A fully Unsupervised approach for mining parallel data from comparable corpora

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

[1]LIG Laboratory, CNRS/UMR-5217, Grenoble, France
[2]MICA Center, CNRS/UMI2954, HUT, Hanoi, Vietnam
thi-ngoc-diep.do@imag.fr

A parallel bilingual corpus

- Statistical machine translation (SMT): a large parallel bilingual text corpus.
- To build parallel corpora:
  - Collect from parallel document pairs (Resnik and Smith, 2003; Kilgarriff and Grefenstette, 2003)
  - Apply alignment methods at document level, sentence level for the source and target monolingual corpora (Koehn, 2005; Gale and Church, 1993, Patry and Langlais, 2005)
  - Mine a comparable corpus (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2006)
  - etc.

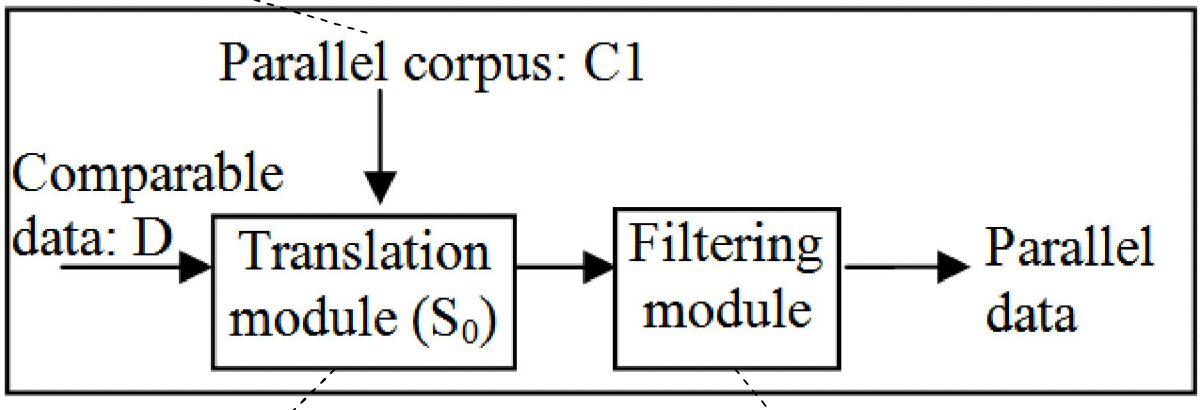
Mining a comparable corpus

- A comparable corpus:
  - "closely related by conveying the same information" (Zhao and Vogel, 2002)
  - "mostly bilingual translations of the same document" (Fung and Cheung, 2004)
  - "various levels of parallelism, such as words, phrases, clauses, sentences, and discourses…" (Kumano et al., 2007).
- Source: News domain
- "comparable" $\xrightarrow{IR}$ "noisy parallel"

Advanced IR approaches are outside of the scope of this paper

Mining a comparable corpus

- bilingual dictionary
- human translation pairs
- parallel corpus



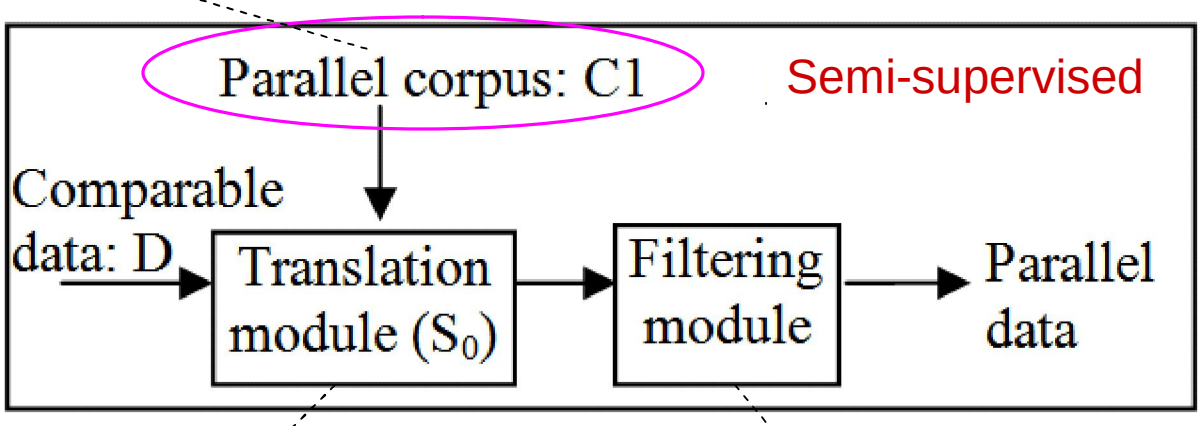
- a translation lexicon model
- a proper statistical machine translation system

- maximum likelihood criterion
- evaluation metric

Ref: Zhao and Vogel (2002), Munteanu and Marcu (2006), Abdul-Rauf and Schwenk (2009), Sarikaya et al. (2009)

