Improving Bilingual Lexicon Induction for Low Frequency Words

March 26, 2022

1 Motivation

现有的基于 Procrustes 和余弦相似度等方式的词典抽取方法，在低频词对齐问题上的表现非常不好，正确率和 hubness 问题都非常突出。这篇工作首先定量的分析了翻译的正确率和 hubness 关于词频和词典大小的变化情况，其次提出了两个方法对于该问题进行改进。

2 Lexicon Induction at Low Frequency

本篇工作首先设计了三个实验，来探究现有的词典抽取方式的效果随词频的变化。因为在不同的语言中，一对互为翻译的词的词频可能会有比较大的差距，所以这篇工作在单语数据上进行词频的实验。

2.1 Monolingual Lexicon Induction (MLI)

本文利用两个不同的单语数据集分别训练词向量，得到了单语数据上的两种词向量表示，通过对齐这两个集合在单语数据上进行实验。本篇文章首先将 500K 个词按照频率分成 50 个组（bins），频率分别由高到低排列。在每一个 bin 中，选取不同的大小进行分割，分别作为测试集和种子词典。本篇文章分别在不同的 bin 中利用 NN 距离进行实验，同时利用不同大小的种子词典进行测试，实验结果如图1a 所示。可以看到，随着词频的下降，准确率下降的非常严重，同时种子词典的规模在 1K 变化到 10K 的过程中对正确率有比较大的影响，在 10K 变化到 300K 的过程中，正确率几乎不再改变，可见种子词典的作用是存在上限的。

2.2 Cosine Similarities and Margin

定义词 x_i 所对应的利用 NN 得到的翻译为 $trans(x_i)$，我们希望在所有的目标集合 Y 中，$trans(x_i)$ 是 Wx_i 中最接近 Y 的，即：

$$cos\left(Wx_i, trans\left(x_i\right)\right) \geq cos\left(Wx_i, y_j\right) \tag{1}$$

定义 $margin$ 为：

$$M\left(x_i\right) = cos\left(Wx_i, trans\left(x_i\right)\right) - \max_j cos\left(Wx_i, y_j\right), y_j \neq trans\left(x_i\right) \tag{2}$$

当 $M(x_i) < 0$ 时，就说明翻译错误。$M(x_i) < 0$ 的结果随词频的变化如1b 所示，可以看到错误率随词频的降低而上升。

2.3 Hubness and Tail of k-occurrence

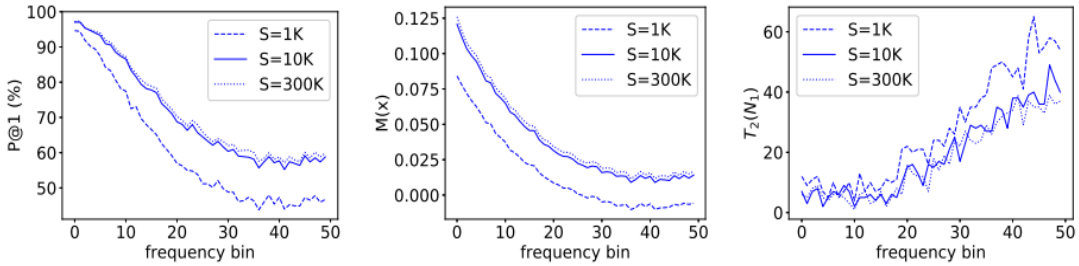
定义 $N_k(y; Q)$ 为在查询集合 Q 中，词 y 作为翻译的次数，这里采用 $k-NN$ 的方式作为查询方式，其具体的定义如下：

$$N_k(y; Q) = |\{x \in Q : y \in k - NN(x)\}| \tag{3}$$

定义 $T_n(N_k)$ 为 N_k 中大于 n 的数量：

$$T_n(N_k) = |y : \{N_k(y; Q) > n\}| \tag{4}$$

所以 $T_n(N_k)$ 越大说明 hubness 问题越严重，图1c 中展示了 $T_n(N_k)$ 随频率变化的结果，可以看到，$T_n(N_k)$ 随频率降低而升高，说明低频词的 hubness 现象非常严重。



