Unsupervised Joint Training of Bilingual Word Embeddings

Benjamin Marie Atsushi Fujita
{bmarie, atsushi.fujita}@nict.go.jp
National Institute of Information and Communications Technology (NICT), Kyoto, Japan

Summary

Key points

- Joint training **directly learns bilingual word embeddings (BWE) on parallel data** (cf. mapping methods)
- **Unsupervised machine translation generates synthetic parallel data** by translating monolingual data
- Unsupervised joint training uses **synthetic parallel data for pseudo-supervision**

Findings

- **Largely outperforms** BWE obtained through unsupervised mapping in bilingual lexicon induction tasks
- **Robust** to noisy synthetic parallel data and **takes advantage of monolingual and bilingual contexts simultaneously**

Assumptions and Research Question

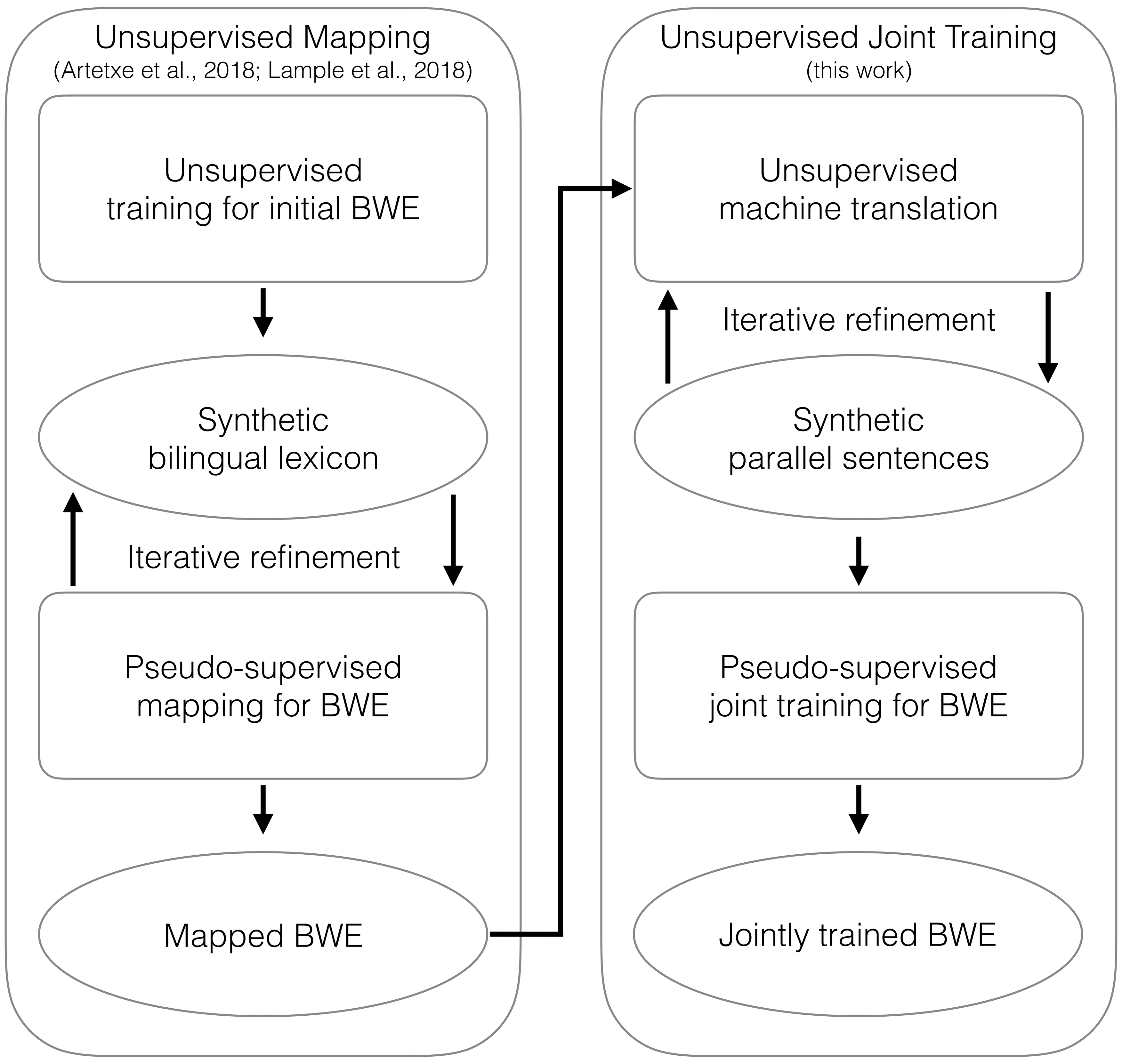
Assumptions

1. Mapping methods for BWE are **limited** by the dissimilarity between the original word embedding spaces to be mapped (Søgaard+, 16)
2. Supervised joint training algorithms (Upadhyay+, 16) are **robust to noisy data**
3. Synthetic parallel data of a **reasonable quality can be generated** through unsupervised machine translation (Artetxe+, 18a; Lample+,18)

Research question

Do synthetic sentence pairs, generated without supervision, supply useful bilingual contextual information for jointly learning better BWE?

Unsupervised Joint Training



Evaluation in Monolingual word Analogy

Does unsupervised joint training improve or preserve monolingually the quality of the word embeddings (as observed for supervised bilingual skipgram)?

Settings

- Task: English word analogy (Mikolov+, 13)
- Unsupervised systems: `VECMAP` (bilingual mapping), `BIVEC` (bilingual joint training), and `fastText` (monolingual)

Results

Method	Data used	Accuracy
`VECMAP`	239M(en)-38M(fr)	77.8
`BIVEC`	10M(en)-10M(synthetic fr)	65.7
