A Feasibility Study for Chinese-Spanish Statistical Machine Translation

Rafael E. Banchs

Josep M. Crego, Patrik Lambert, José B. Mariño

Univesitat Politècnica de Catalunya, Barcelona, Spain

Motivation

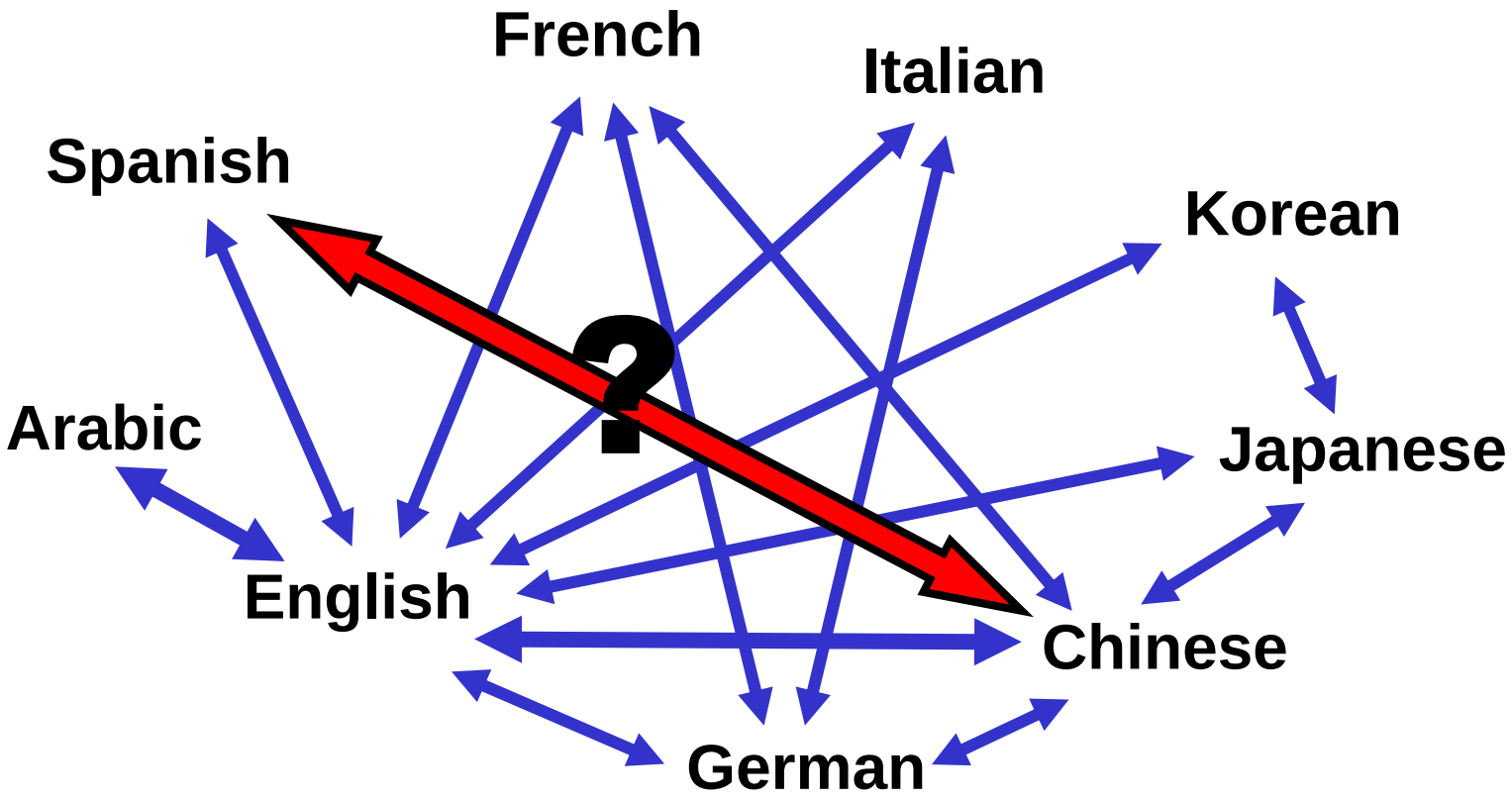
- *The most spoken languages in the world*

Language	*Native Speakers*
Chinese	**874.000.000**
Spanish	**332.000.000**
English	**322.000.000**
Hindi	**291.500.000**
Arabic	**193.000.000**
Bengali	**189.000.000**

Language	*Including 2da Language*
Chinese	**1.051.000.000**
Hindi	**594.000.000**
English	**510.000.000**
Spanish	**420.000.000**
Russian	**255.000.000**
Arabic	**230.000.000**

Motivation

- *Little research in Chinese-Spanish machine translation*



Objectives of the Work

Main problem: There is not a Chinese-Spanish parallel corpus for training purposes (at least as a freely available resource).

