LIG/GETALP–MICA

Improved Vietnamese-French Parallel Corpus Mining Using English Language

Do Thi Ngoc Diep[1,2], Laurent Besacier[1], Eric Castelli[2]

(1) LIG Laboratory, CNRS/UMR-5217, Grenoble, France
(2) MICA Center, CNRS/UMI-2954, Hanoi, Vietnam
Laurent.Besacier@imag.fr

Abstract

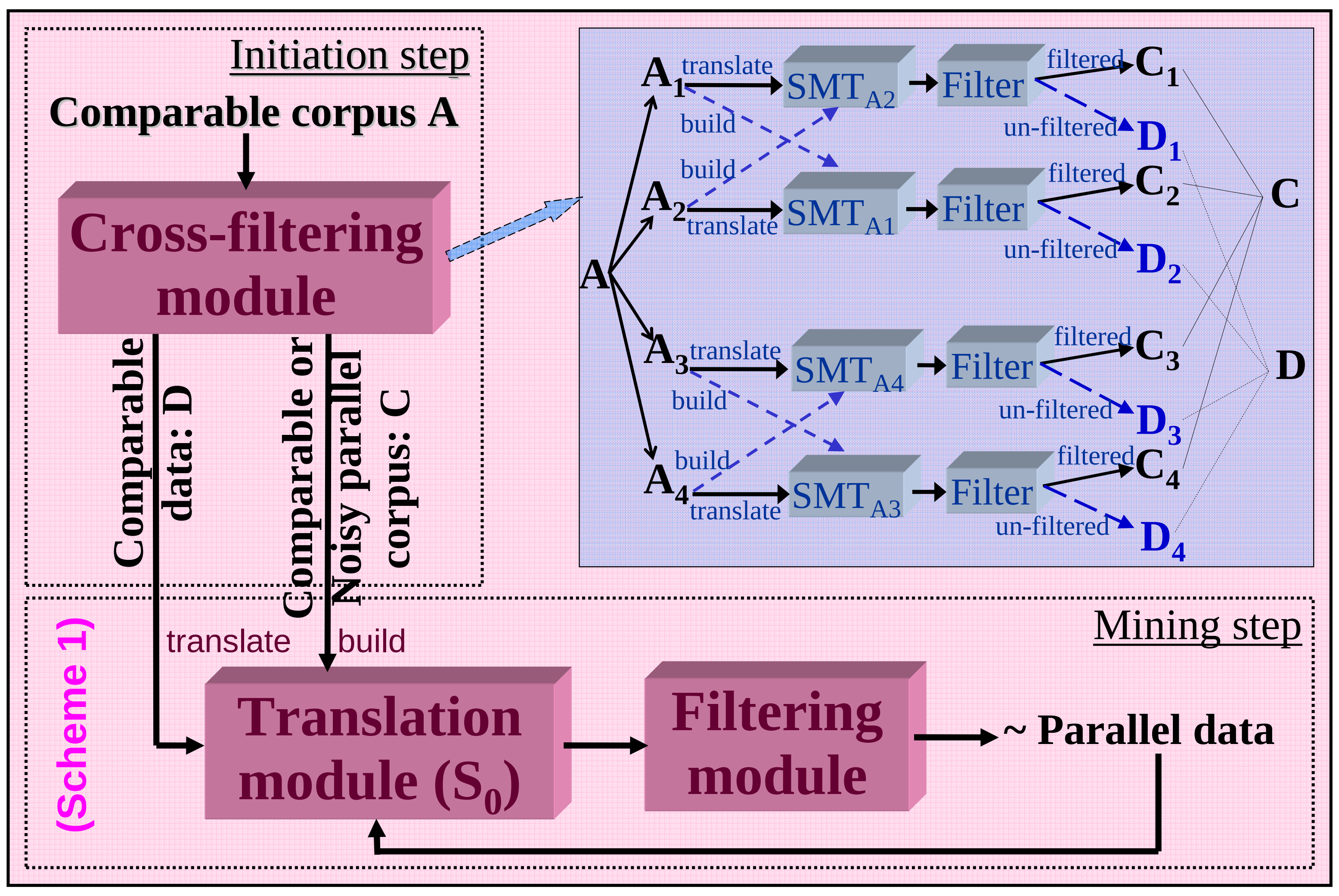
Extracting parallel sentence pairs from a comparable corpus using:

❖ Un-supervised method [1]: start with a comparable corpus, overcome the problem of lacking parallel data.