Mining a comparable corpus

- bilingual dictionary
- human translation pairs
- parallel corpus



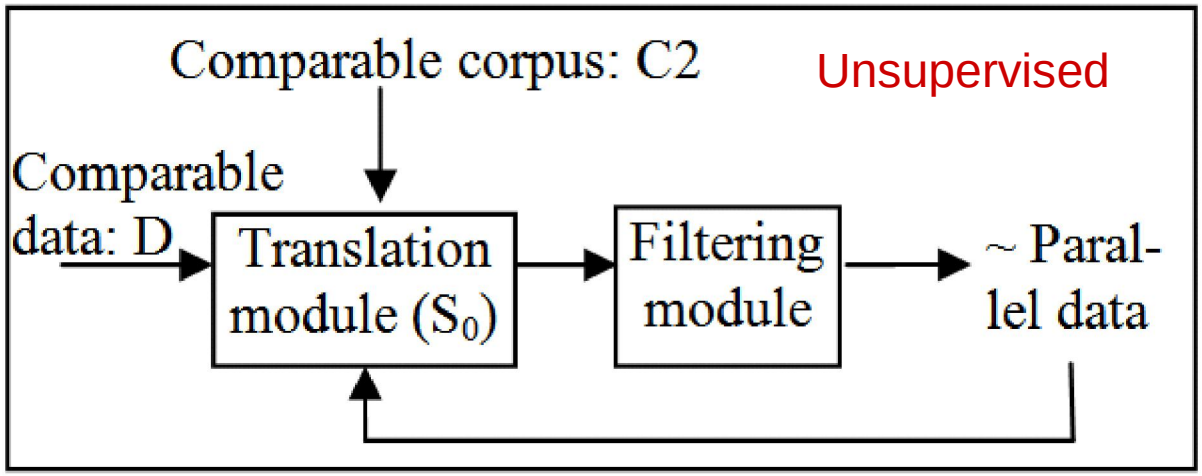
- a translation lexicon model
- a proper statistical machine translation system

- maximum likelihood criterion
- evaluation metric

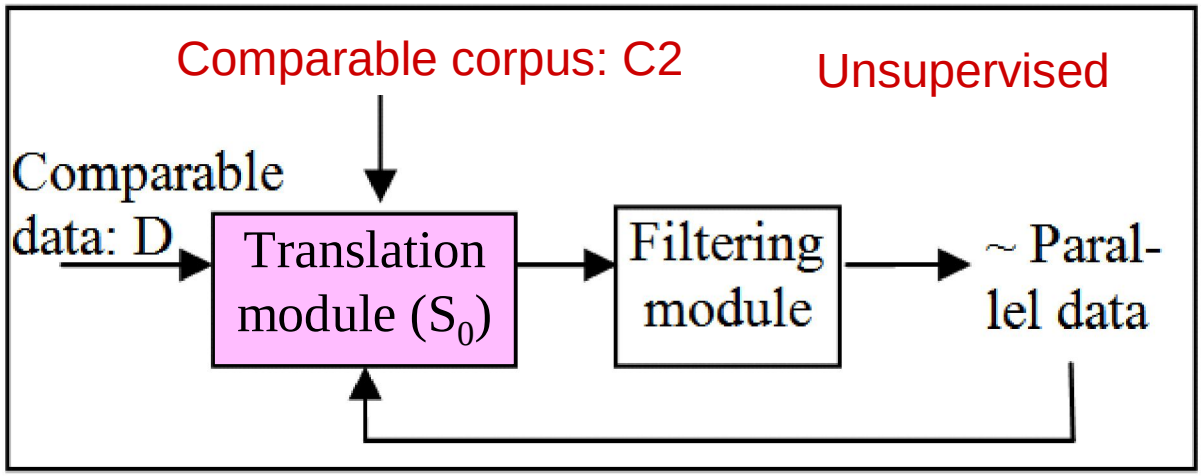
Ref: Zhao and Vogel (2002), Munteanu and Marcu (2006), Abdul-Rauf and Schwenk (2009), Sarikaya et al. (2009)

=> Does a fully unsupervised method, starting with a comparable corpus, allow us to overcome the problem of lacking parallel data?

