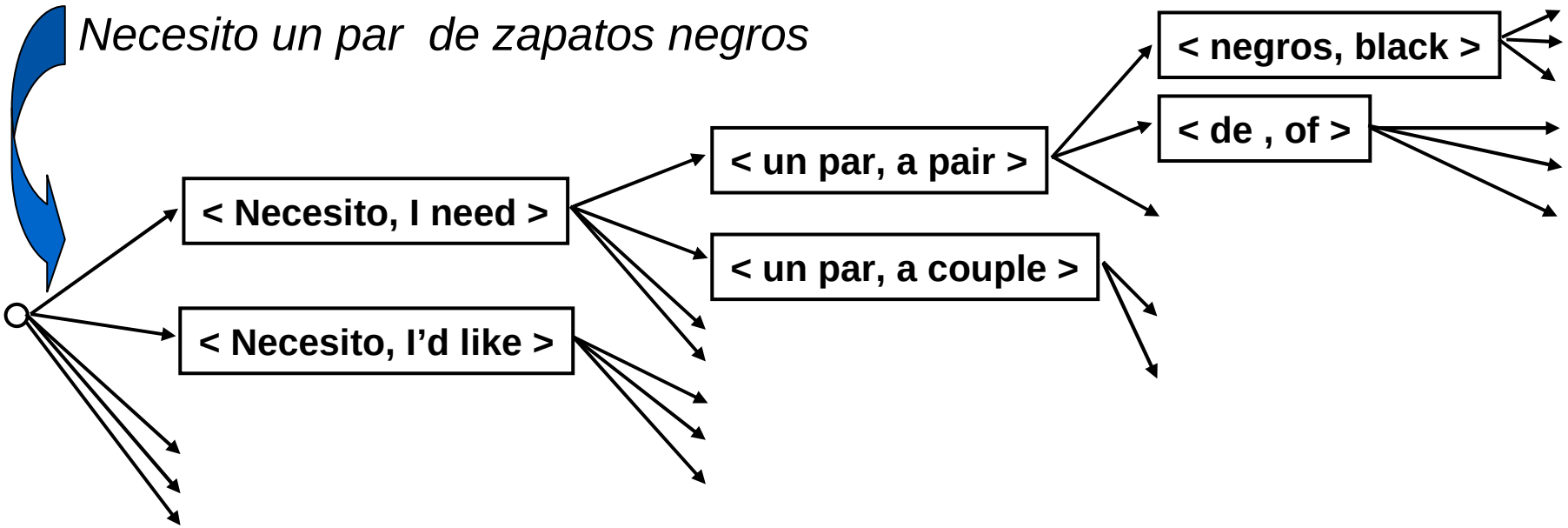
- Morphology (lexicon, lemmas, POS)
- Syntax (parsing, dependencies)

3.- Inspired on semantics:

- Style/topic classification and/or detection

Phrase-based decoding

The target-sentence space is explored by means of a search algorithm *(Wang y Waibel 1997, Tillman et al. 1997, Koehn 2004)*



N-best lists and re-ranking

In addition to the “best” translation output, some decoders can provide a list of the *n* best translation hypothesis.

Such an n-best list can be re-ranked or post-edited by using some additional information in order to further improve the translation results.

Some times the best translation is not necessarily the first decoder’s choice.

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

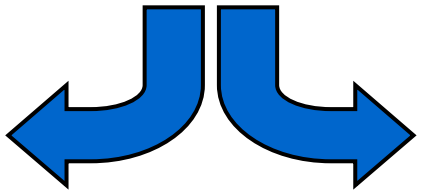
5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

Evaluating the translation quality

Types of evaluation procedures



Human evaluation

- Based on human judgments
- Subjective scores

Automatic evaluation

- Based on comparative metrics
- Use translation references

Human evaluation

A subjective numeric scale from 1 (worst) to 5 (best) is used

1.- Adequacy: measures how much information from the original source sentence is preserved in the translation.