❖ Triangulation as an extension of the unsupervised method

[1] Do, T.N.D, L. Besacier, E. Castelli, “A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora”, European Association for Machine Translation 2010

A Fully Unsupervised Method to Mine Parallel Data from Noisy Parallel Corpora



+ Translation module: statistical machine translation system.

+ Filtering module: A pair is considered as parallel if its PER* metric > a threshold.

$$PER^* = \frac{2 * \text{number of identical words}}{(\text{length of hypothesis} + \text{length of reference})}$$

Other metrics like TER, BLEU, NIST were investigated but PER* achieved the best performance [1]

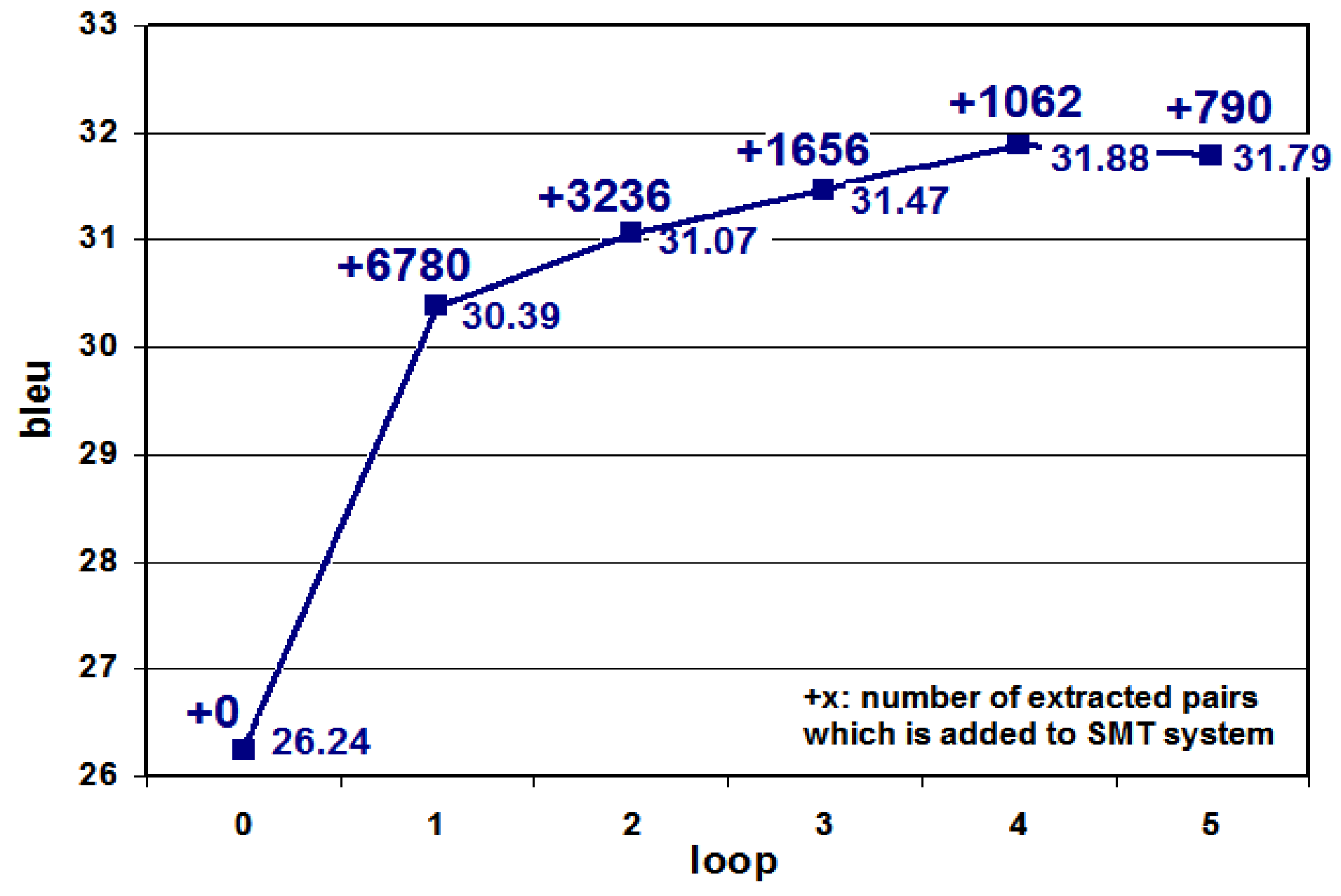
Experiments – Scheme 1

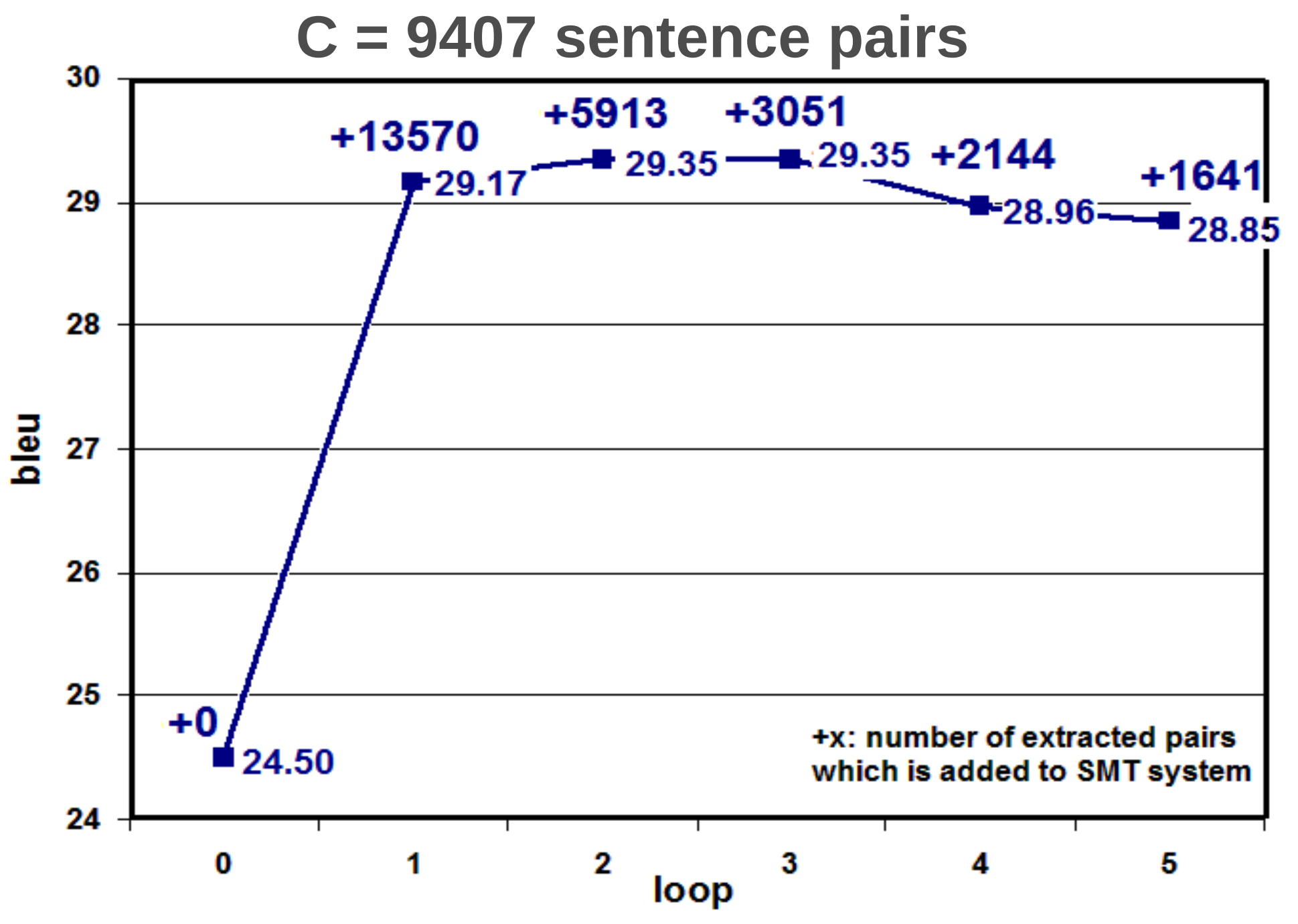
Vietnamese – French comparable corpus

After cross-filtering:
D = 341,178 sentence pairs;
C = 4076 sentence pairs