Main Objectives:

• To evaluate a method for developing a Chinese-Spanish SMT system by using an artificially constructed corpus

• To compare direct translation by using the developed system with indirect translation by using English as a “bridge language”

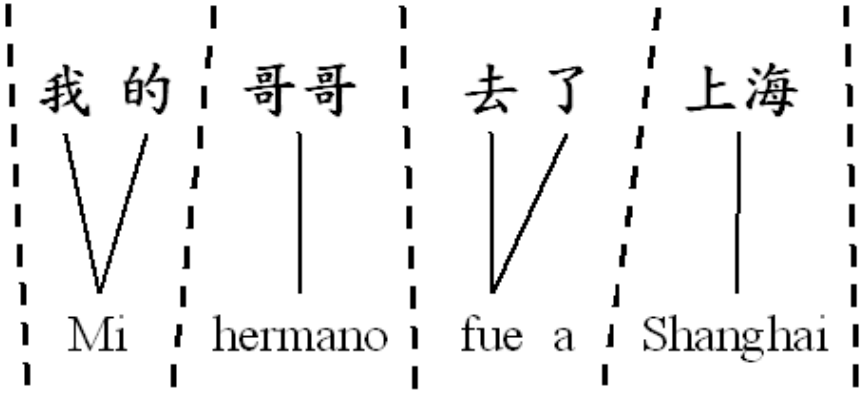
• To compare the developed statistical system with a Rule-based MT system

The Considered SMT Approach

N-gram Based SMT (Mariño *et al.*, 2006)

Translation model: n-gram language model of bilingual units

$$p_{TM}(T,S) \approx \prod_{k=1}^{K} p\left((t,s)_k \mid (t,s)_{k-1}, (t,s)_{k-2}\right)$$



Example of sentence segmentation in bilingual units (*tuples*)

The Considered SMT Approach

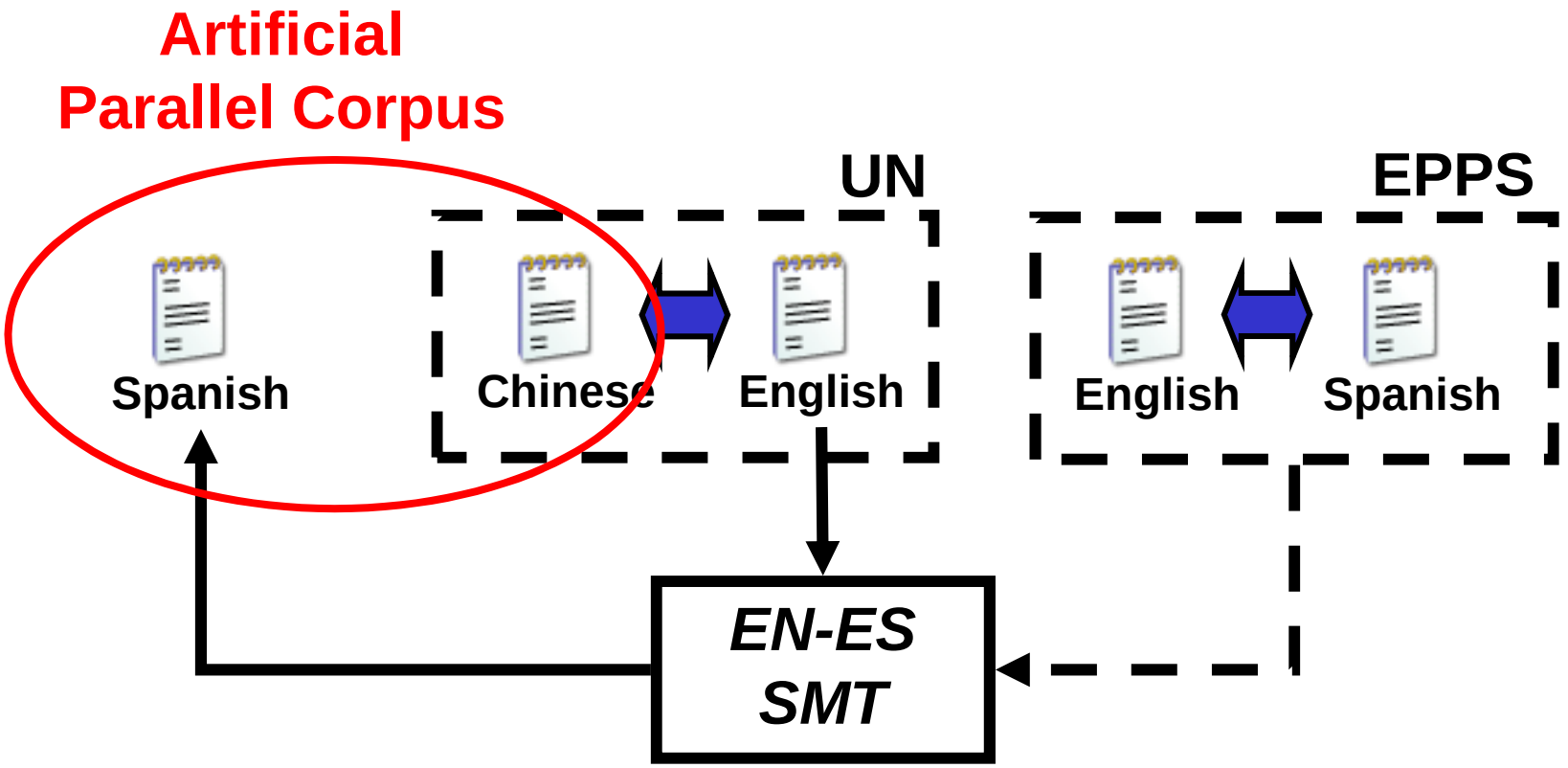
N-gram Based SMT (Mariño *et al.*, 2006)