2.- Fluency: measures whether the translation is well-formed or not in terms of the target language without taking into account the meaning of the source sentence.

Automatic evaluation

Computed translations are compared to human translations

1.- Accuracy metrics:

- BLEU: geometric mean of n-gram precisions ⇨ $[0, 1]$
- NIST: arithmetic mean of n-gram precisions ⇨ $[0, \infty)$

2.- Error rates:

- WER: word error rate
- PER: position independent error rate

} ⇨ $[0\%, 100\%]$

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

Experimental framework

- TC-STAR project: http://www.tc-star.org
- Integration of the three technologies involved in speech-to-speech translation: ASR, MT and TTS.
- Wide domain real life data: official transcriptions of the European Parliament Plenary Sessions (EPPS).

	Sentences	Words	Vocabulary
English	**1.220.000**	**33.400.000**	**105.000**
Spanish	**1.220.000**	**34.800.000**	**169.000**

Systems evaluated

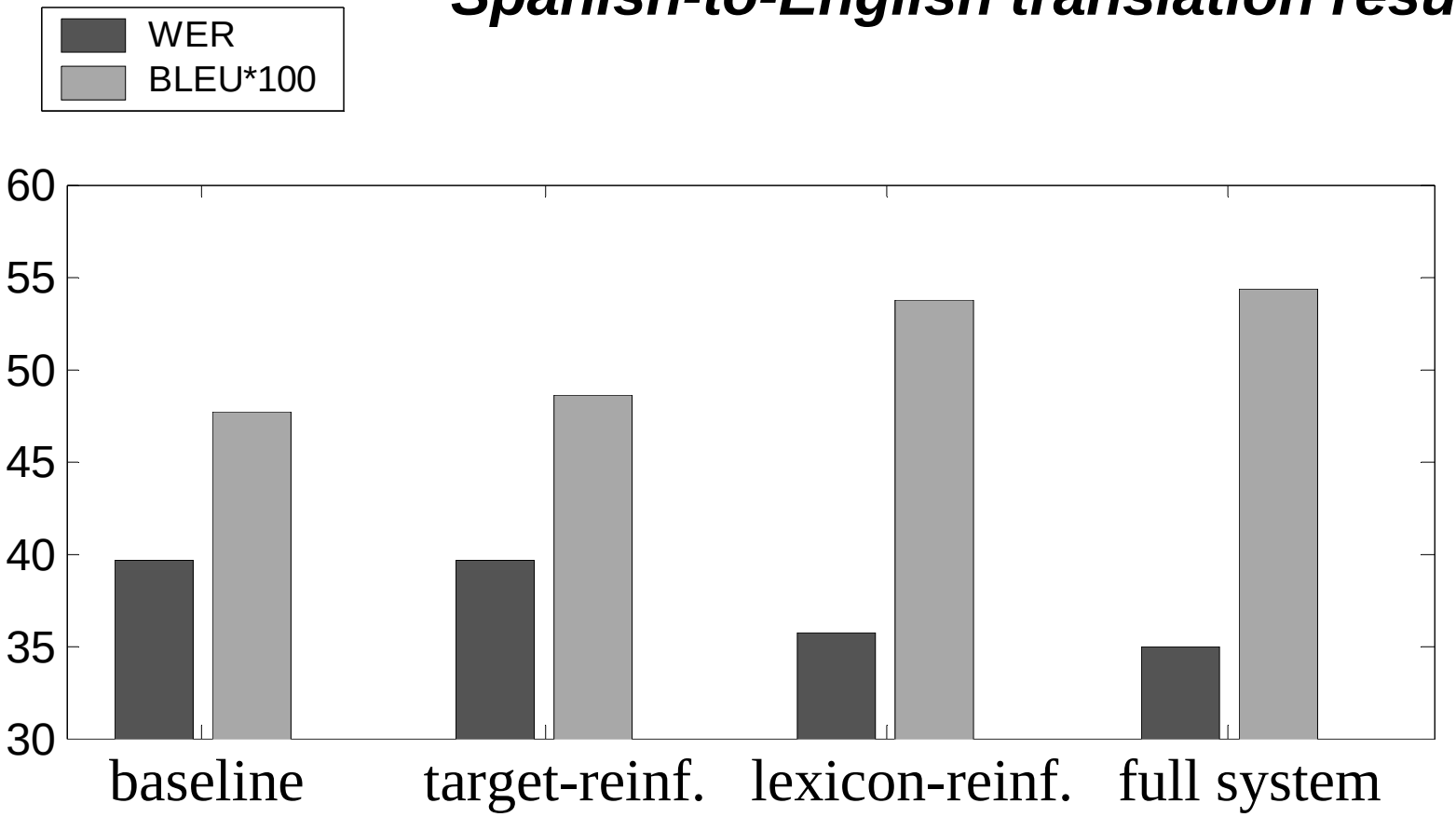
- **Baseline**
  translation model only: n-gram based ***(Mariño et al. 2005)***