Vietnamese – English comparable corpus

After cross-filtering:
D = 470,458 sentence pairs;
C = 9407 sentence pairs





BLEU score of SMT systems after each loop

+The text corpus used: a multilingual daily news website of 4 languages (Vietnamese, English, French, and Spanish) - The Vietnam News Agency (http://www.vietnamplus.vn)

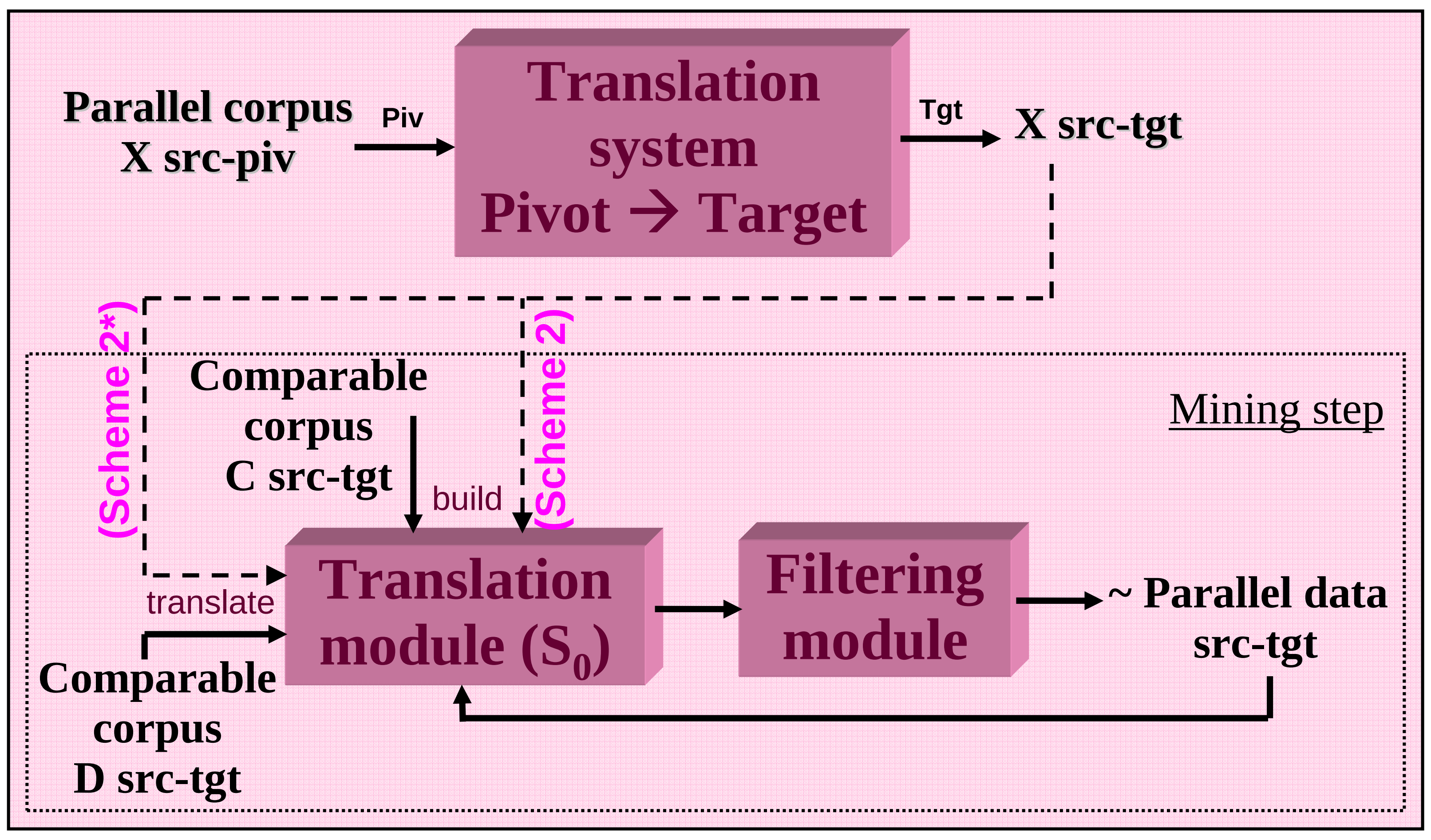
+Test set: 400 manually extracted parallel sentence pairs