Our unsupervised learning method

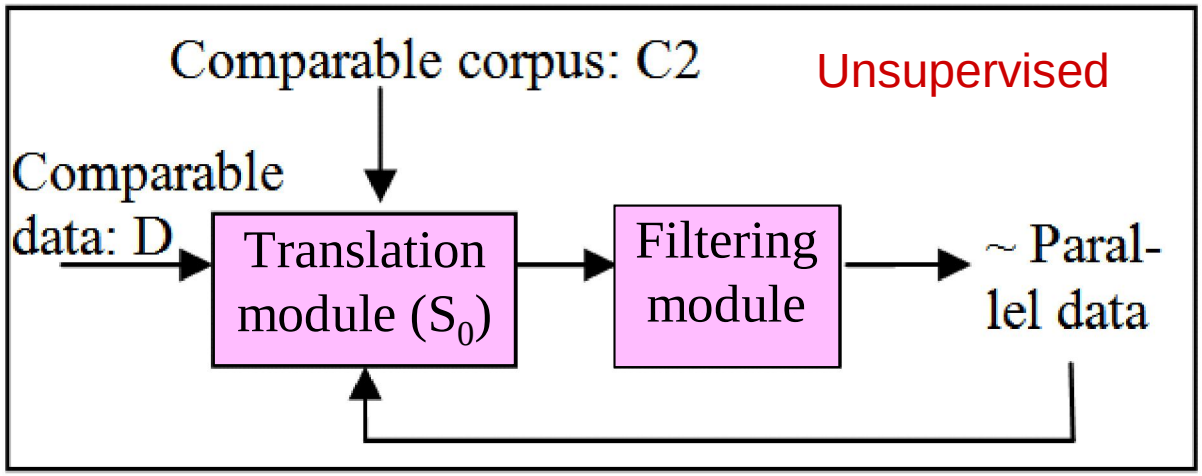


Our unsupervised learning method



- Translation module
  - A statistical machine translation system
  - Start with a simple noisy comparable corpus (named C2), without using additional parallel data

Our unsupervised learning method

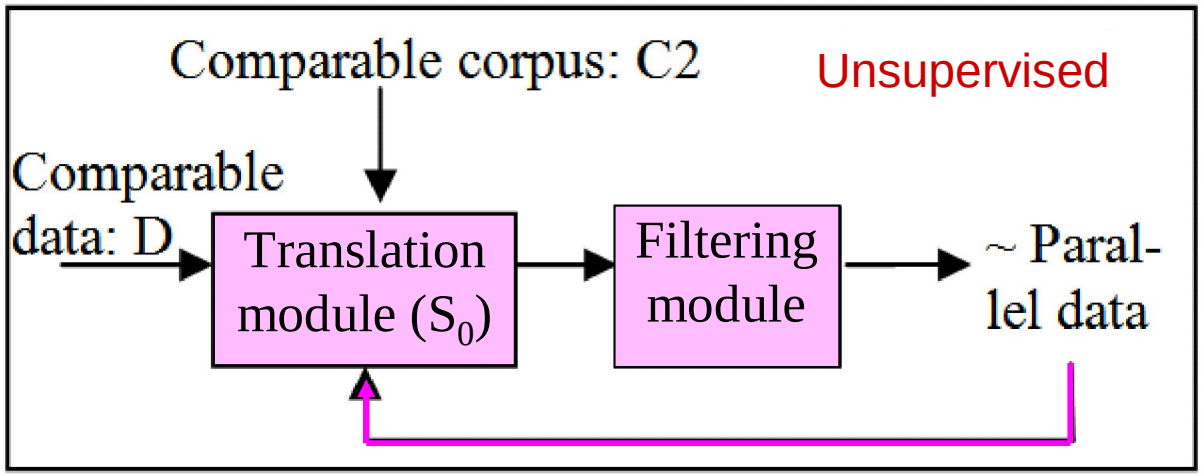


- Filtering module:
  - Use evaluation metric estimated for each sentence pair
  - Which one ? Bleu, Nist, Ter, Per* (based on the similarity of two sentences)

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

  - A pair is parallel if score > threshold (for Bleu, Nist, Per*) or < threshold (for Ter)

Our unsupervised learning method



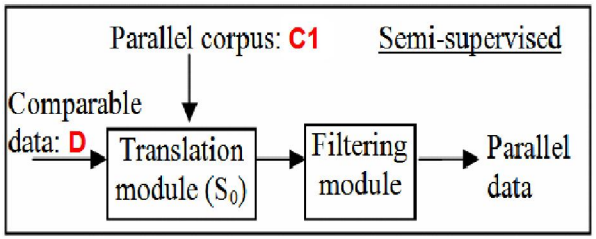
- Iterative scheme:
  - combine the extracted pairs with the translation module => new one
  - Re-translate D à re-calculate score à re-filter data à re-combine …
- Different combinations at iteration i :
  - W1: S_0 is retrained on C2 and E_{i-1}
  - W2: S_0 is retrained on C2 and $E_0+E_1+\ldots+E_{i-1}$
  - W3: E_{i-1} à a new separate phrase-table. Decode using phrase-table of S_0 and this new one (log-linear model) without weighting them.
  - W4: the same combination as W3, but the phrase-table of S_0 and the new one are weighted, e.g. 1:2.