- **Target-reinforced**
  translation model + language model + word bonus

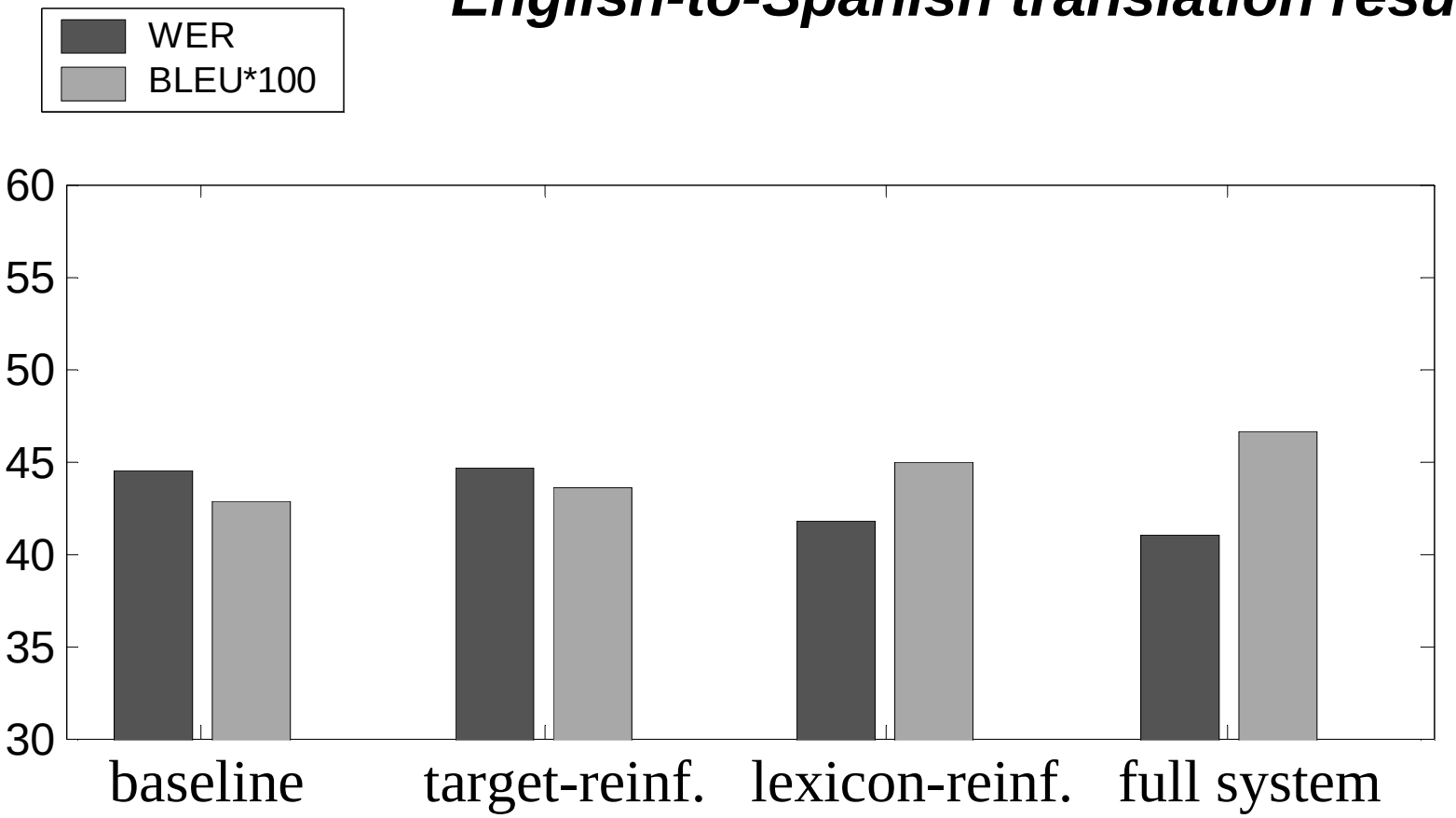
- **Lexicon-reinforced**
  translation model + forward and backward IBM1 models