`fastText`	239M(en) 10M(en)	79.1 64.6

- `BIVEC` outperforms `fastText` when they are trained on the same data

⇒ Joint algorithms take advantage of **noisy but bilingual contexts** to **monolingually improve word embeddings**

Evaluation in Bilingual Lexicon Induction

Are BWE unsupervisedly and jointly trained on noisy synthetic data better than unsupervised mapped BWE?

Data

- Monolingual training data:
  News crawl for en-de and en-fr, Common Crawl for en-id
- Test sets: "full" Muse Wikipedia bilingual lexicons

Baseline systems: unsupervised mapping

- `VECMAP`: `fastText` word embeddings mapped without supervision (Artetxe+, 18b)

Evaluated systems

- Training data: synthetic parallel data generated with unsupervised statistical machine translation (Lample+, 18) by translating source and/or target sentences
- `BIVEC`: bilingual skipgram using pre-trained word alignments (Luong+, 15)
- `SENTID`: skipgram on a word/sentence-ID matrix (Levy+, 17)

Results (acc@1)

Method	Data used src-tgt	en→de	de→en	en→fr	fr→en	en→id	id→en
`VECMAP`	all-all	42.4	59.0	67.7	70.0	58.9	59.5
`BIVEC`	10M-0	**45.8**	59.2	73.9	71.3	**70.4**	69.7
`SENTID`	10M-0	**45.8**	60.1	**74.4**	71.8	69.8	69.2
`BIVEC`	0-10M	43.7	63.4	72.0	74.3	67.3	72.3
`SENTID`	0-10M	43.5	**63.5**	72.6	**74.8**	67.5	**73.4**
`BIVEC`	10M-10M	44.9	54.9	73.9	73.8	69.5	72.1
`SENTID`	10M-10M	45.4	62.1	74.2	74.0	69.4	73.0

- **Unsupervised joint training outperforms unsupervised mapping** by a large margin
  - **More than 10 points of improvement** for en-id
  - Higher accuracy when synthetic parallel data do not contain synthetic English (e.g., "10M-0" for en→de, en→fr, and en→id)
- `BIVEC` and `SENTID` performs **similarly**
  - pre-trained word alignments are unnecessary

⇒ Joint algorithms are **robust to noise and learn better BWE** for bilingual lexicon induction