Experiments for French-English SMT

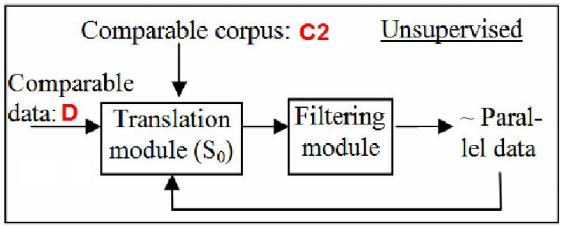
Compare the semi- and un- supervised methods

Data preparation

- Two systems were constructed (using the Moses toolkit (Koehn et al., 2007)) to mine a comparable corpus D:
  - semi-supervised method (Sys1)
  - unsupervised method (Sys2)
- Create "simulated" noisy parallel corpus:
  - C1: 50K parallel sentence pairs from the Europarl v.3
  - C2: 25K correct parallel sentence pairs (withdrawn from C1) and 25K wrong sentence pairs
  - D: 10K parallel sentence pairs from the Europarl v.3 (marked) and 10K wrong sentence pairs, which were different from sentence pairs of C1 and C2



Sys1

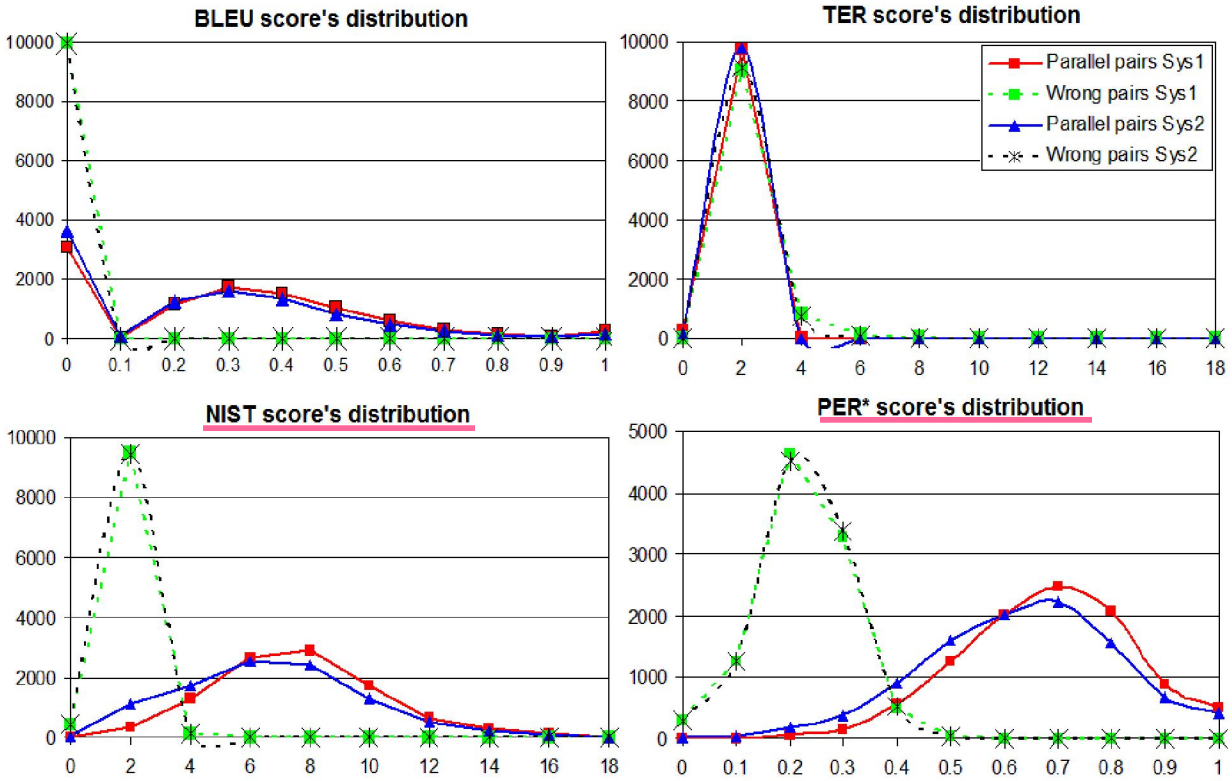


Sys2

Experiments

- Whether Sys2 can be used to filter the input data in the same fashion as Sys1 does?
  - Translate the French side of corpus D by Sys1 and Sys2
  - Calculate the scores BLEU, NIST, TER and PER* for the translated output with the English side of the corpus D
  - Display the distributions of evaluation scores for correct parallel sentence pairs and wrong sentence pairs

Score distributions

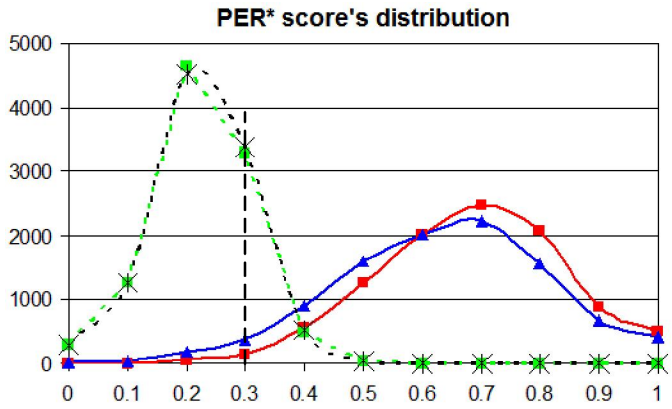


Sys1 – semi-supervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	**6908**	**6892**	**99.76**	**68.92**	**81.52**
Nist=0.4	**8350**	**8347**	**99.96**	**83.47**	**90.97**
Per*=0.3	**10342**	**9785**	**94.61**	**97.85**	**96.20**
Per*=0.4	**9390**	**9333**	**99.39**	**93.33**	**96.27**
Sys2 – unsupervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	**6233**	**6218**	**99.75**	**62.18**	**76.61**
Nist=0.4	**7110**	**7108**	**99.97**	**71.08**	**83.08**
Per*=0.3	**10110**	**9468**	**93.65**	**94.68**	**94.16**
Per*=0.4	**8682**	**8629**	**99.38**	**86.29**	**92.37**