Additional feature functions:

- Target language model: a 4-gram model of target words
- Word bonus: a sentence length bonus factor
- Lexicon models: source-to-target and target-to-source IBM-1 lexical probabilities

Beam search with monotonic decoding (no reordering)

Parallel Corpus Construction



Parallel Corpus Statistics

Original Chinese-English and Spanish-English Corpora

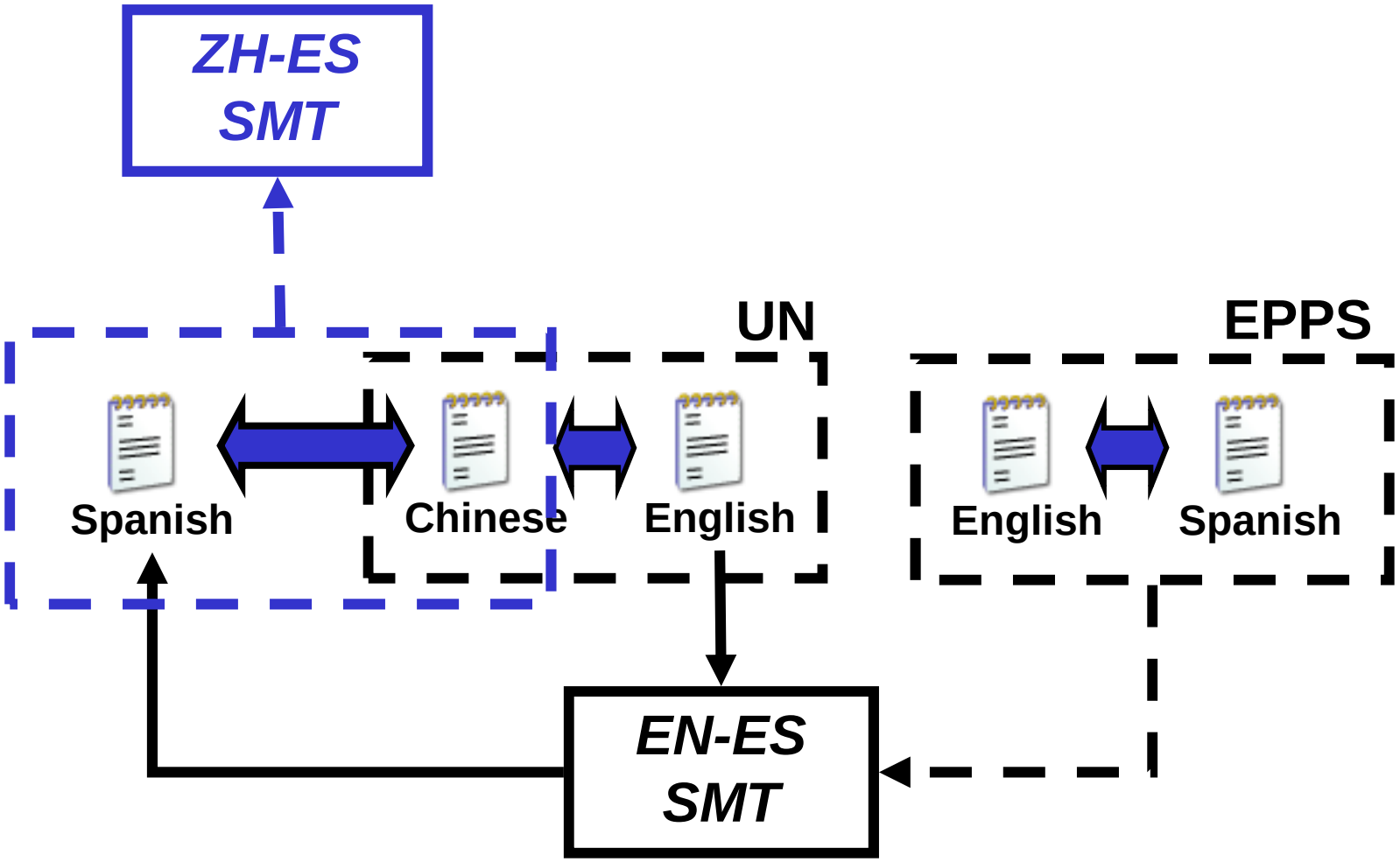
