Improving Bilingual Lexicon Induction on Distant Language Pairs

Wenhao Zhu[1] Zhihao Zhou[1] Shujian Huang[1(✉)] Zhenya Lin[2]
Xiangsheng Zhou[2] Yaofeng Tu[2] Jiajun Chen[1]

[1] Nanjing University
{zhuwh, zhouzh}@smail.nju.edu.cn {huangsj, chenjj}@nju.edu.cn
[2] ZTE Corporation
{lin.zhenya, zhou.xiangsheng, tu.yaofeng}@zte.com.cn

Sept 28, 2019

Outline

1. **Introduction**
2. **Experiment**
3. **Conclusion**

Outline

Bilingual Lexicon Induction

- Aligning the representation spaces of two languages to conduct bilingual lexicon induction (BLI) achieves attractive results on European language pairs. [Mikolov et al. 2013]

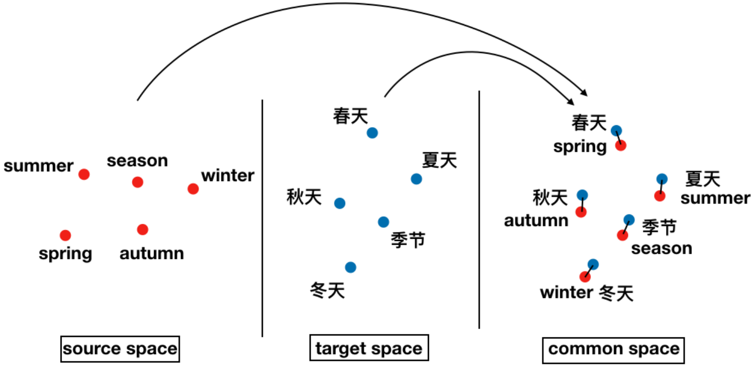


Figure 1: Illustration of bilingual lexicon induction

Bilingual Lexicon Induction

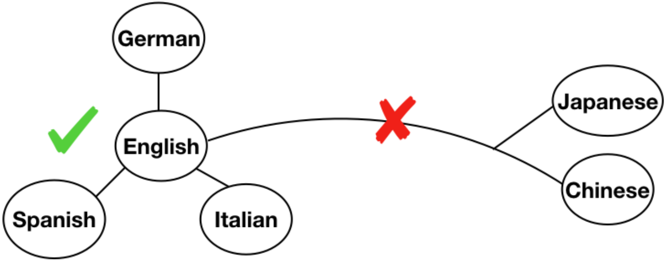
There are two popular branches in researches of BLI

- Supervised methods: seed dictionary
- Unsupervised methods: self-learning, GAN-based methods (unstable)

Thus we mainly discuss the **supervised methods** in this paper.

Motivation

- Can't be applied directly to distant language pairs (e.g. EN-ZH, EN-JA) and perform terribly on these language pairs
- Less attention is paid on this problem



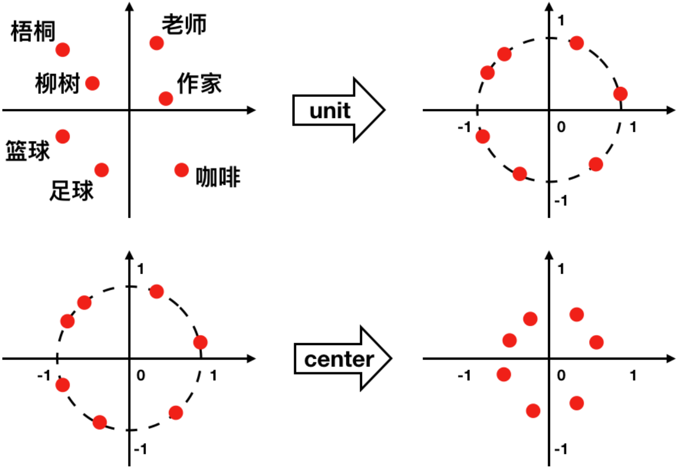
Background

Supervised Method

- Preprocessing
- Mapping
- Inference

Step 1 - Preprocessing

- Transform the representation space before mapping, such as “unit”, “center”, etc.