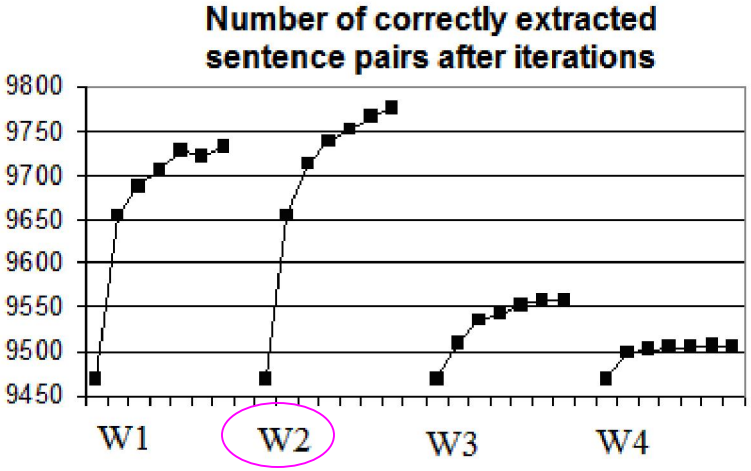
- The distributions of scores have the same shape between Sys1 and Sys2
  - In particular, the distributions of scores for the wrong pairs were nearly identical in both systems.
- PER* can be considered as the most suitable score

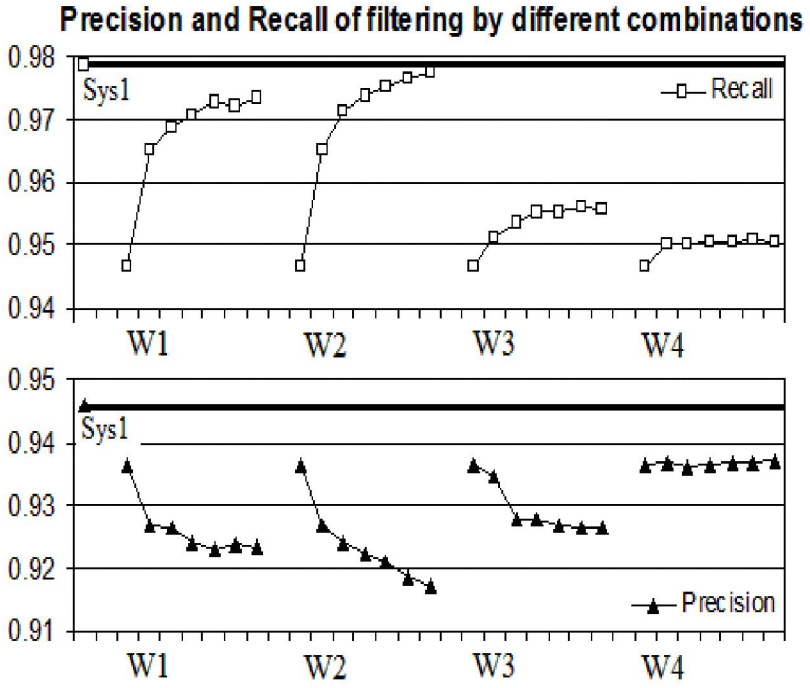
Iterations

- The iterations of the unsupervised method
  - improve the quality of the translation system
  - increase the number of correctly extracted sentence pairs
- Combined the extracted sentence pairs in 4 ways: w_1, w_2, w_3, w_4
- Chose the score PER* and the threshold=0.3



