MATMT-2008 – Donostia

Exploring Spanish-morphology effects on Chinese-Spanish SMT

Rafael E. Banchs *

Barcelona Media Innovation Center

Haizhou Li

Institute for Infocomm Research

Motivation

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

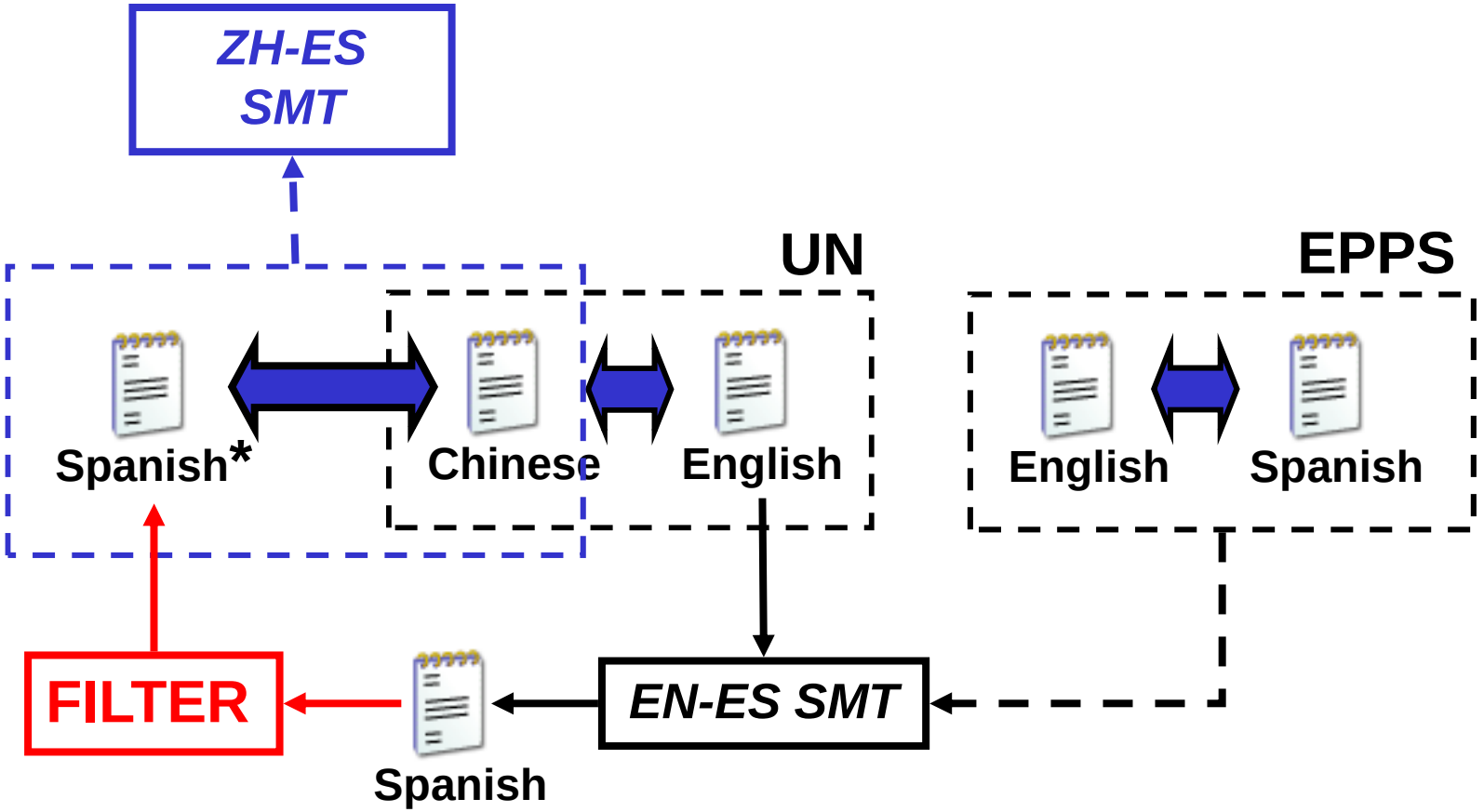
Particularities of the task

Main problem: no Chinese-Spanish parallel corpus available for training purposes *(at least as a free resource)*