❖ Each iteration brings us a number of extracted sentence pairs.

❖ The quality of the translation system increases in the first few iterations and decreases after that. (in the first iterations, a lot of new parallel sentence pairs are extracted and included to the translation model. However, in subsequent iterations, as the amount of truly parallel sentences decreases, more wrong sentence pairs are added to the system so the quality of the translation system is reduced)

❖ However, the quality of the translation system built by extracted data from this unsupervised method is comparable with that of another method which requires better quality data for bootstrapping (bilingual dictionary, etc.) (see more in [1])

Using Triangulation through English Extension of the Unsupervised Mining Method



Experiments – Scheme 2, 2*

SRC: Vietnamese, TGT: French; PIV: English

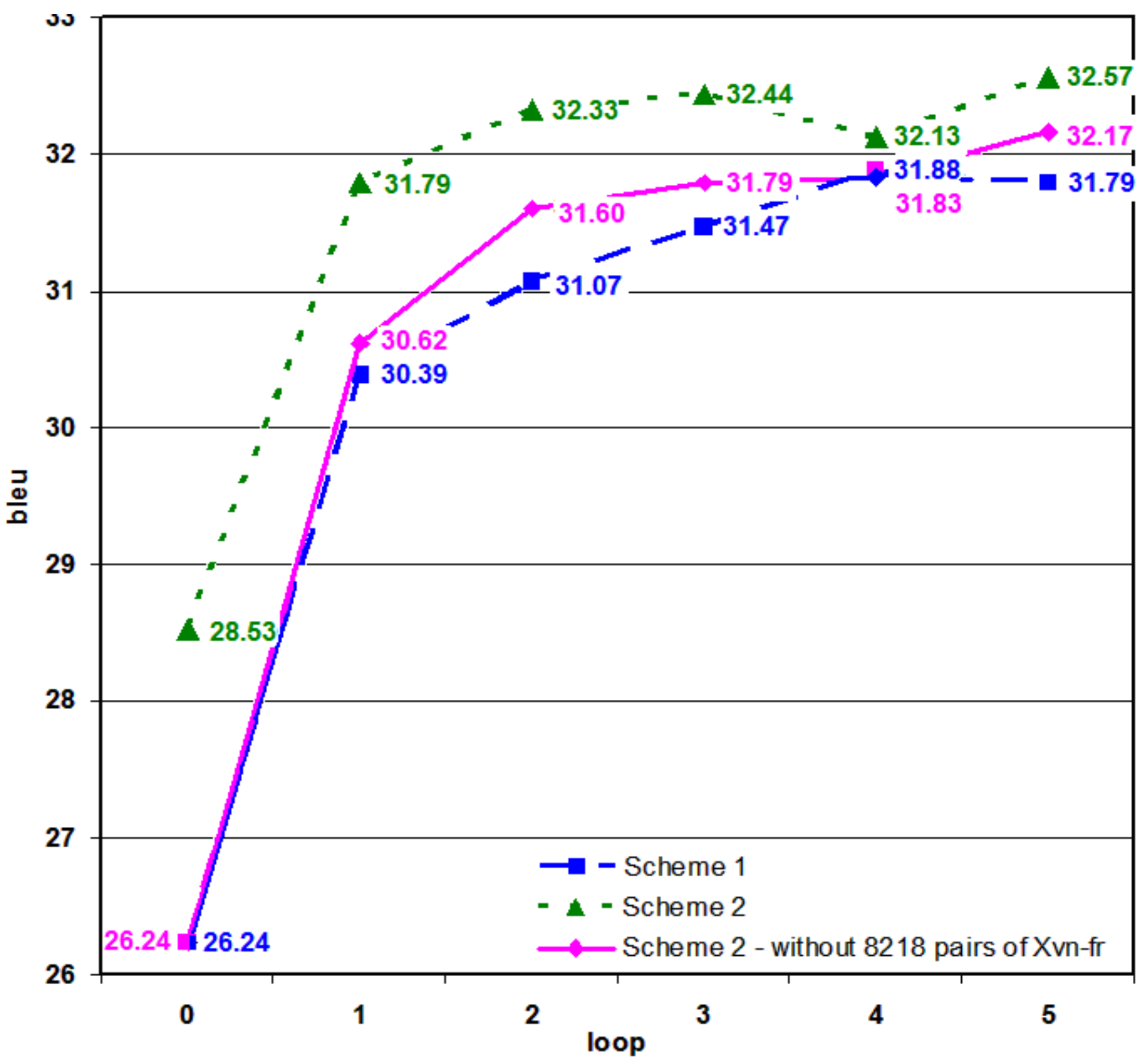
❖ **SMTEnglish – French**: build from the Europarl and News corpora (WMT, IWSLT2010). BLEU on test set WMT 2009 = 23.74)

❖ **Vietnamese - English data**:
+ Scheme 1, PER*=0.4 → Xvn-en=8218 sentence pairs.

❖ **Vietnamese - French data**:
+ C vn-fr = 4076 sentence pairs
+ D vn-fr = 341,178 sentence pairs.

❖ Test set: 400 manually extracted parallel sentence pairs

Scheme 2 (S0: C=4076+ X=8218) and Scheme 1 (S0: C=4076)