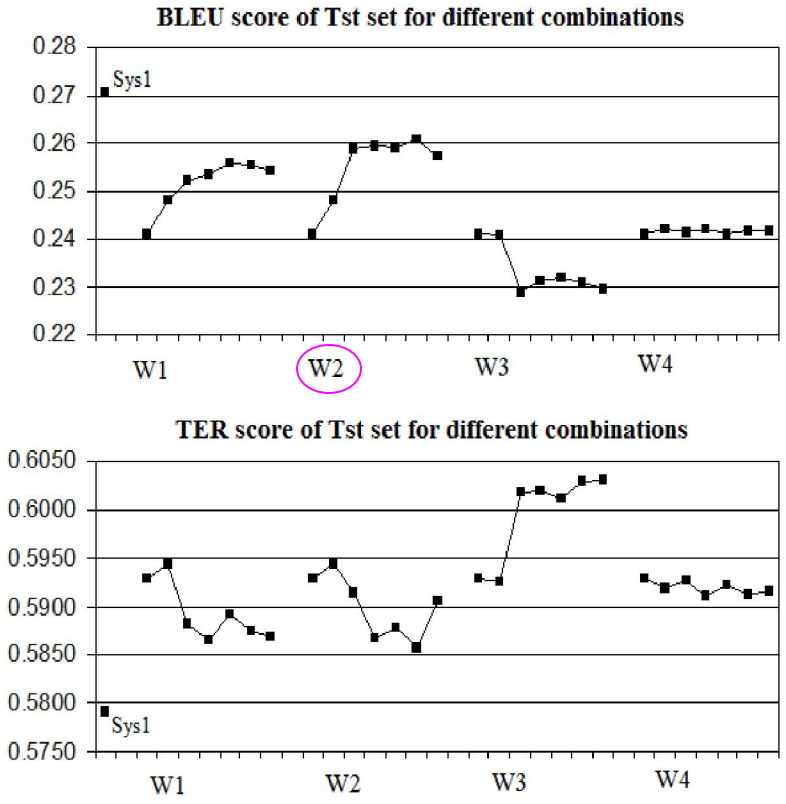
Iterations





- The number of correct extracted pairs was increased in all cases
- W2 brought the largest number of correct extracted sentence pairs.

Iterations



The quality of the translation systems

- A test set: 400 French-English parallel sentence pairs from Europarl corpus.
- Use one reference.
- The quality of the translation system was increased quickly during the first few iterations, but decreased after that.

APPLICATION FOR FRENCH-VIETNAMESE LANGUAGE PAIR

A truly comparable corpus

Preparing the data

- Vietnamese daily news website, the Vietnam News Agency[1] (VNA): tends to contain parallel sentences or rough translations of sentences on the same topics
  - 20,884 French documents
  - 54,406 Vietnamese documents
  - 10 sentences per document
  - 30 words per sentence

(from 12 April 2006 to 14 August 2008)

[1]. http://www.vnagency.com.vn/

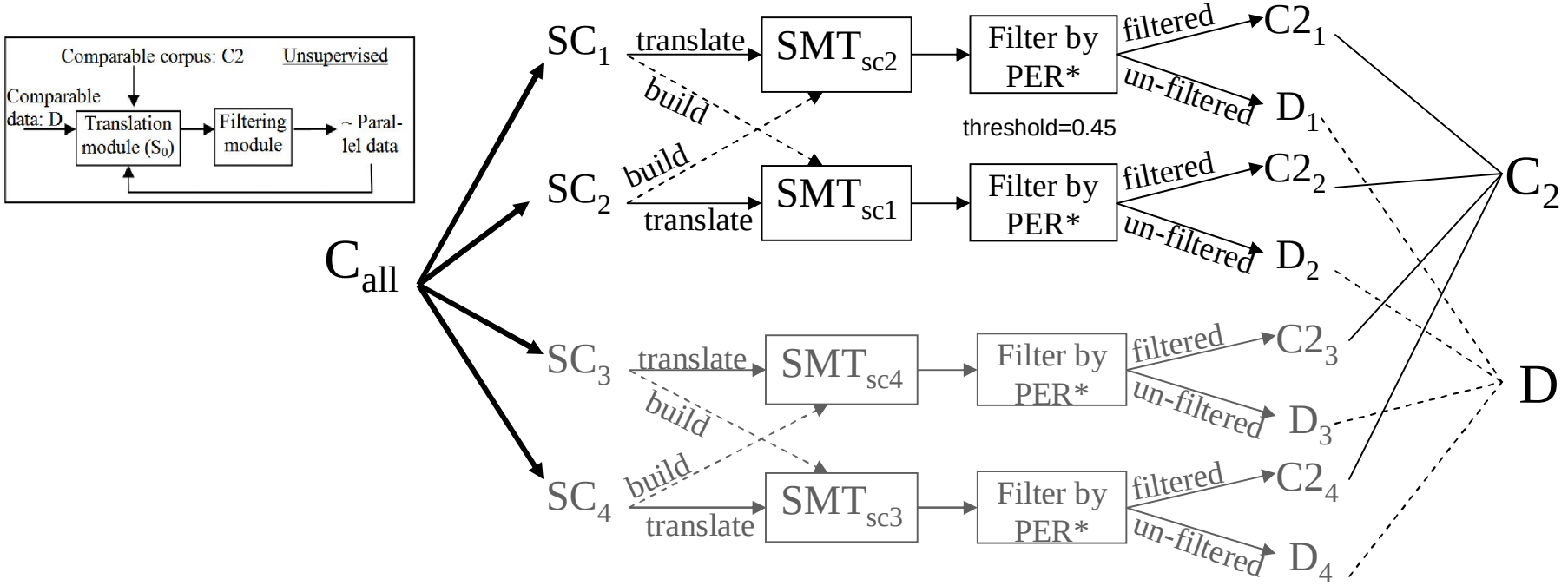
A noisy comparable corpus

- A noisy comparable corpus
  - Apply a publishing date filter
  - Merge sentence: a *m*-sentence Vietnamese document and a *n*-sentence French document => *m x n* pairs of sentences.
  - From VNA => 1,442,448 pairs of sentences: really noisy parallel
  - Filter by the ratio of the French sentence's length to the Vietnamese sentence's length = 0.8 ÷ 1.3

=> 345,575 pairs of sentences (named Call).