Additional problem: Chinese and Spanish are very distant languages with strong differences in morphological variations and word ordering

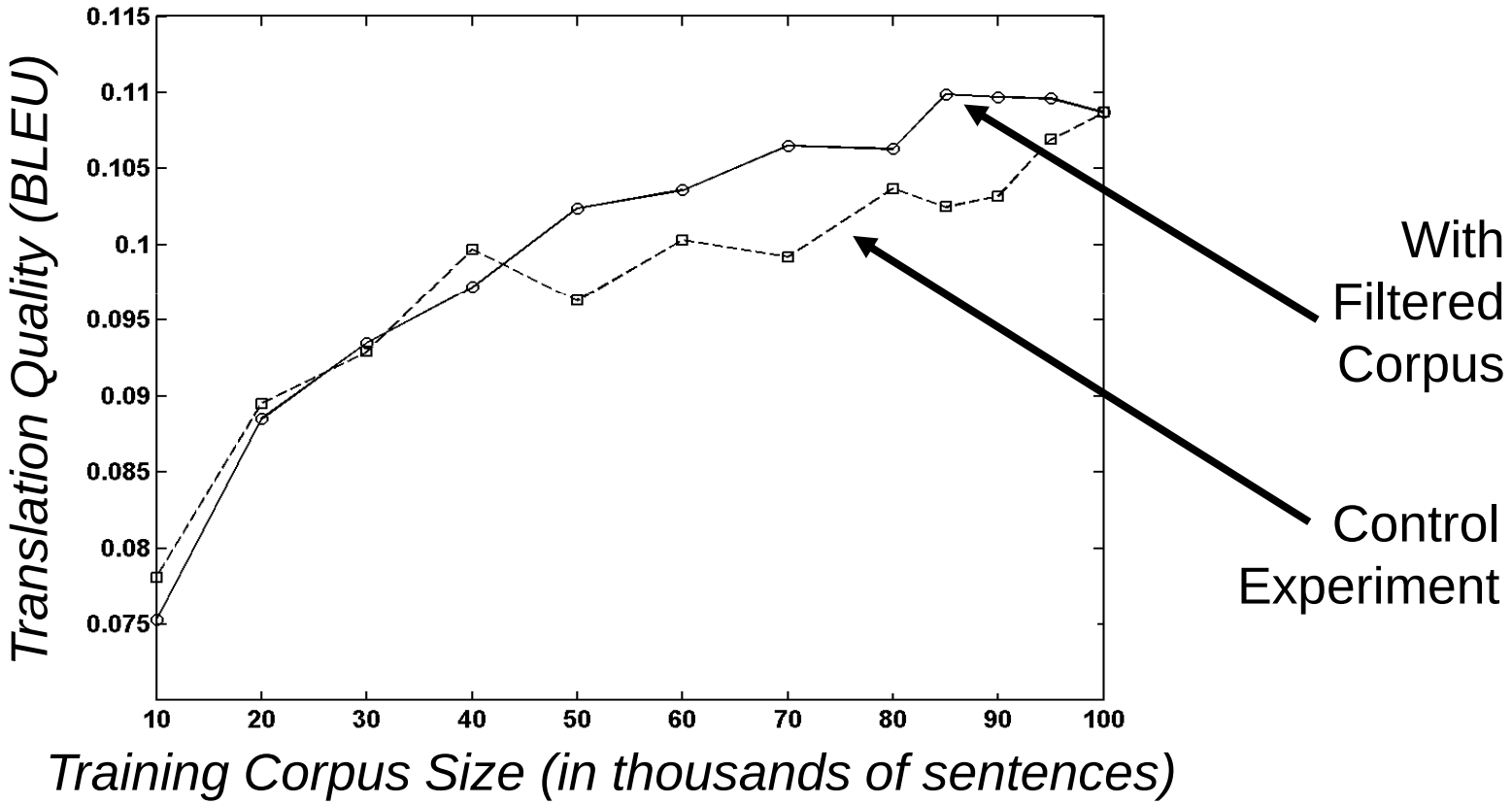
Previous Work

Artificial corpus construction



Previous Work

Artificial corpus construction



Some initial conclusions

Proposed strategy does not lead to a better translation system than the naive approach of cascading a Chinese-English system with an English-Spanish system

What are the alternatives?

