A Bilingual Graph-based Semantic Model for Statistical Machine Translation

Rui Wang[1], Hai Zhao[1], Sabine Ploux[2], Bao-Liang Lu[1], and Masao Utiyama[3]

[1]Shanghai Jiao Tong University
[2]Centre National de la Recherche Scientifique
[3] National Institute of Information and Communications Technology

Bilingual Word Embedding

- Bilingual word embedding can enhance many cross-lingual NLP tasks, such as word translation, cross-lingual document classification and SMT.
- According to the *cross-lingual* projection step, there are mainly three types of bilingual embedding methods.
  - 1) Each language is embedded separately at first, and transformation of projecting one embedding onto the other. [*Mikolov, 2013*]
  - 2) Parallel sentence/document-aligned corpora are used for learning word or phrase representation directly, such as a series of NN methods.
  - 3) Monolingual and bilingual objectives are optimized jointly, such as BiLBOWA [Gouws et al. 2015]

Bilingual Graph-based Semantic Model

- **Motivation**
  - Most of the existing methods for bilingual word embedding only consider shallow context or simple co-occurrence information.
  - Sense information gives more exact meaning representation than word information itself.
  - Dynamic representation: A word may have multiple senses.
- **Hypotheses:**
  - Bilingual Contexonym Clique (BCC) as smallest bilingual sense unit.
  - Construct the cross-lingual relationship before the projection step.
  - To embed words dynamically according to contextual information.
  - Apply word embedding to phrase translation and generation.

Graph Constructing

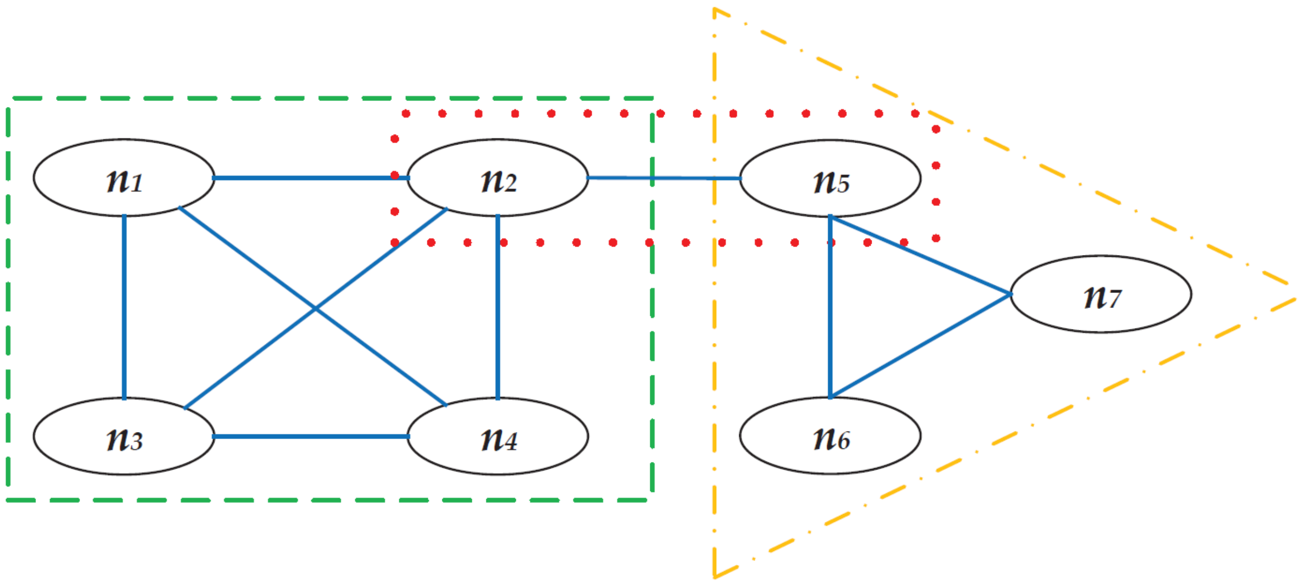
- Formally, words are considered as nodes (vertices) and co-occurrence relationships of words are considered as the edges of graph. An edge-weighted graph derived from a bilingual corpus is defined as,

$$G = \{W, E\},$$

- The *Edge Weight* (*EW*) connecting nodes n_i and n_j is defined by a modified PMI measure,

$$EW = \frac{Co(n_i, n_j)}{fr(n_i) \times fr(n_j)}$$

Context-Dependent Clique Extraction



- Clique in this thesis: a maximum, complete sub-graph.
- Only the co-occurrence nodes n_{ij} of each n_i (including n itself) are useful and kept.

