跨语言分布表示学习方法概述

郭江

2016.10.07

分布表示（distributed representation）与深度神经网络（deep neural network）极大地推动了近几年自然语言处理研究的发展。我们知道，分布表示指的是对于一个客观描述对象的低维、稠密、连续向量表示。不同于符号表示（symbolic representation），分布表示可以自然地联结表现形式不同但语义空间相同（或相似）的对象，比如不同模态数据（图像、文本、语音）以及不同语言的数据等。近年来，很多研究致力于跨语言分布表示学习，即：将不同语言的词语嵌入至一个统一的向量空间之内，使得语义相似（单语及跨语言）的词在该向量空间内距离接近，如图1所示。跨语言分布表示为不同语言的数据资源之间建立了一座桥梁，为跨语言的迁移学习提供了一种有效的途径，进而为很多跨语言应用（比如对于资源稀缺语言的文本分类、句法分析，以及机器翻译等）带来显著的推动作用。

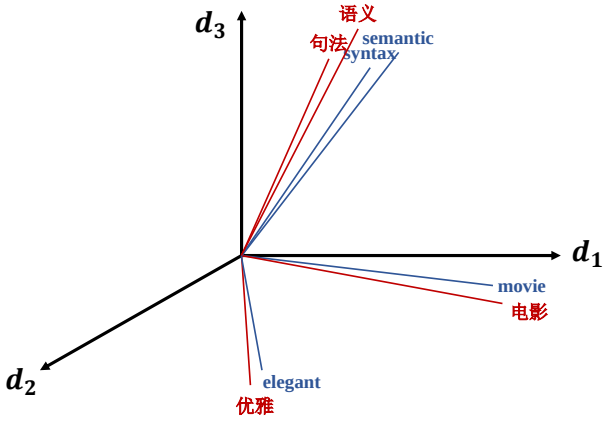

Figure 1: 跨语言词汇分布表示（三维空间下的示意）

本文首先描述典型的跨语言分布表示学习方法，再简单介绍其应用场景。

1　跨语言分布表示学习

根据学习方式的不同，可以将现有的跨语言分布表示学习分为两类，分别是：1. 基于线下处理的方法；2. 基于联合学习的方法。接下来我们对这两类方法中较为典型的模

型进行介绍。

1.1 线下处理方法

线下处理方法的基本思路是先独立学习各个语言的词汇分布表示，然后对两者进行对齐。Mikolov等发现，使用*word2vec*学习得到的不同语言的分布表示之间存在一定程度上的线性映射关系（类似于同种语言下的word analogy性质）（见图2），于是提出“翻