The initial translation system

- A cross-filtering process to extract C2 and D



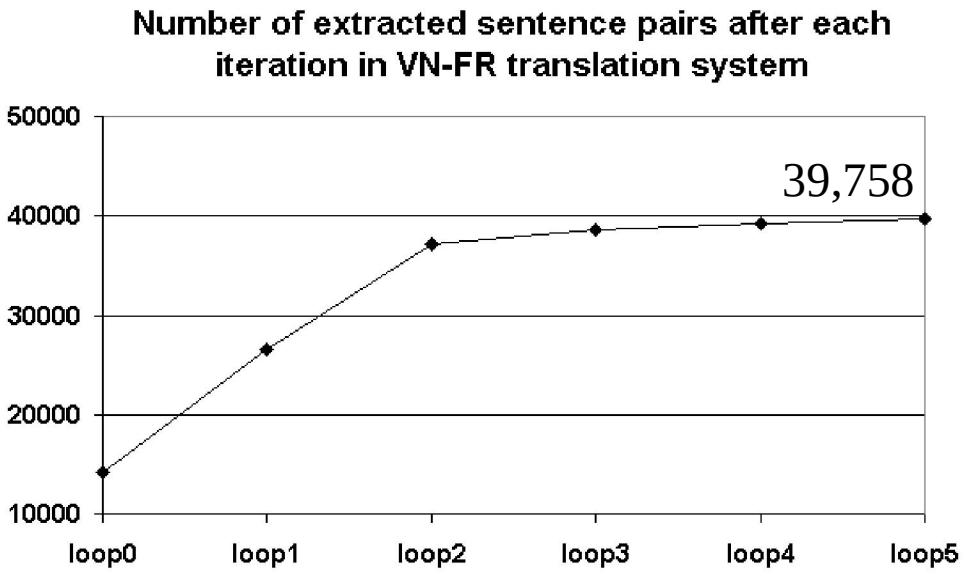
Sub corpus	# pairs	# $C2_i$	# D_i
SC_1	85,011	2916	82,095
SC_2	85,008	3495	81,513
SC_3	86,529	3820	82,709
SC_4	89,027	3892	85,135

à

C2: 14,123 pairs

D: 331,452 pairs

Applying the unsupervised method



SMT iter.	# extracted pairs	Bleu	Nist	Ter
0	14,123	30.67	6.45	0.59
1	26,517	32.18	6.70	0.57
2	37,210	32.42	6.75	0.56
3	38,530	32.45	6.77	0.55
4	39,254	32.14	6.73	0.56
5	39,758	31.85	6.68	0.56

Test set: 400 manually extracted Vietnamese-French parallel sentence pairs

- The number of extracted sentence pairs increased with each iteration
- The quality of the translation system was increased quickly during the first few iterations, but decreased after that.

The former method

- *Method1 (Do et al. 2009):*
  - Mining method:
    - Filter possible parallel document pairs by *publishing date and special words* (numbers, attached symbols, named entities).
    - *Align sentences* in a possible parallel document pair *using lexical information* (lexemes, stop words, a bilingual dictionary, etc.).
    - Extract sentence pairs based on the sentence alignment information, which combines *document length information and lexical information*
  - From VNA => extracted 50,322 “parallel” sentence pairs

Compare unsupervised method and *Method1*

Mining method	# extracted pairs	Bleu	Nist	Ter
Lexical info. + Heuristics (*Method1*)	**50,322**	**32.74**	**6.78**	**0.55**
Unsupervised method	**38,530**	**32.45**	**6.77**	**0.56**

The same test set of 400 manually extracted Vietnamese-French parallel sentence pairs

- The number of extracted sentence pairs is lower than that in the *Method1*
- The quality of the SMT systems are comparable

Conclusion and perspectives

- An unsupervised method for extracting parallel sentence pairs from a comparable corpus
  - based on a comparable corpus, instead of a parallel corpus
  - using iterative scheme
- The quality of the translation system
  - can be improved during the first iterations, but it becomes worse later because of adding the noisy data into the statistical models.
  - is comparable with that of another method which requires better quality data for bootstrapping (bilingual dictionary, etc.).
- This method may be applied successfully even in those cases where parallel data are lacking.

Conclusion and perspectives

- Our future works:
  - deeper analysis of the filtering and data inclusion techniques
  - experiments at a larger scale
  - human evaluations to confirm improvements obtained with our unsupervised method

Thank you !