$$|N_{extracted}| = \left| \bigcup_{\forall i,j} \{n_{ij}\} \right|$$

Bilingual Contexonym Clique (BCC)

- As the clique is to represent a fine grained bilingual sense of a word given a set of its contextual words, it is called **Bilingual Contexonym Clique (BCC)**.

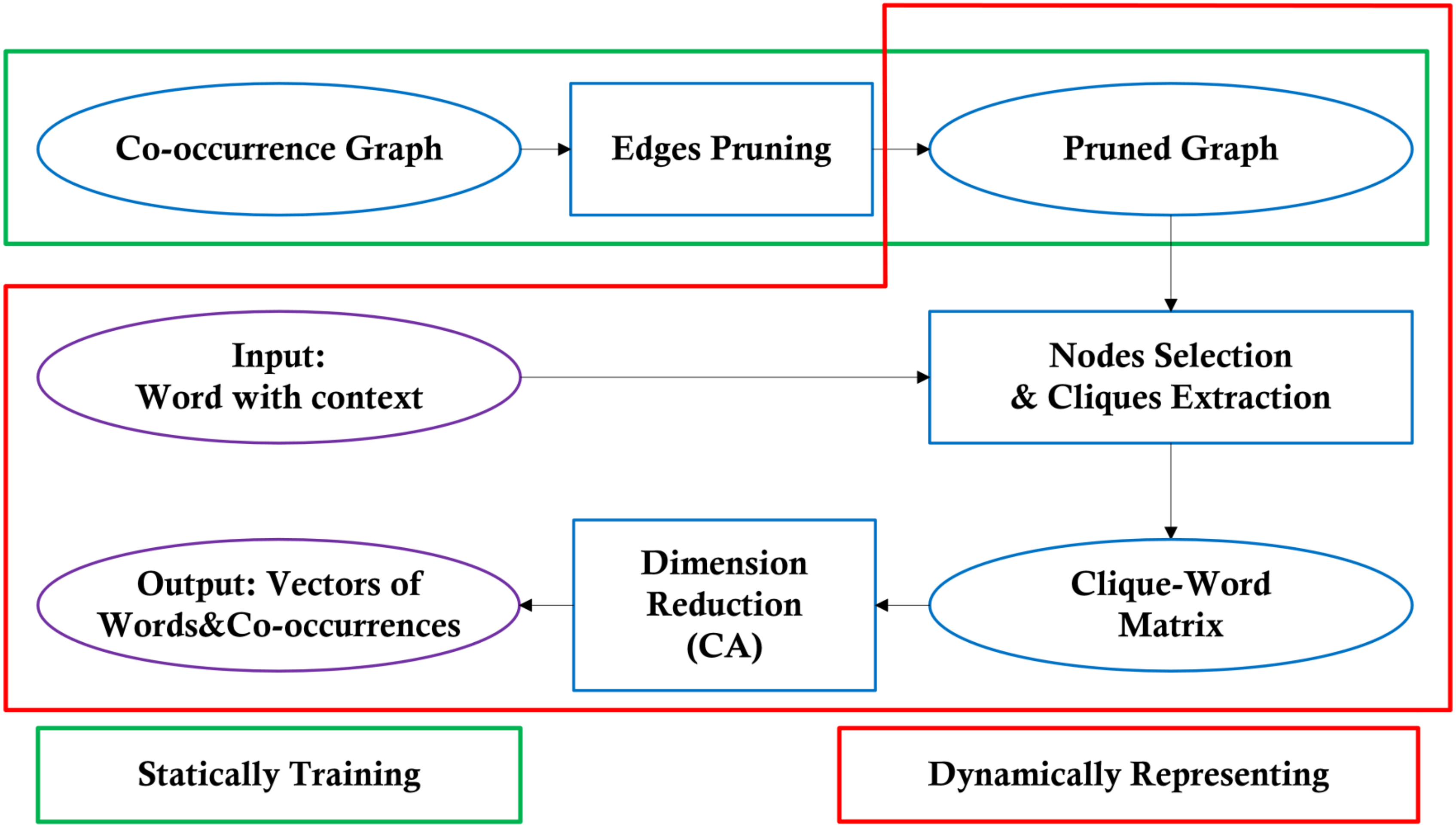
Words	BCCs
work_e	{*employees_e*, *travail_f* (work), *unemployed_e*, *work_e* } {*heures_f* (hours), *travaillent_f* (to work, third-person plural form), *travailler_f* (work), *week_e*, *work_e* } {*readers_e*, *work_e* }...
readers_e	{*informations_f* (information), *journaux_f*(newspapers), *online_e*, *readers_e*} {*journaux_f* (newspapers), *lire_f* (read), *newspaper_e*, *presse_f* (press), *readers_e*, *reading_e*} {*readers_e*, *work_e*}...