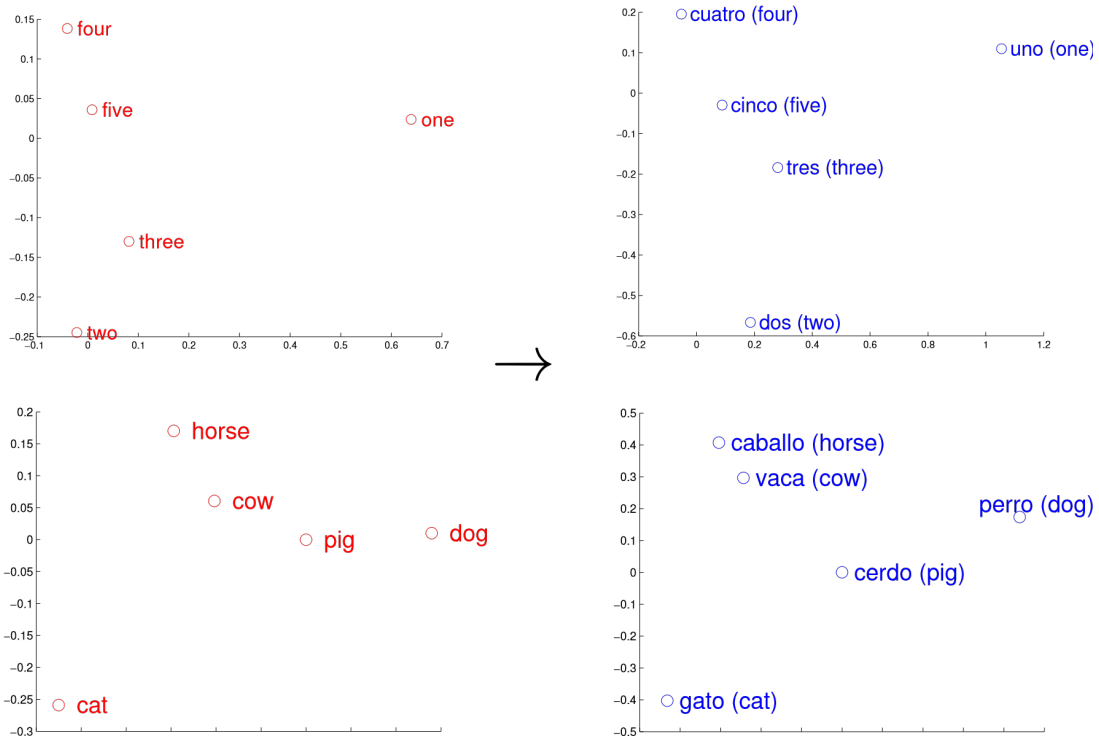

Figure 2: 英语与法语词表示示例（数字、动物）

译矩阵”学习的方法来实现跨语言分布表示的映射 (Mikolov et al., 2013)。具体地，给定一个翻译词对的集合$\mathbb{D} = \{x_i, z_i\}_{i=1}^{n}$（即双语词典，其中$x_i$为源语言中第$i$个词，$z_i$为目标语言中与$x_i$互为翻译的词），$\mathbb{D}$中词对所对应的分布表示矩阵分别记为$\boldsymbol{E}_{\mathbb{D}}^{src}$以及$\boldsymbol{E}_{\mathbb{D}}^{tgt}$。作者假设存在一个从源语言到目标语言的线性映射矩阵$\boldsymbol{W}$，使得平方误差最小：

$$\min_{\boldsymbol{W}} \| \boldsymbol{E}_{\mathbb{D}}^{src} \cdot \boldsymbol{W} - \boldsymbol{E}_{\mathbb{D}}^{tgt} \|^2 \tag{1}$$

再将$\boldsymbol{W}$应用于整个源语言词表进行跨语言映射。这种方法也能够很自然地扩展到多语言（≥ 2）的情形。

另外一种方法是典型关联分析（CCA）。CCA是一种度量两个多维变量之间线性相关性的统计分析方法。对于两个多维变量，CCA寻找两个映射矩阵，将原始的变量

分别映射至新的子空间（一般情况下维度更低），使得两个变量之间的相关性最大。如果把两种语言的词表示看作是两个多维变量，则可利用CCA对这两种语言的词表示矩阵进行变换 (Faruqui and Dyer, 2014)。我们仍然考虑双语词典$\mathbb{D}$，并将词表示矩阵记为$\boldsymbol{E}_{\mathbb{D}}^{l_1}, \boldsymbol{E}_{\mathbb{D}}^{l_2}$，CCA优化以下目标:

$$\max_{\boldsymbol{W},\boldsymbol{V}} \texttt{Corr}(\boldsymbol{E}_{\mathbb{D}}^{l_1} \cdot \boldsymbol{W}, \boldsymbol{E}_{\mathbb{D}}^{l_2} \cdot \boldsymbol{V}) \tag{2}$$

Corr表示相关系数。$\boldsymbol{W}$与$\boldsymbol{V}$则可用于对两种语言的词表示矩阵进行映射。可见，CCA仍然隐含了线性变换的前提假设。然而，由于语言的复杂性，线性变换对于词表示向量之间的映射关系刻画得并不理想，鉴于此，Lu等人提出深度典型关联分析（Deep CCA），先对原始词向量进行多层非线性变换，再使用CCA处理 (Lu et al., 2015)。