- **Full system**
  combines all five models

Spanish-to-English translation results



English-to-Spanish translation results



Important observations

- Spanish-to-English translations are both significantly and consistently better than English-to-Spanish translations.
- The additional feature models considered provide important improvements in the translation quality.
- The IBM1-based lexical models are the ones having the largest impact on translation quality improvement.

Translation error analysis

- " The policy of the European Union on Cuba ***NULL*** **must** [must not] change . "
- " To achieve these purposes , it is necessary *NULL* for the governments **to be allocated** [to allocate] , at least , 60 000 million *NULL* dollars a year ... "
- " In the UK we have ***NULL*** [already] **laws enough** [enough laws] , but we want to encourage *NULL* other States ... "

The most commonly encountered type of errors

	Eng-to-Spa	Spa-to-Eng
• Incorrect verbal forms	31.3%	29.9%
• Omitted translations	22.0%	26.1%
• Incorrect word order	15.9%	19.7%
• Concordance problems	10.8%	4.6%
• Other type of problems	20.0%	19.7%

Demo on speech translation technology

- "*off-line" processing*
- Integrates Automatic Speech Recognition (ASR) and Spoken Language Translation (SLT) technologies.
- Translation direction: Spanish ⇨ English

CONTENTS

1.- Brief historical notes on machine translation

2.- Different approaches to the problem of machine translation

3.- A more detailed discussion on the statistical approach

4.- The state of the art in statistical machine translation

5.- Evaluation metrics for machine translation performance

6.- Statistical machine translation of European Parliament speeches

7.- Future perspectives for the statistical machine translation approach

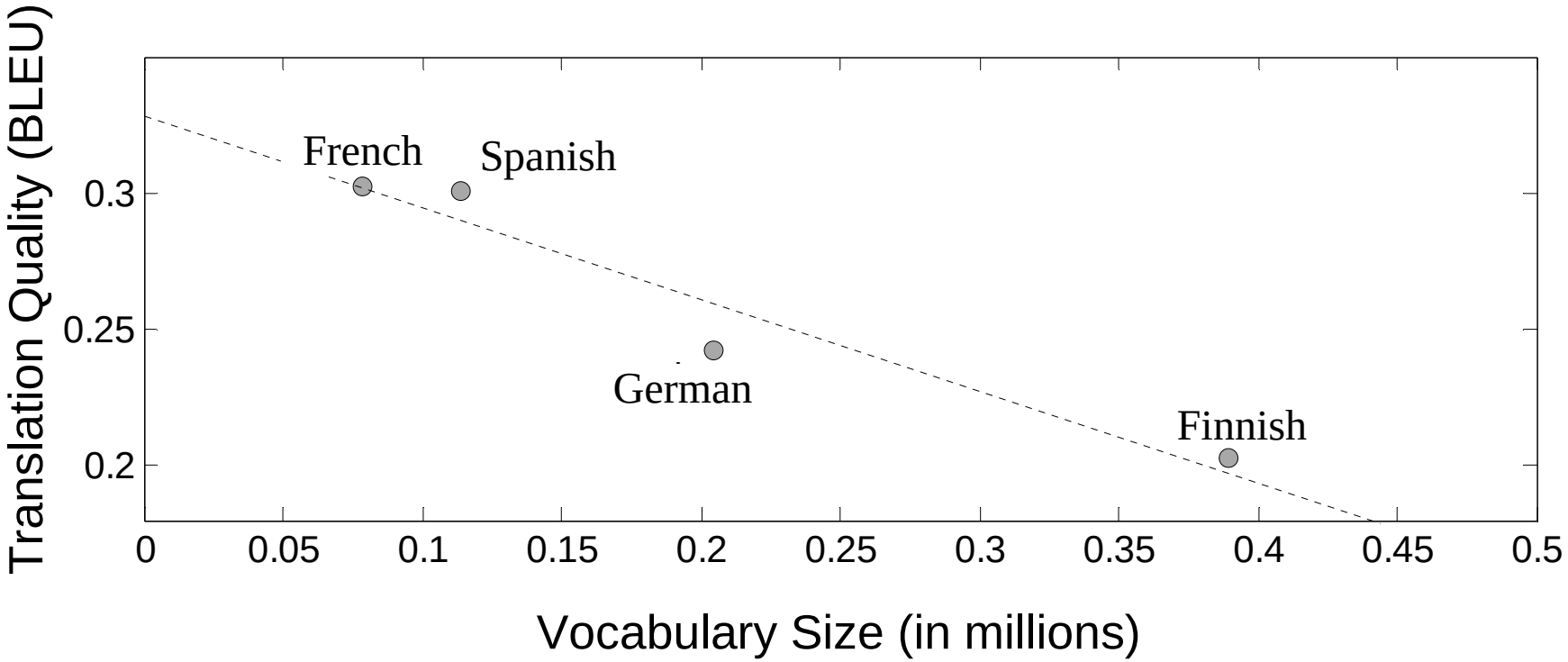
The most important challenges

1.- Morphology: it directly affects the vocabulary sizes generating data sparseness problems.