Correspondence Analysis (CA)

- CA (Benzécri, 1980), which is based on SVD, measure and assess semantic variations of principal axes.
- To project words/BCC onto lower dimensional semantic space, CA is conducted over the clique-word matrix constructed from the relation between BCCs and words.

	w_1	w_2	w_3	…
BCC_1	0	0	1	
BCC_2	1	1	0	
BCC_3	0	0	1	
…				

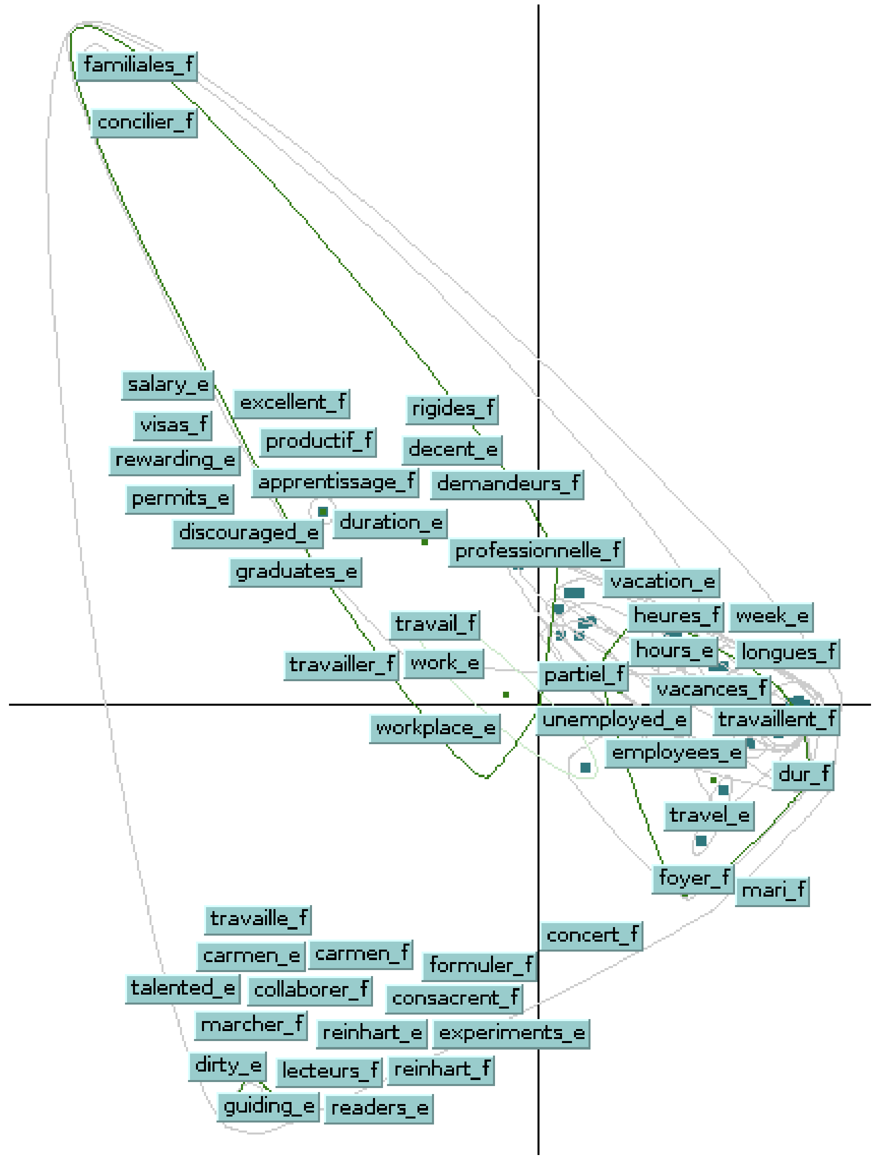
Entire Pipeline



[*Ours IJCAI-2016*]