1.- to combine the translation tables in order to make the cascade system independent of the English LM

2.- to collect the Chinese-Spanish corpus (news, embassy, ?)

3.- to find a Chinese-Spanish parallel corpus…

While looking for a corpus

The Bible corpus

- Chinese, Spanish and English versions are freely available
- About 31,000 sentences and 800,000 words per language
- Already aligned at the sentence level, although some manual edition was required
- Restricted to a specific domain, old language style, too many proper nouns, and relatively small

Corpus preprocessing

Tokenization: standard tokenization of Spanish and English and word segmentation of Chinese with ICTCLAS *(Zhang,2003)*

Sentence length restriction: a maximum of 80 words per sentence were allowed

Fertility filtering: a maximum ratio of 1:9 words per sentence pair was allowed

Format conversion: from GB2312 (Chinese) and ISO88591 (Spanish) to UTF-8

Corpus statistics

		Sentences	Tokens	Vocabulary	OOVs
train	**English**	28,887	848,776	13,216	–
	Chinese	28,887	760,451	12,670	–
	Spanish	28,887	781,113	**28,178**	–
dev	**English**	1,033	30,199	3,234	170 (5,3%)
	Chinese	1,033	27,235	3,404	127 (3,7%)
	Spanish	1,033	27,862	4,634	433 (9,3%)
test	**English**	1,035	30,008	3,158	132 (4,2%)
	Chinese	1,035	26,794	3,396	143 (4,2%)
	Spanish	1,035	27,368	4,652	415 (8,9%)

SMT system description

Alignment: GIZA++ *(Och & Ney, 2003)*

Translation units: maximum phrase length of 4 tokens, different alignment sets were considered

Decoding: Pharaoh *(Koehn, 2004)*, only four models used: relative frequency in the source-to-target direction, target language model, distortion model and word penalty

Language model: SRILM *(Stolcke, 2002)*, 4-grams of words

Optimization: bleu maximization by using Simplex algorithm

Baseline experiment results

BLEU scores were computed for all six translation tasks and translation tables extracted from four different alignment sets

	Source2target	Intersection	Union	Symmetrized
Chinese-to-Spanish	14.1	**14.3**	12.9	13.8
Spanish-to-Chinese	16.7	17.4	15.2	**17.5**
Chinese-to-English	18.6	19.2	17.2	**19.6**
English-to-Chinese	**20.2**	20.1	19.3	19.7
English-to-Spanish	30.6	**31.5**	30.6	30.5
Spanish-to-English	34.4	34.2	34.3	**34.5**