Step 1 - Preprocessing

- Transform the representation space before mapping, such as “unit”, “center”, etc.

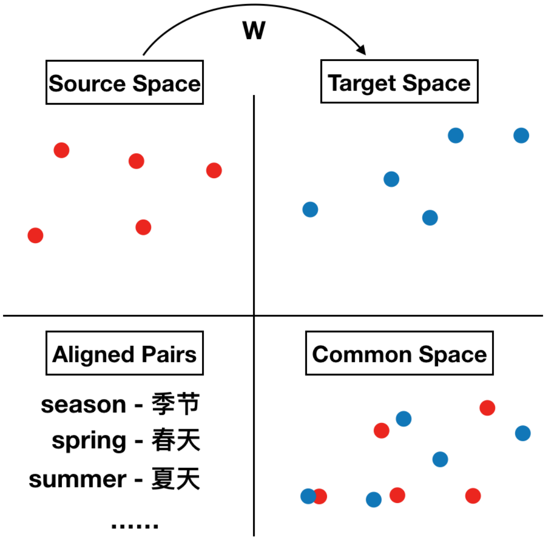
Limitation

- There is no guidance on using these transformations for distant language pairs. Simply stacking them can’t ensure the same effect on these language pairs

Our Work

- We make empirical analysis of these transformations on English-Chinese
- Our hypothesis
  - **"unit"** and **"center"** are the most important operations
  - other transformations do not bring obvious improvement
  - ...

Step 2 - Mapping



Step 2 - Mapping

- Make aligned pairs stay as close as possible with matrix W

$$\arg\min_{W} \sum_{i} ||X_{i*}W - Y_{i*}||^2 \tag{1}$$

- Orthogonal constraint is proposed to be added into (1)

$$W^T W = I \tag{2}$$

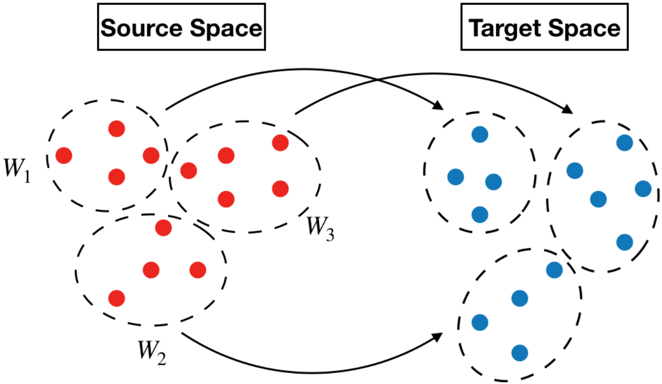
- Neural mapping suffers severe overfitting problem

Limitation

- Using a single matrix W as transformation function has an idealized assumption: vector spaces have similar geometric arrangement. We find it's not held for distant language pairs

Our Work

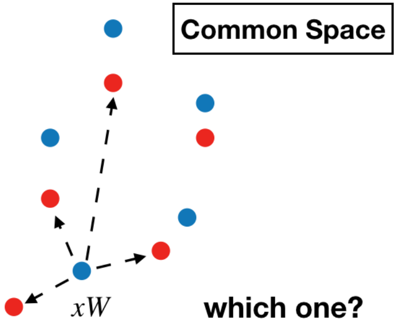
- Multiple Local Mappings
  - Similar geometric distribution may only happens **locally**
  - A set of multiple local mappings $\{W_i\}_{i=1}^{m}$ rather a single mapping W better model BLI on distant pairs



Step 3 - Inference

- Obtain translation pairs from the mapped space with retrieval method
- For a given word x, its induction translation y is:

$$\arg\min_{y} f(xW, y) \tag{3}$$



Step 3 - Inference

- Nearest neighbour (NN) suffers a severe problem [Dinu et al. 2015]
- Hubness problem
  - some meaningless target words (for example: aaaa, 1988-03) which appear as the nearest neighbour of many source words