(a) Induction accuracies (b) margins (c) tailness of N_1

Figure 1: NN 方法实验结果

3 Two Methods

3.1 Hinge Loss for Learning Transformation

定义公式：

$$\min_{W \in O(d)} \sum_i \sum_{j: y_j \neq trans\left(x_i^s\right)} max\left\{0, \gamma - cos\left(W x_i^s, y_i^s\right) + cos\left(W x_i^s, y_i\right)\right\} \tag{5}$$

其中 $\gamma > 0$，该公式训练后一项的值尽可能大于 $gamma$，以此来提升对齐的质量。

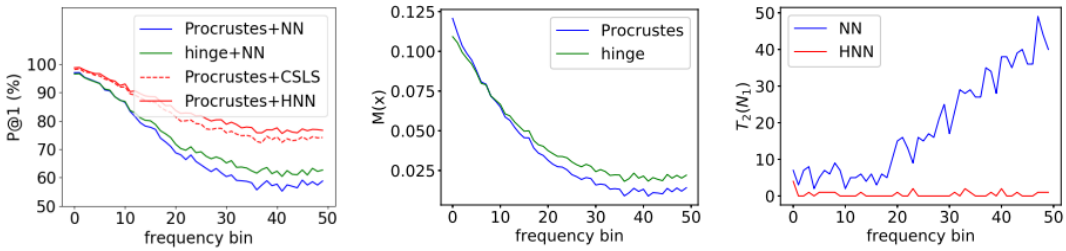
3.2 Hubless Nearest Neighbor (HNN) Search

定义矩阵 $P_{i,j}$ 表示单词 x_i 翻译为 y_j 的概率，为了减小 hubness 情况，对 P 采取限制：$\sum_j P_{i,j} = 1$ 和 $\sum_i P_{i,j} = \frac{m}{n}$，这两个约束分别保证了 x_i 对应的所有翻译的概率和为 1，y_j 作为翻译的概率时均等的。第二个约束可以有效的解决 hubness 问题。利用如下公式训练得到矩阵 P，并利用 P 进行候选词筛选。

$$\min_{P \in O[0,1]^{m*n}} \sum_{i,j} P_{i,j} cos\left(W x_i, y_i\right), s.t. \sum_j P_{i,j} = 1, \sum_i P_{i,j} = \frac{m}{n} \tag{6}$$

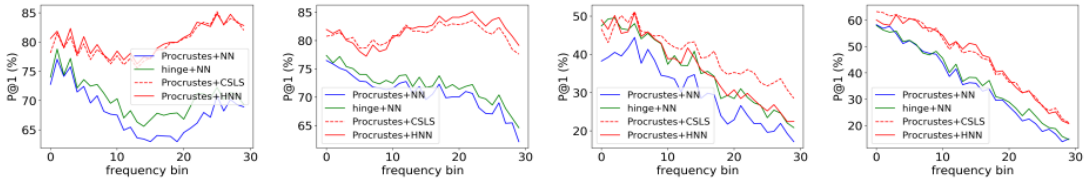
4 Experiment

如图2和图3所示，该方法在 Procrustes 和 NN 的基础上有效的解决了 hubness 问题。



(a) Induction accuracies (b) margins (c) tailness of N_1

Figure 2: 单语数据实验结果



(a) en-fr (b) fr-en (c) en-fi (d) fi-en

Figure 3: 双语数据实验结果

5 Conclusion

本篇文章的研究角度比较新颖，探究了词频和种子词典大小对于词抽取正确率的影响以及对 hubness 问题的影响。该文章的方法可以较为有效的解决 hubness 问题，其次其在低频词上的效果也有比较大的提升。最后从其实验中也可以发现种子词典的大小的作用也是有一定的上限的。

总体来说，这篇工作的实验是比较有亮点的。