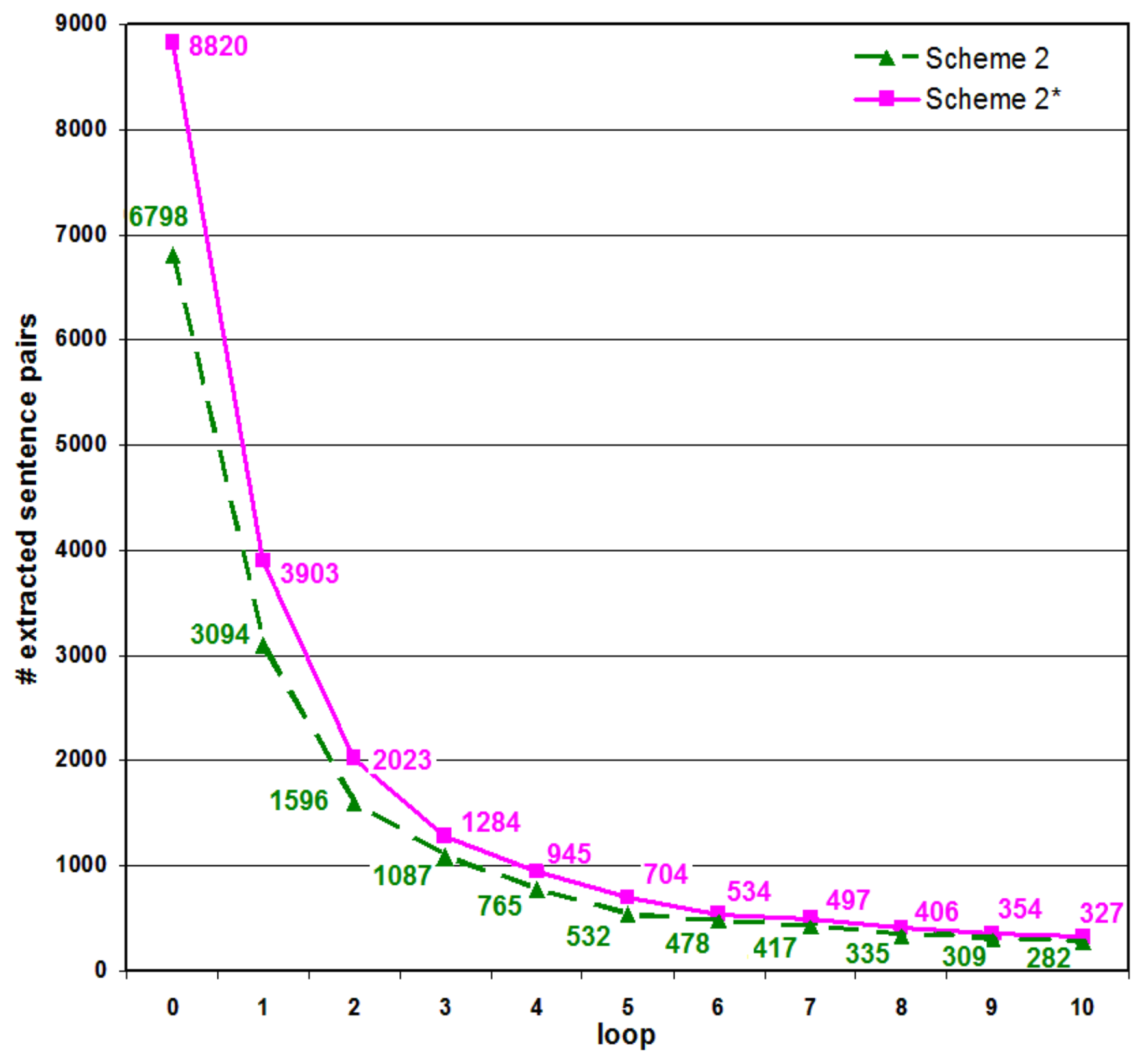


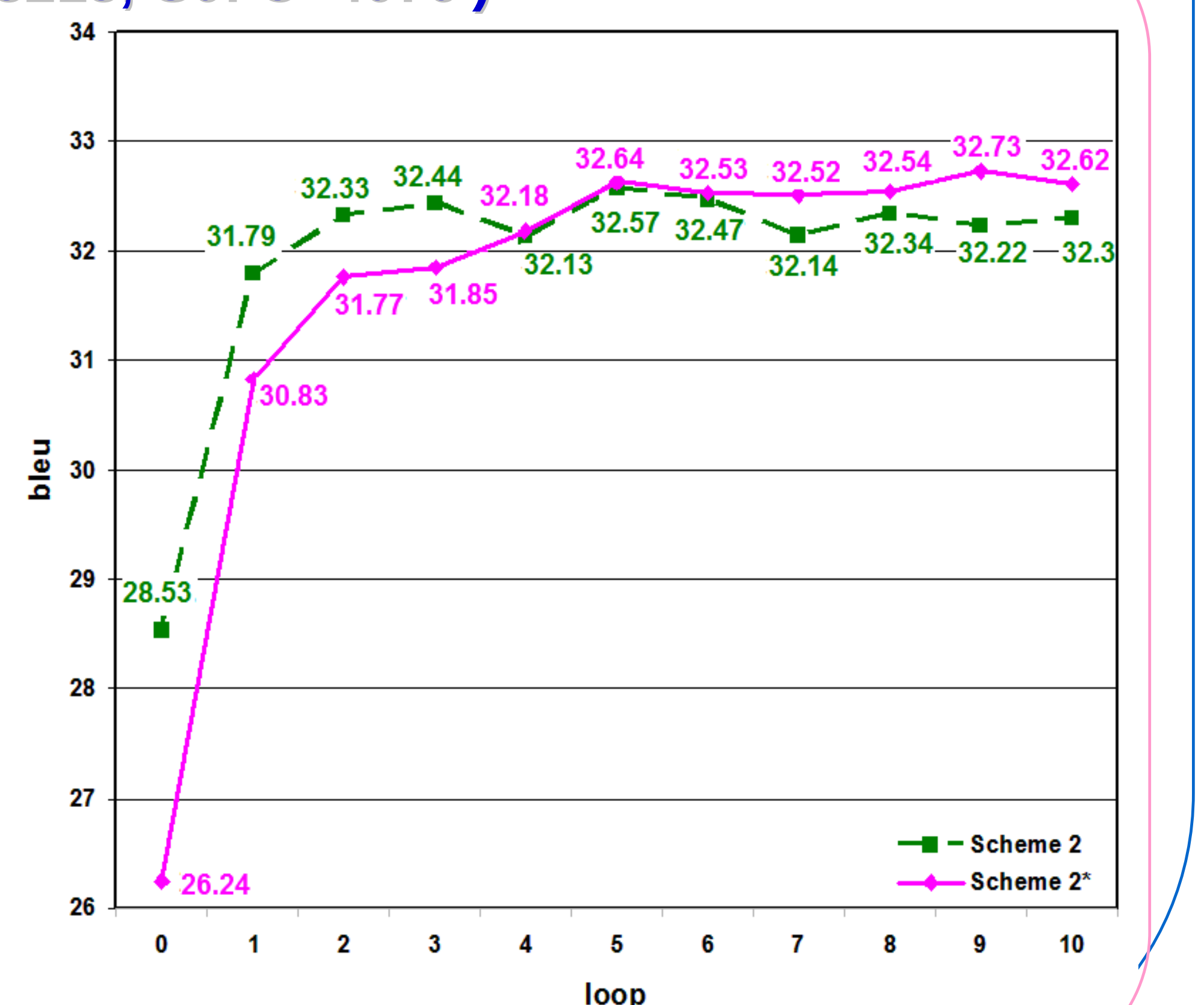
Evaluation of the SMT systems through loops

Loop	Sch 1	Sch 2	Loop	Sch 1	Sch 2
0	6780	6798	6	460	478
1	3236	3094	7	409	417
2	1656	1596	8	392	335
3	1062	1087	9	324	309
4	790	765	10	239	282
5	576	532			

The number of extracted data through loops

Scheme 2 (D: 341,178; S0: C=4076 + X=8218) and Scheme 2* (D: 341,178 + X=8218; S0: C=4076)





Extraction result and Evaluation score of the SMT systems through loops

International Workshop on Spoken Language Translation

Paris - France | 2-3 December 2010