Figure 2: Illustration of hub word

Step 3 - Inference

- *Invnn*, *Invsoftmax*, *CSLS** are proposed to cope with "hub word"
- Retrieval formula: [Conneau et al. 2018]

$$CSLS(xW, y) = 2\cos(xW, y) - r_T(xW) - r_S(y) \tag{4}$$

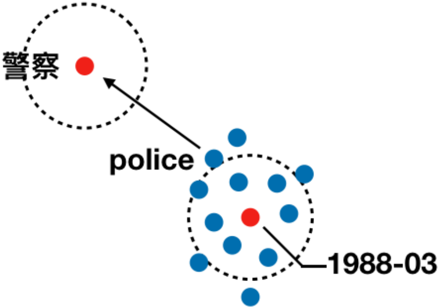


Figure 3: Illustration of inducing word pairs with *CSLS*

Step 3 - Inference

- Topic word
  - represents a broad concept
  - also has great similarity with surrounding words

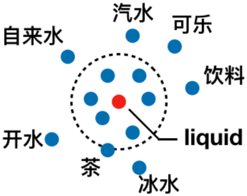


Figure 4: Illustration of topic word

Limitation

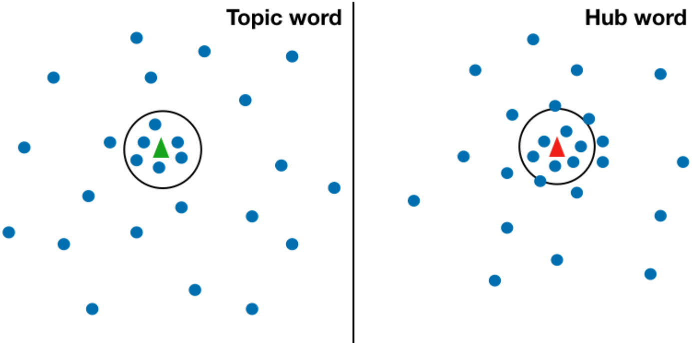
- Though *CSLS* enjoys success in its efficiency and low computation expense, it still faces some problems in practice.
- We find *CSLS* always confuses "**topic words**" with "hub words"

Our Work

We find...

Tuning hyper-parameter K in *CSLS* enables the model to distinguish between "topic word" and "hub word"

- Explanation: "hub word" always has high similarity with surrounding words while "topic word" not.

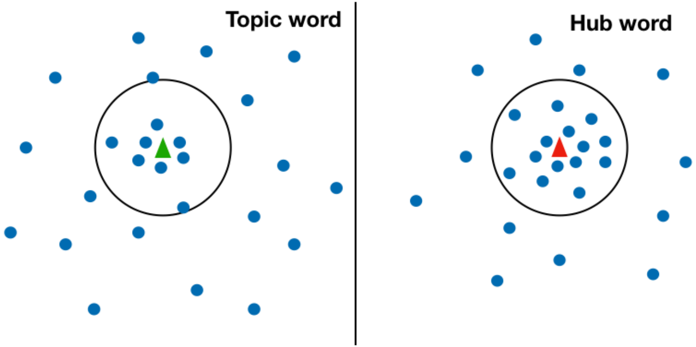


Our Work

We find...

Tuning hyper-parameter K in *CSLS* enables the model to distinguish between "topic word" and "hub word"

- Explanation: "hub word" always has high similarity with surrounding words while "topic word" not.



Our Work

We propose...

- an **approximated searching algorithm** to determine K.
  - Increase K in step of 10 and compute model accuracy on the training set;
  - Once induction performance declines, we choose K in the last step as optimal value.

Outline

Settings

- Dataset
  - Fasttext dataset built by Facebook Inc.
  - Pretrained on Wikipedia corpus by skip-gram model
  - Five language pairs:
    English $\rightarrow$ Chinese, Japanese, Korean, Finish, German
  - Training set: the most frequent 5000 words
    Test set: following 1500 words
- 300-dims word embedding
- For other detailed settings please refer to our paper

Empirical Study of Transformations

Oberservation

- "Unit" plus "center" is the optimal combination for distant language pairs, "center" brings the most performance gain
- "Unit" and "center" are the most effective way to make distribution similar without need of supervised signal

unit	center	whiten	de-whiten	re-weight	reduction	Acc.
						27.33%
✓						27.13%
✓	✓					**42.47%**
✓	✓	✓				42.47%
...						42.47%
✓	✓	✓	✓	✓	✓	42.47%

Table 1: Different combinations' accuracy on English-Chinese.