事实上，一种更为简单有效的方法是根据词对齐直接进行跨语言映射。首先，我们从双语平行数据中统计出词对齐矩阵（矩阵元素为对齐次数，需要作归一化处理），再根据该矩阵对源语言词表示矩阵进行变换，如图3所示。

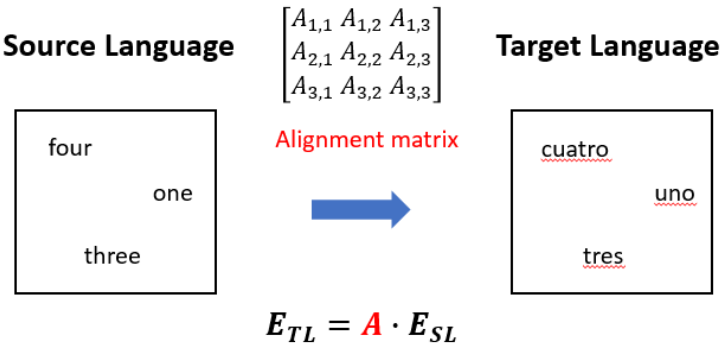

Figure 3: 跨语言映射方法

我们发表在ACL 2015的工作中对该方法进行了改进，增加了一步单语传播的过程，以缓解目标语言中OOV的问题 (Guo et al., 2015)。该方法可解释性较强，并且在跨语言依存句法分析任务上取得了最好的性能。

1.2 联合学习方法

联合学习方法的目标是同时学习多语言的词语分布表示。Klementiev等人最早提出基于多任务学习（Multi-task Learning）来同时更新不同语言的词语分布表示矩阵 (Klementiev et al., 2012)。以英文-法文为例，该方法首先从英-法双语平行数据中根据词对齐信息获得双语词汇相似度矩阵（对齐频率）。分布表示的学习仍然依赖单语数

据下的神经网络语言模型，每当计算英语中w_i^{en}的梯度时，都会根据相似度矩阵为相应法语词赋予梯度（根据相似度进行加权），从而达到联合学习的目的。该方法可以视为上述跨语言映射方法的联合学习版本。

Zou等人则借鉴前文所述“翻译矩阵”的思想，将双语词汇相似度矩阵作为跨语言映射矩阵，而以相应的平方误差损失作为单语分布表示学习目标的正则项 (Zou et al., 2013)。另一个比较独特的思路是使用跨语言AutoEncoder (Chandar A P et al., 2014)，也称相关性网络（Correlational Neural Network）。该方法将AutoEncoder的重构思想应用于双语平行数据之间的语义重构，基本思路是将一种语言的句子表示（bag-of-words）作为输入，期望重构出另一种语言的句子表示。对应的AutoEncoder权值矩阵则为所学到的跨语言分布表示。

受跨语言AutoEncoder所启发，我们还可以通过最小化双语平行数据中句对表示之间的距离，来学习跨语言分布表示。在词分布表示的基础之上通过语义组合（semantic composition）得到句子表示的方法有很多，为了保证分布表示学习的效率，一般使用较为简单的语义组合模型。 Hermann与Blunsom提出双语组合语义向量模型（BiCVM），考虑了两种基本的语义组合方式，分别是ADD模型：$f(x) = \sum_{i=1}^{n} \boldsymbol{x}_i$，与BI模型：$f(x) = \sum_{i=1}^{n} \tanh(\boldsymbol{x}_{i-1} + \boldsymbol{x}_i)$ (考虑Bigram信息)。