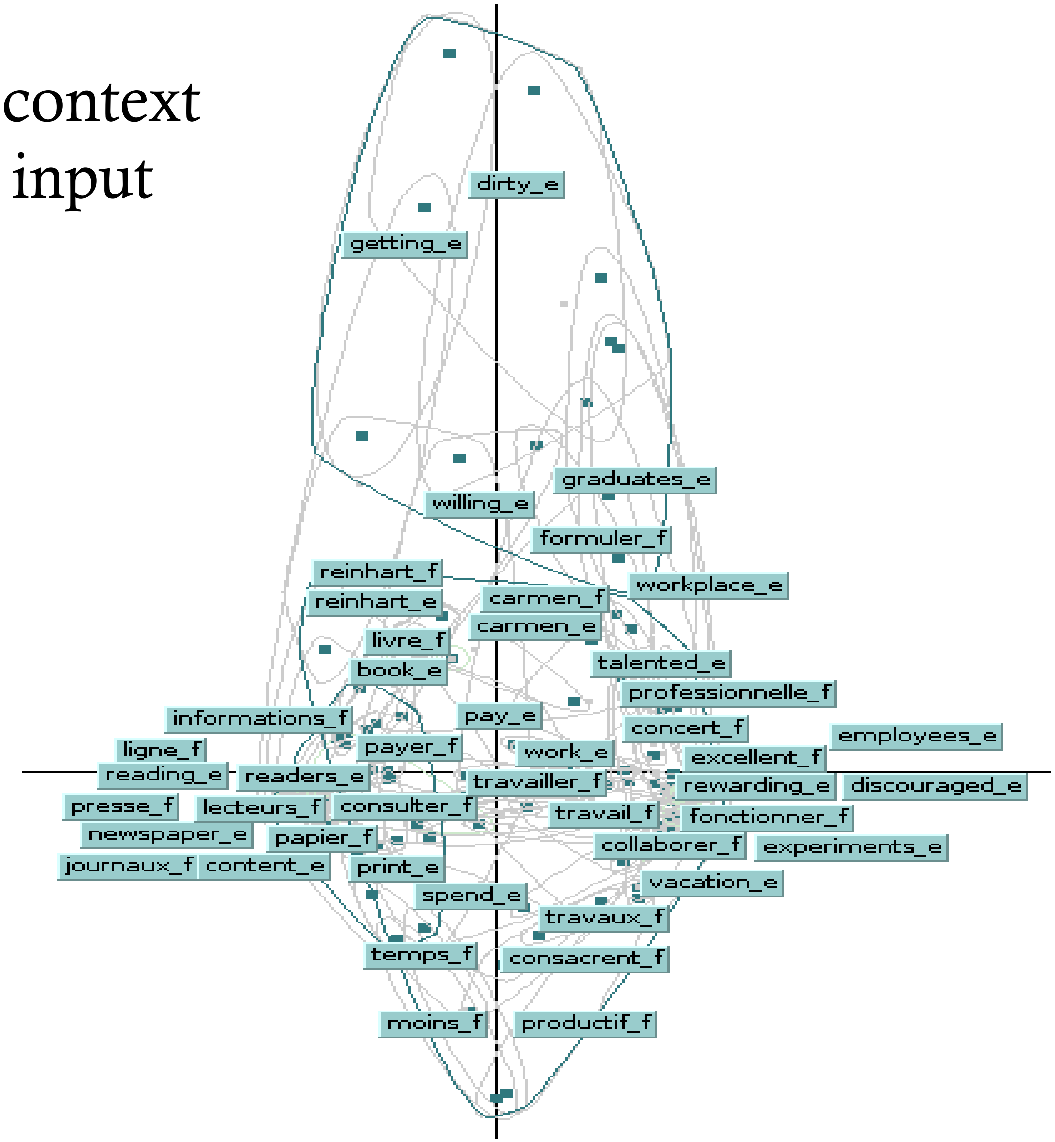
Semantic Spatial Representation

"Work" as input



Contexts as input

"Work" with context "readers"as input



Phrase Translation

- The phrase-table of phrase-based SMT model can be simply formalized as:

$$(P_F, P_E, \text{scores, word-alignment})$$

  - Strategy-A: only the source words in P_F are used as contextual words.
  - Strategy-B: both the source words in P_F and target words in P_E are used as contextual words.

Semantic Similarity Measurement

- Because the lengths of phrases are different, *Normalized Euclidean Distance* (*NED*) is adopted to measure the distance between source and target phrases incorporated with word-alignment model:

$$NED(P_F, P_E) = \sqrt{\frac{\sum_{align(i,j)} ED^2(V_{wf_i}, V_{we_j})}{\sum_{i,j} align(w_{f_i}, w_{e_j})}}$$

- NED is added as additional feature of phrase based SMT.

Bilingual Phrase Generation

- Word w and its co-occurrence words are represented as vectors. For a aligned word pair (w_{fi}, w_{ej}), they are represented as vectors (V_{fi}, V_{ej}) and their co-occurrence words *fwcog* are represented as vectors Vco. We need to find new translation candidate w'_{ej} in w_{co} to form new phrase pair (w_{fi}, w'_{ej}).