Employing Multiple Mapping Function

Observation

- The baseline model acts poorly on training set which indicates that a single mapping is far from perfect
- The accuracy of using multiple local mappings is substantially better than a single global map for different groups.

topic word	train dict size	ACC_{tr}	test dict size	ACC_{te}
"animal"	1230	94.74	471	51.15
"culture"	1331	92.95	342	52.34
"education"	1315	92.60	351	51.24
average		93.43		51.58
single mapping		45.14		32.47

Table 2: Train set accuracy (ACC_{tr}) and test set accuracy (ACC_{te}) of high quality local mappings on English-Chinese

Employing Multiple Mapping Function

Note

- Automatically choosing the number of local mappings and selecting reasonable topic words for each mapping are difficult
- At the current stage, this method is not integrated into our final system

Inference with Approximated Searching

Observation

- Setting $K = 10$ is not the best choice
- Tuning K enables the model to distinguish between "topic word" and "hub word"

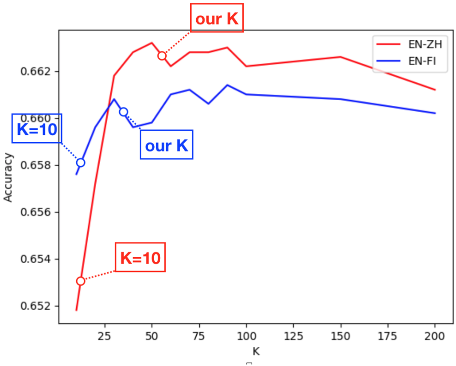


Figure 7: BLI accuracy on English-Chinese, English-Finnish

Experimental setting

Our Final Framework

- Preprocessing: unit + center
- Mapping: orthogonal matrix
- Inference: *CSLS* with our searching K

Baseline

- No preprocessing
- Mapping: orthogonal matrix
- Inference: *CSLS*

Results

- We conduct experiments on both distant and close language pairs

	distant pairs				close pairs
	EN-ZH	EN-JA	EN-KO	EN-FI	EN-DE
Mikolov et al., 2013	13.27	14.16	16.11	32.47	61.20
Xing et al., 2015	27.13	2.54	24.64	38.67	68.13
Dinu et al., 2015	27.00	32.49	25.32	43.33	66.33
Artetxe et al., 2016	42.47	45.65	27.03	42.93	70.30
Smith et al., 2017	12.47	1.10	25.05	44.60	71.40
Nakashole et al., 2018	43.27	-	-	-	68.50
baseline	32.47	1.71	31.47	47.60	73.37
baseline + uc	45.33	51.68	31.54	65.76	79.02
baseline + uc + *CSLS'*	**45.80**	**51.68**	**32.29**	**66.08**	**79.34**

Table 3: Precision for BLI task compared with previous work.

Further analysis

We may ask...

- What prevents the model inducing perfect lexicon ?

Source Word	Predicted Word	Ground Truth
ear	舌头 (tongue)	耳朵 (ear)
myanmar	泰国 (thailand)	缅甸 (myanmar)
honey	柚子 (Pomelo)	蜂蜜 (honey)
plural	单数 (singular)	复数 (plural)

Table 4: Some representative wrong translation pairs made by our improved framework on English-Chinese

Outline

Conclusion

Contribution

- Make deep analysis on the English-Chinese word translation task.
- Propose three methods to address observed problems.
- Present an improved framework on distant language pairs.

Future Work

- Complete the algorithm of multiple local mappings
- Eliminate the effect brought by words with similar context

Reference

- Exploiting Similarities among Languages for Machine Translation. Mikolov et al. arXiv 2013.
- Improving zero-shot learning by mitigating the hubness problem. Dinu et al. ICLR 2015.
- Word Translation Without Parallel Data. Conneau et al. ICLR 2018.