Corpus	Language	Sentences	Words	Vocab.
ZH-EN	Chinese	105 K	1.9 M	29.5 K
	English	105 K	2.1 M	34.8 K
ES-EN	Spanish	105 K	2.0 M	40.0 K
	English	105 K	2.0 M	27.0 K

Constructed Chinese-Spanish Parallel Corpus

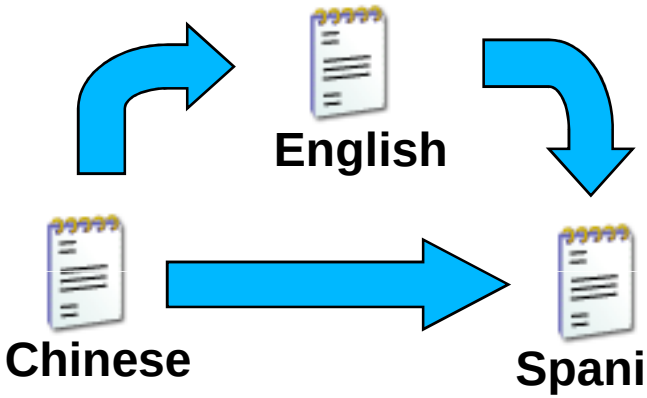
Corpus	Language	Sentences	Words	Vocab.
TRAIN	Chinese	105 K	1.9 M	29.5 K
	Spanish	105 K	2.0 M	34.8 K
DEV*	Chinese	330	6.0 K	1.6 K
	Spanish	330	6.8 K	2.0 K
TEST*	Chinese	100	1.9 K	813
	Spanish	100	2.1 K	908