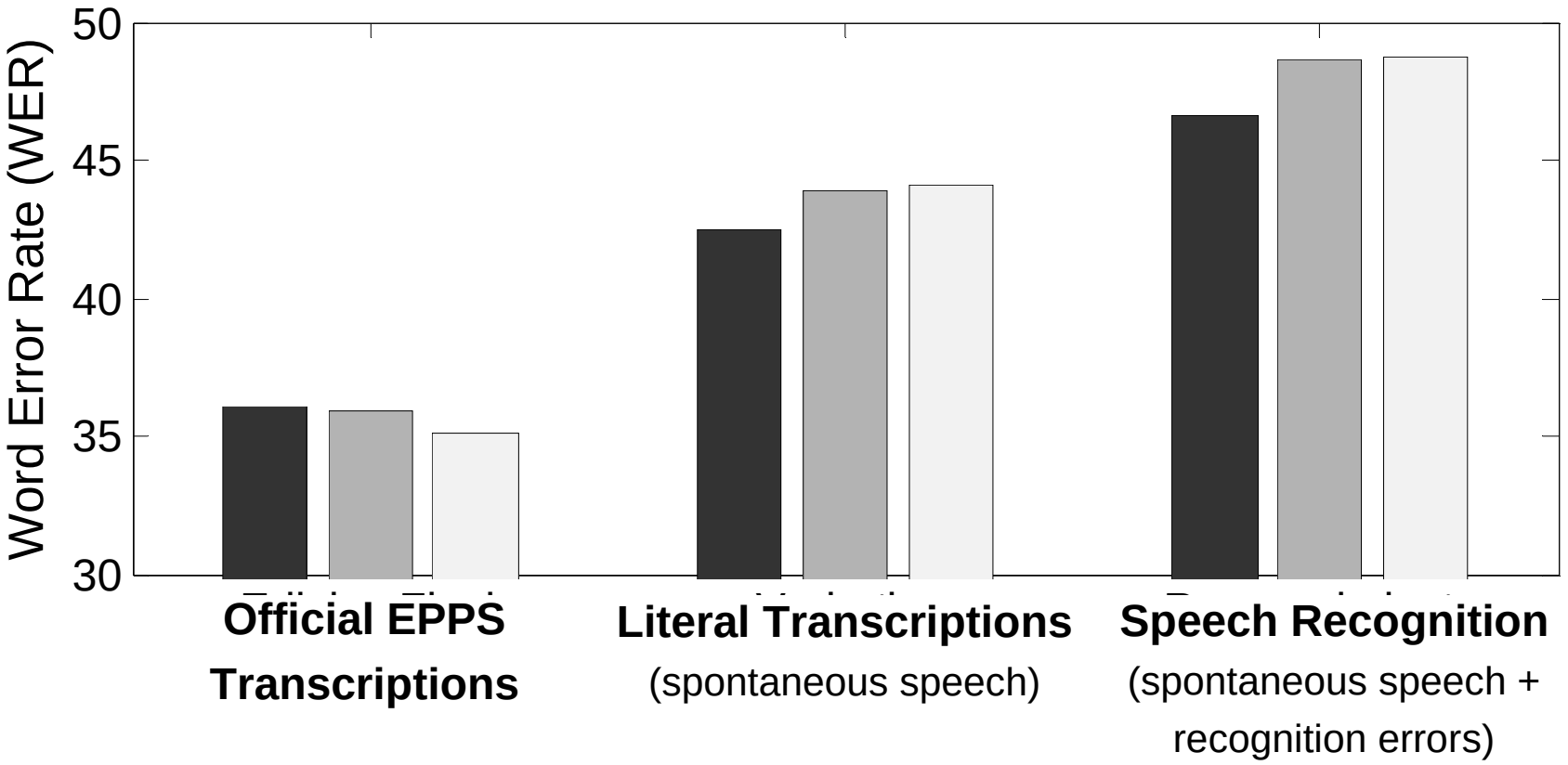
2.- Word reordering: it affects translations between distant languages, it actually is a very expensive problem from the computational point of view.

3.- Spoken language: spontaneous speech effects (such as repetitions, interruptions and ungrammaticality) severely affect translation quality.

Comparison of translation qualities when translating into English from four different source languages



Spoken language effects on translation performance



Limitations of evaluation metrics

1.- Human evaluation is slow and expensive.

2.- The existent automatic evaluation metrics:

- are suitable for globally evaluating a set of translations,
- are not reliable for evaluating individual translations,
- strongly depend on the available set of references.

Further research

During the next years, MT research will focus on the:

1.- incorporation of linguistic knowledge into the statistical machine translation framework,

2.- development of efficient strategies for dealing with the problem of word reordering,

3.- development of new models for taking into account spontaneous speech effects,

Further research

4.- design of more appropriate interfaces for speech recognition and machine translation systems,

5.- development of methods for a more effective identification and usage of paralinguistic information,

6.- development of more efficient and reliable automatic evaluation metrics.

How to get started in SMT…

- Bilingual corpora: http://www.europarl.eu.int/
- Word alignments: http://www.fjoch.com/GIZA++.html
- Language modeling: http://www.speech.sri.com/projects/srilm/
- Decoding: http://www.isi.edu/licensed-sw/pharaoh/
- Other links: http://www.statmt.org/wpt05/mt-shared-task/

  http://gps-tsc.upc.es/veu/soft/soft/marie/

  http://www.isi.edu/licensed-sw/carmel/

http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html

Nanyang Technological University, Singapore
August 31, 2007

Statistical Machine Translation: State of the Art and Future Challenges

Rafael E. Banchs

Universitat Politècnica de Catalunya – Barcelona Media Centre d'Innovació