Important observations

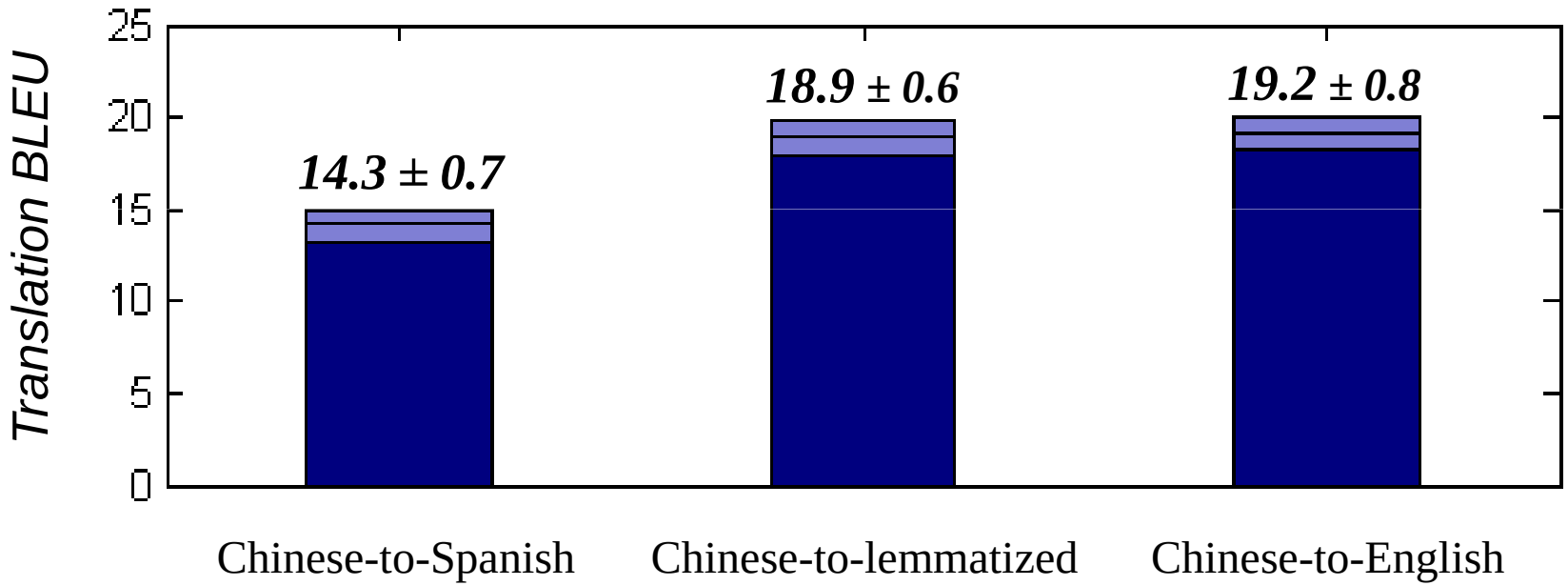
- The lowest translation qualities are obtained for the Chinese-Spanish language pair, while the highest ones are obtained for the English-Spanish language pair
- Extracting translation units from intersection seems to be the best strategy when Spanish is the target language
- Having Spanish as the target language seems to be adding a significant degree of complexity, which is most probably due to its high morphological variations…

Spanish morphology reduction

The morphological analyzer FreeLing *(Carreras et al., 2002)* was used to reduce Spanish full forms to lemmas *(this processing implied: lowercasing and retokenization of original corpus)*

		Sentences	Tokens	Vocabulary	OOVs
train	**Full forms**	28,887	781,113	**28,178**	–
	Lemmatized	28,887	784,398	**14,229**	–
dev	**Full forms**	1,033	27,862	4,634	433 (9,3%)
	Lemmatized	1,033	27,986	2,882	186 (6,5%)
test	**Full forms**	1,035	27,368	4,652	415 (8,9%)
	Lemmatized	1,035	27,452	2,864	181 (6,3%)