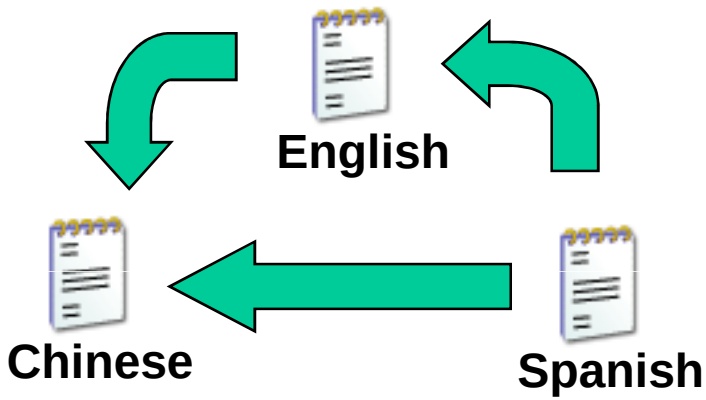
** Manually constructed*

Chinese-Spanish SMT Training



Direct Vs. Indirect Translation

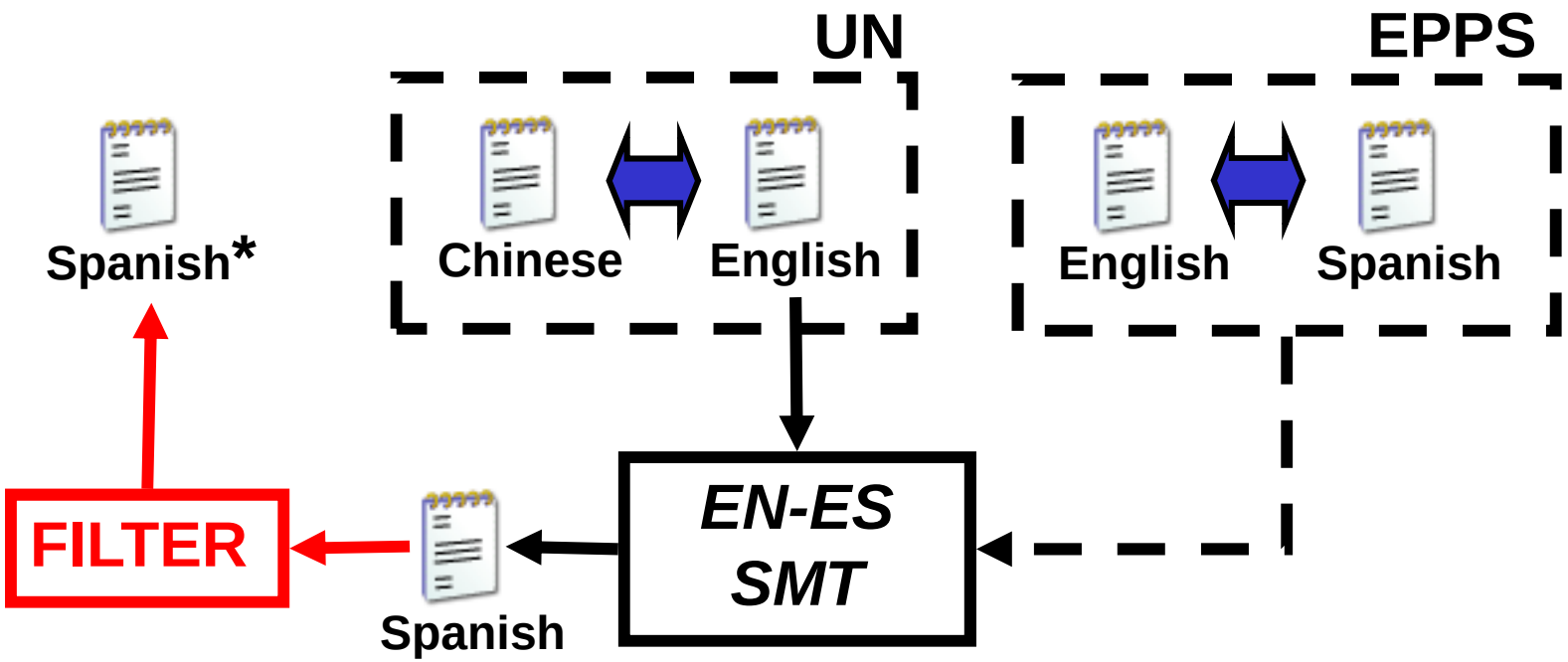




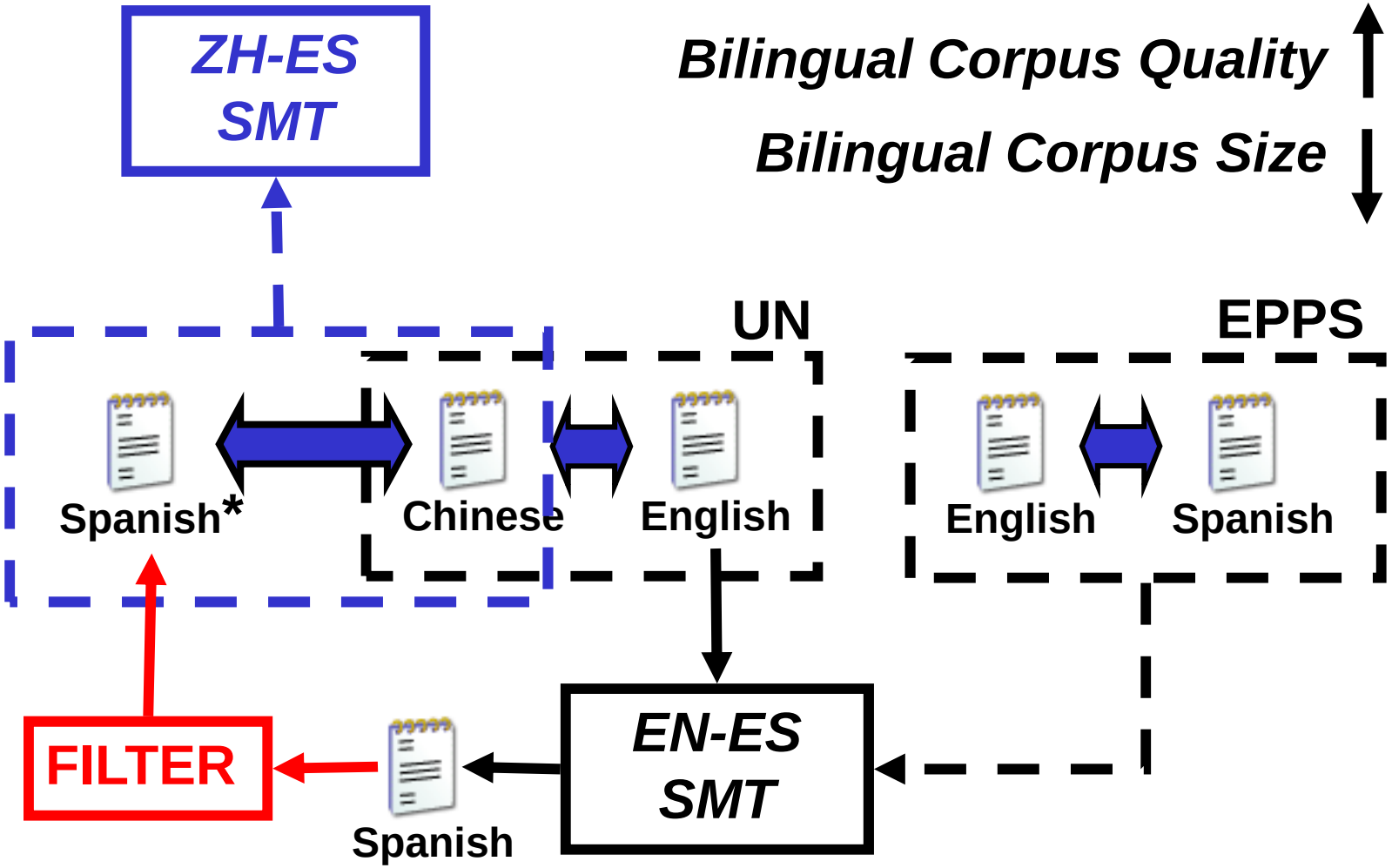
	Strategy	BLEU	NIST	WER	PER
Direction ZH ⇨ ES	Direct	0.1087	4.157	83.81	62.14
	Indirect	0.1145	4.413	78.04	58.21
Direction ES ⇨ ZH	Direct	0.0391	3.946	76.16	58.54
	Indirect	0.0397	3.378	75.62	59.22