References

- Abdul-Rauf, S. and H. Schwenk. 2009. On the use of comparable corpora to improve smt performance, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.
- Brown, P.F., S.A.D. Pietra, V.J.D. Pietra and R.L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. Vol. 19, no. 2.
- Doddington G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Human Language Technology Proceedings.*
- Fung, P., P Cheung. 2004. Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM. *Conference on Empirical Methods on Natural Language Processing.*
- Gale, W.A. and K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*.
- Ho, T.B. 2005. Current status of machine translation research in vietnam, towards asian wide multi language machine translation project. *Vietnamese Language and Speech Processing Workshop*.
- Hutchins, W.J. 2001. Machine translation over fifty years. *Histoire, epistemologie, langage.* ISSN 0750-8069.
- Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics, volume 29*.
- Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *Machine Translation Summit*.
- Koehn, P., F.J. Och and D. Marcu. 2003. Statistical phrase-based translation. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* Vol. 1.
- Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, R. Zens, M. Federico, N. Bertoldi, B. Cowan, W. Shen and C. Moran. 2007. Moses: open source tool-kit for statistical machine translation. *Proceedings of the Association for Computational Linguistics*.

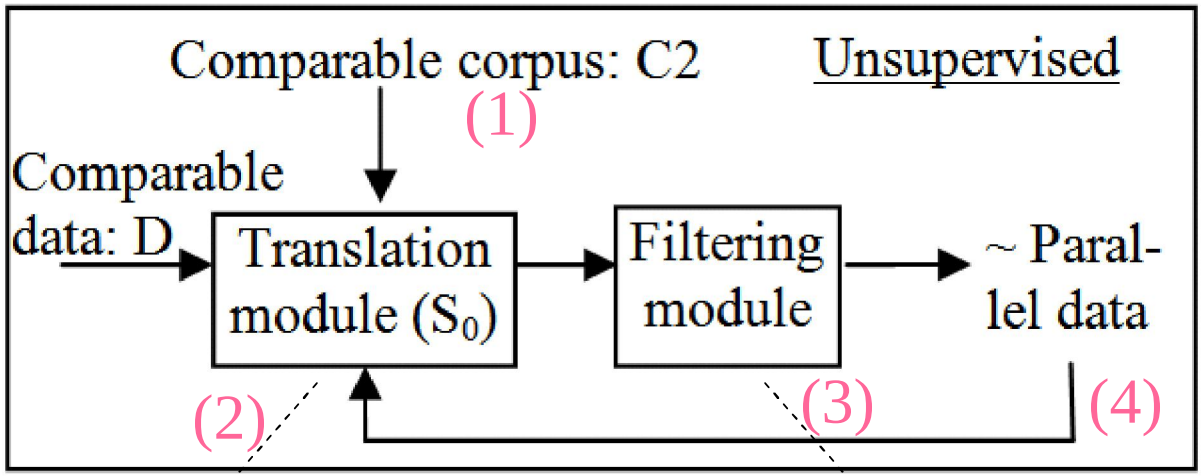
References

- Kumano, T., H. Tanaka, T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. *Conference on Theoretical and Methodological Issues in Machine Translation.*
- Ma, Xiaoyi. 2006. Champollion: A robust parallel text sentence aligner. *LREC: Fifth International Conference on Language Resources and Evaluation.*
- Munteanu, D.S. and D. Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. *44th annual meeting of the Association for Computational Linguistics*.
- Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29.1*
- Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU:a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.
- Patry, A. and P. Langlais. 2005. Paradocs: un système d'identification automatique de documents parallèles. *12e Conference sur le Traitement Automatique des Langues Naturelles*.
- Resnik, P. and N.A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics.*
- Sarikaya R., S. Maskey, R. Zhang, E. Jan, D. Wang, B. Ramabhadran, S. Roukos. 2009. Iterative sentence–pair extraction from quasi–parallel corpora for machine translation. *Interspeech*.
- Snover M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*.
- Stolcke, Andreas. 2002. SRILM an extensible language modeling toolkit. *Intl. Conf. on Spoken Language Processing.*
- Tillmann C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based search for statistical translation. *In 5th European Conf. on Speech Communication and Technology*.
- Do T.N.D., V.B. Le, B. Bigi, L. Besacier, E. Castelli. 2009. Mining a comparable text corpus for a Vietnamese-French statistical machine translation system. *4th Workshop on Statistical Machine Translation*.
- Zhao B., S. Vogel. 2002. Adaptive parallel sentences mining from Web bilingual news collection. *International Conference on Data Mining*.


L I G


MICA

Our unsupervised learning method



- a proper statistical machine translation system

- evaluation metrics (Bleu, Nist, Ter, Per*)

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

- Iterative scheme with different combinations:
  - W1: S_0 at step *i* is retrained on C2 and E_{i-1}
  - W2: S_0 at step *i* is retrained on C2 and $E_0+E_1+\ldots+E_{i-1}$
  - W3: at iteration *i*, a new separate phrase-table is built based on the extracted data E_{i-1}. System decodes using both phrase-table of S_0 and this new one (log-linear model) without weighting them.
  - W4: the same combination as W3, but the phrase-table of S_0 and the new one are weighted, e.g. 1:2.