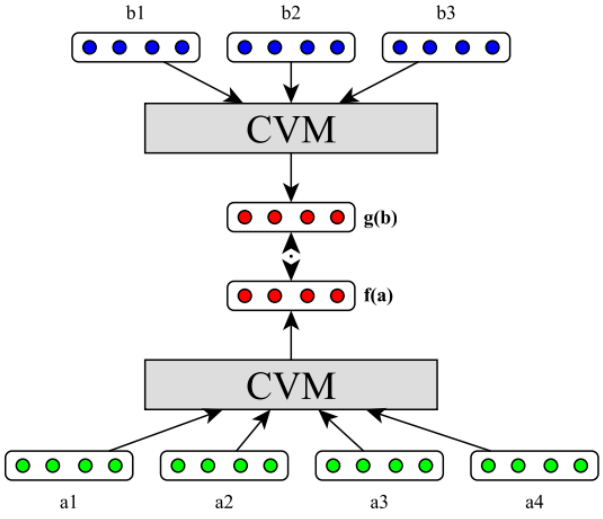

Figure 4: 双语组合向量模型

注意到，双语平行数据对于很多语言对而言都是非常有限的，因此，Gouws等人进一步提出Bilbowa模型 (Gouws et al., 2015)，将单语分布表示学习的优化目标与双语平行约束进行融合，使得模型能够同时利用大规模单语数据以及有限的双语数据。

总体而言，基于句子表示对齐的方法粒度较粗，所得到的跨语言表示在词级别

的语义对齐上表现得并不是很理想。实际上我们完全可以先进行双语词对齐，再利用词对齐信息来改进跨语言词表示的学习过程。我们发表在AAAI 2016的工作中对word2vec中的skip-gram模型进行了扩展并提出multilingual skip-gram模型 (Guo et al., 2016)，基本思想是根据词对齐获取跨语言上下文，进而在skip-gram模型中引入对跨语言上下文的预测，如图5所示。该模型能够自然地扩展至多于两种语言的情形，而且充分利用了单语数据以及双语数据。

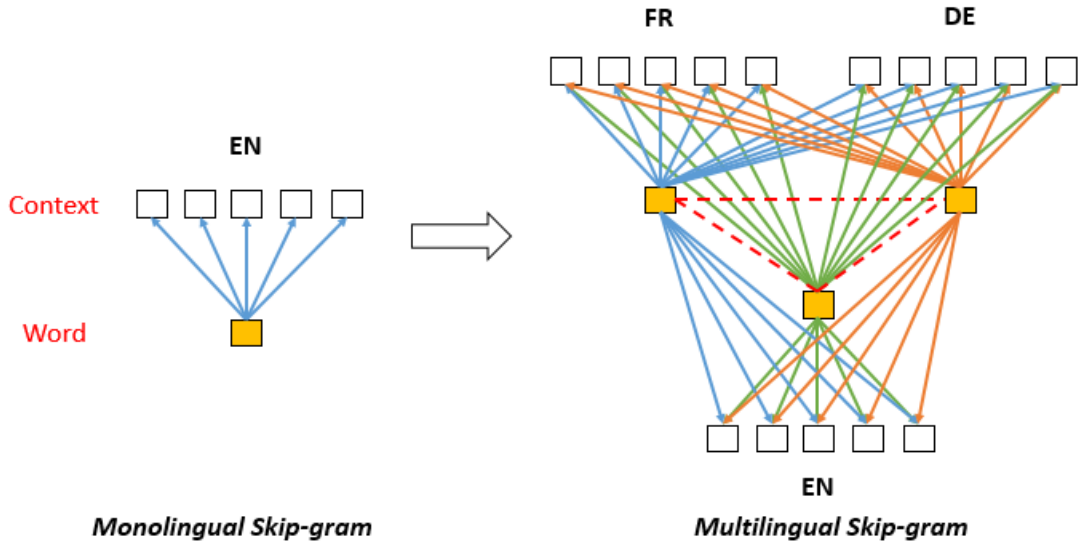

Figure 5: 跨语言skip-gram模型

另一方面，考虑到skip-gram模型与矩阵分解模型的紧密联系 (Levy and Goldberg, 2014)，我们也可以通过对多语言“词–上下文”共现矩阵进行分解来获得跨语言分布表示，Shi等人采取了这种做法 (Shi et al., 2015)。