Effects of lemmatization on translation quality

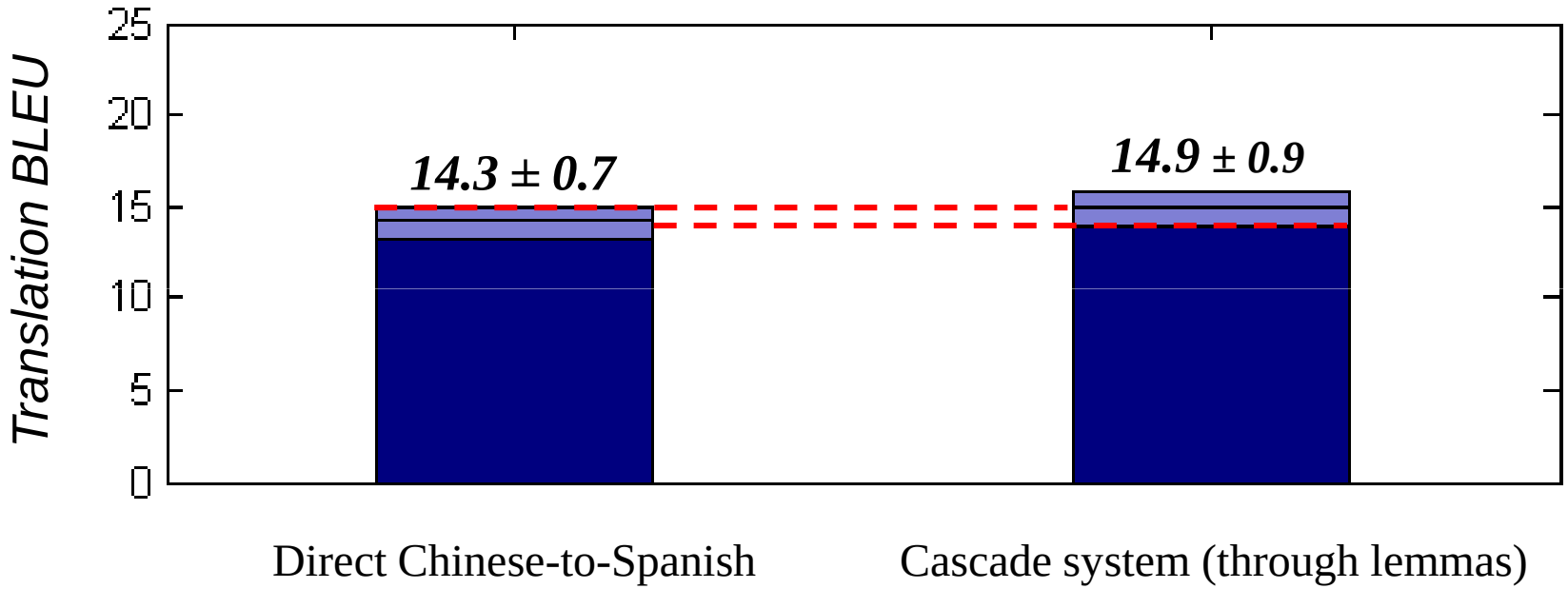


Recovering Spanish morphology with SMT

As a naive approach to recover Spanish morphology, we trained a lemmatized-Spanish to full-form-Spanish translation system