Source	Original Target	CSTM Generated	BGSM Generated
la bonne réponse	*the right answer*	*1. a right answer* *2. all right answer* *3. the right reply*	*1. the correct answer* *2. the right response* *3. the good answer*
nettoyer le jardin	*clean the garden*	*1. clean a garden* *2. clean the yard* *3. clean an garden*	*1. clean the yard* *2. clean the ground* *3. tidy the garden*

$$DR(P'_E, P_E) = \frac{NED(P_F, P'_E)}{NED(P_F, P_E)}$$

Experiments (Chapter 5.4)

- Corpora

Corpus	IWSLT	NCTIR	NIST
training	186.8K	1.0M	2.4M
dev	0.9K	2.0K	1.6K
test	1.6K	2.0K	1.3K

- Phrase Translation: BLEU

	IWSLT	NTCIR	NIST
Baseline	31.80	32.19	30.12
+Zou	N / A	N / A	30.36
+CSTM	32.19	32.37	30.25
+BGSM-A	32.32+	32.56	30.38
+BGSM-B	**32.61++**	**33.04++**	**30.44+**

Experiments

- Phrase Generation

Corpora	Methods	Phrase Pairs	BLEU
IWSLT	Baseline	9.8M	31.80
	+CSTM	23.1M	32.19
	+Saluja	31.5M	32.35
	+BPG	25.6M	32.37
	+BPG+BGSM	25.6M	**33.13++**
NTCIR	Baseline	71.8M	32.19
	+CSTM	297.8M	32.42
	+Saluja	341.3M	32.68
	+BPG	312.6M	32.54+
	+BPG+BGSM	312.6M	**33.47++**

- Efficiency Comparison

Methods	Training Time	Calculating Time
CSTM	59.5 Hours	17.1 Minutes
BGSM-A	1.1 Hours	8.9 Minutes
BGSM-B	1.1 Hours	15.6 Minutes

Thank You