Bilingual Corpus Filtering

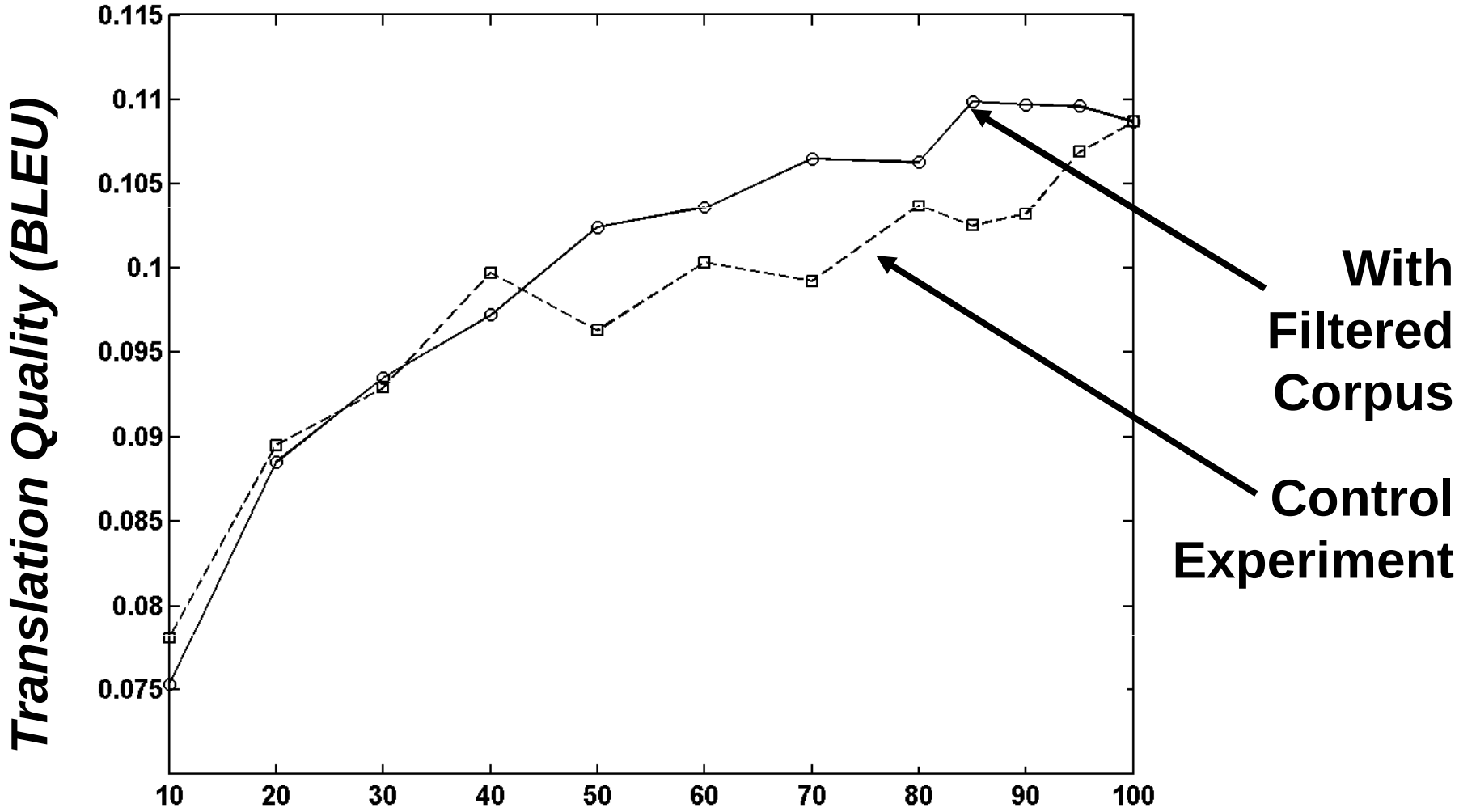
A Spanish language model, based on word 3-grams, was used for filtering the artificially constructed corpus



Filtered Translation System



Translation Quality Vs. Filtering



Comparison with a Rule-based System

System	BLEU	NIST	WER	PER
Statistic	0.1336	4.3101	57.99	79.73
Rule Based*	0.0697	2.8355	74.15	93.71

Note: it is known that automatic evaluation metrics tends to favor statistical systems over rule-based systems, so a human evaluation is required in order to provide a fair comparison.

***** http://www.worldlingo.com/en/products_services/worldlingo_translator.html

Conclusions

• The artificial generation of a Chinese-Spanish parallel corpus does not provide better results than performing indirect translation by using English as a “bridge language”.

• No significant translation quality improvement is achieved by using the proposed corpus filtering technique.

• The obtained Chinese-Spanish translation scores are still to low; so, further work is required to achieve state of the art translation quality.

Further work

• More efficient corpus filtering and/or editing techniques must be designed and evaluated.

• A word reordering feature must be included into the SMT system under consideration in order to significantly improve translation quality.

• Additional alternatives for Chinese-Spanish parallel corpus construction and/or compilation must be considered.

A Feasibility Study for Chinese-Spanish Statistical Machine Translation

Rafael E. Banchs

Josep M. Crego, Patrik Lambert, José B. Mariño

Univesitat Politècnica de Catalunya, Barcelona, Spain