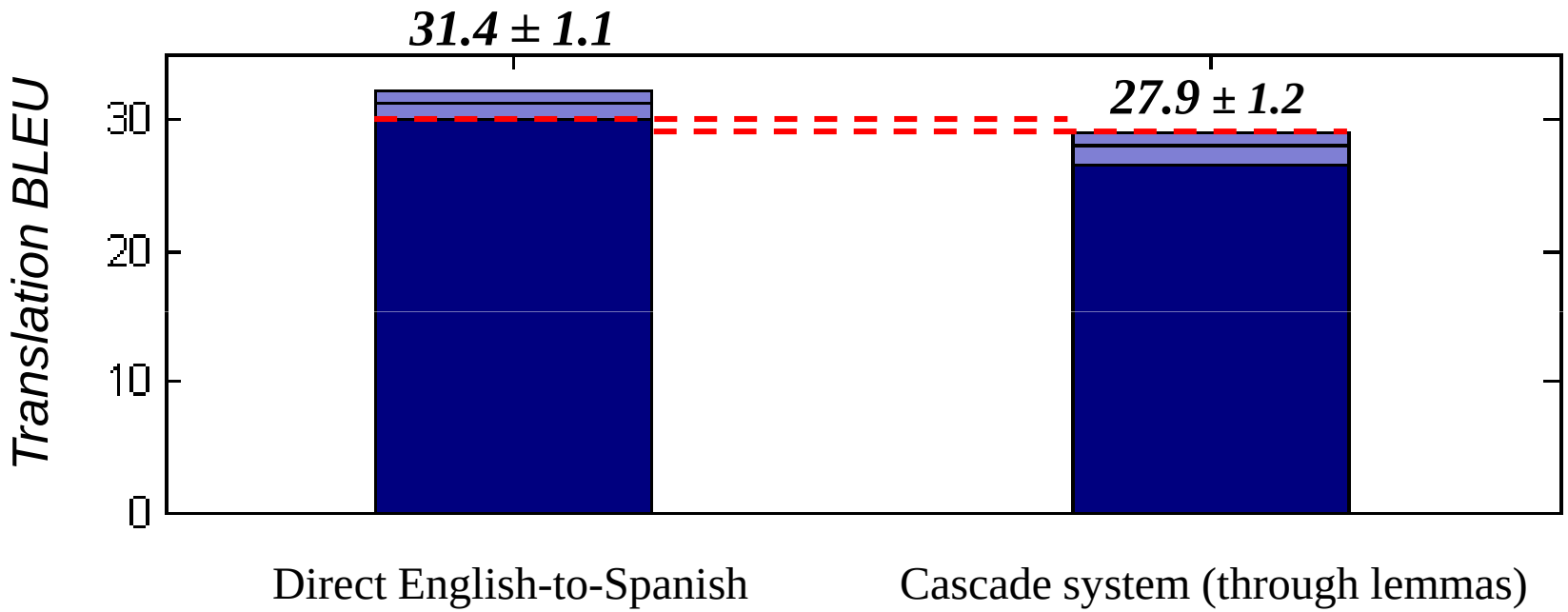
BLEU = 67.3 ± 1.1

Cascading through lemmas vs. direct translation



Does this actually make sense?

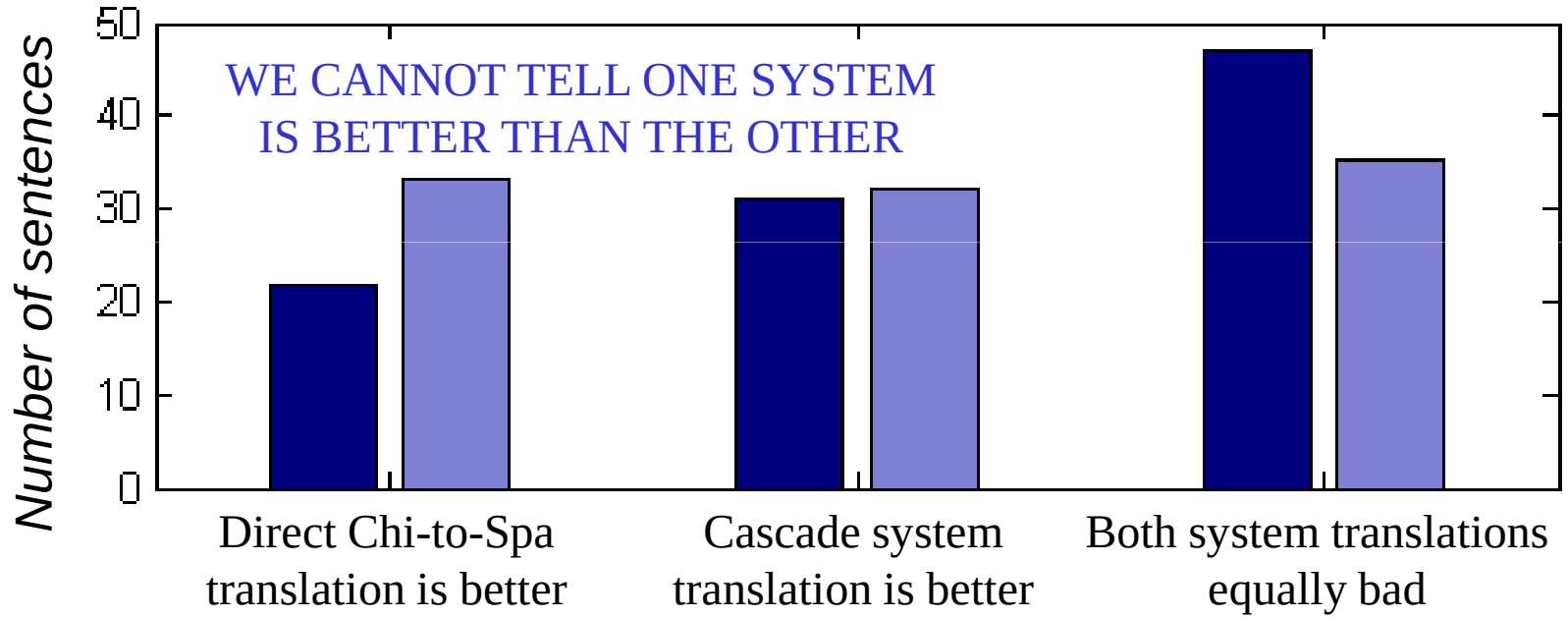
Comparison with the English-to-Spanish task



This is more like what one would actually expect !!!