线下处理与联合学习的方法各有优劣。从资源依赖的角度，线下处理方法使用的是双语词典，而联合学习方法依赖双语平行数据。双语词典既可以从平行数据中通过自动词对齐来获得，也可以从一些在线词典资源中获取，比如PanLex[1]，Wikitionary[2]等；相对而言，对于大部分语言而言，高质量的双语平行数据较难获取。因此，线下处理方法的可扩展性更强。然而也应该注意到，“翻译矩阵”及CCA等线下处理方法对于跨语言映射的线性变换假设不尽合理，这在很大程度上制约了其分布表示学习的质量。联合学习方法则通常不对跨语言分布表示之间的映射关系进行约束，因此更为自由。

[1] https://panlex.org/
[2] https://en.wiktionary.org

2 应用

跨语言数据是一种重要的知识来源，尤其是对于资源稀缺语言而言。跨语言分布表示正好为不同语言之间的知识迁移提供了一座桥梁，使得资源稀缺语言的自然语言处理任务能够充分利用跨语言的标注资源。这种跨语言迁移学习的基本思路是在源语言端构建一个基于分布表示的模型，并使用源语言的特征分布表示进行训练；然后将所得到的模型直接应用于目标语言数据。基于跨语言分布表示的“模型迁移”方法已经被成功应用于文本分类、句法分析、情感分析等任务中。以句法分析为例，使用前文所述的多语言skip-gram模型与跨语言映射的方法相对于“去词汇化特征”模型迁移的结果取得了最高+6.9/+8.9%的LAS提升：

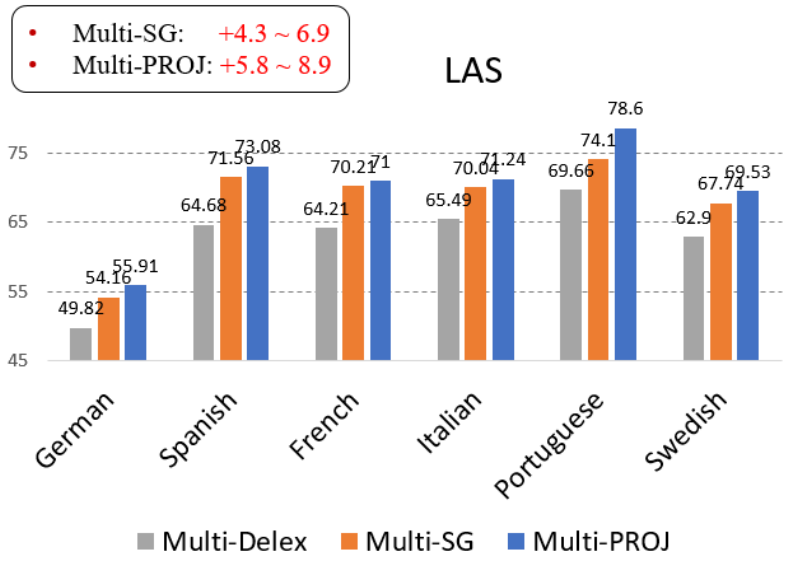

Figure 6: 跨语言依存句法分析上的结果

3 总结与展望

从目前的研究现状来看，不同的跨语言分布表示学习方法在不同自然语言处理任务上的表现各有优劣。比如Bilingual autoencoder模型在跨语言文本分类上表现较好，而对于句法分析则几乎没有帮助。对于情感分析等任务而言，或许情感极性的对齐比词义对齐更为重要。因此，未来的一个研究方向是学习任务相关的跨语言分布表示。另外，目前大部分表现较好的模型仍然严重依赖双语平行资源，而大部分真正的资源稀缺语言的双语数据难以获取。因此，如何充分利用单语数据以及代价相对较低的小规模词典资源来学习高质量的跨语言分布表示，也是一个很有趣的研究思路。

References

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS*

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *EMNLP*

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *ACL-IJCNLP*

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI*

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *NIPS*

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL*

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv*.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL-IJCNLP*

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*