Some human evaluations and oracle

100 randomly selected outputs were compared by two judges

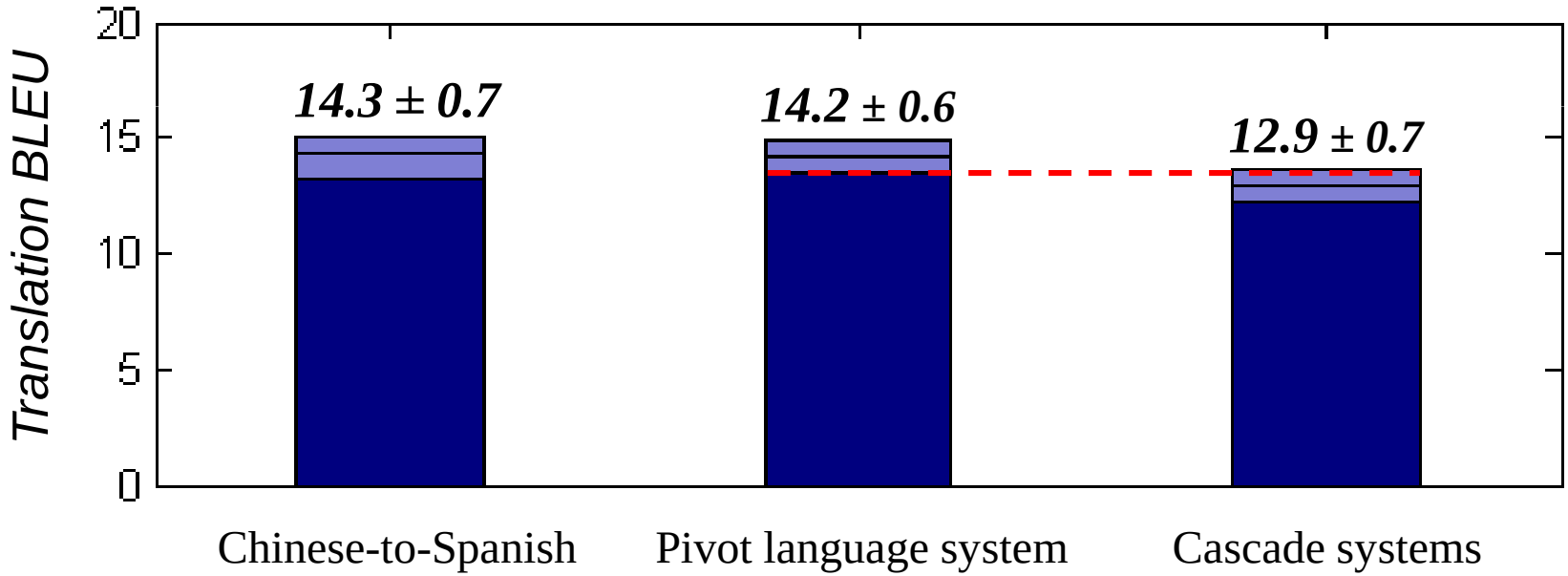


An oracle was computed with both system outputs: $BLEU = 16.7 \pm 0.8$

Some recent results

Translation table combination *(Wu & Wang, 2007)* by using English as a pivot language

$$p(t_i|s_i) = \sum_k p(t_i|p_k)\, p(p_k|s_i)$$



Conclusions and future work

There are not any interesting conclusions at this moment !!!

but future research should focus on:

1.- understanding better the relationships among models involving Chinese and lemmatized-Spanish and full-form-Spanish

2.- exploring methods for combining both systems

Conclusions and future work

3.- extrapolating the translation table combination method to larger corpora

4.- incorporating IBM-1 lexical features into the scenario

5.- exploring techniques for parallel corpus extraction from comparable corpora

6.- continuing our search for a free Chinese-Spanish parallel corpus…

MATMT-2008 – Donostia

Exploring Spanish-morphology effects on Chinese-Spanish SMT

***Rafael E. Banchs* ***

Barcelona Media Innovation Center

Haizhou Li

Institute for Infocomm Research