跨语言自然语言处理笔记

Xiachong Feng

1 摘要

跨语言自然语言处理是当下研究的热点。其中，跨语言词向量（Cross-lingual Word Embedding）可以帮助比较词语在不同语言下的含义，同时也为模型在不同语言之间进行迁移提供了桥梁。[Ruder et al., 2017] 详细描述了跨语言词向量学习方法和分类体系，将跨语言词向量按照对齐方式分为了基于词对齐、基于句子对齐、基于文档对齐的方法。其中基于词对齐的方法是所有方法的核心和基础。在基于词对齐的方法中，又有基于平行语料的方法，基于无监督的方法等。近些年，无监督方法成为研究热点。本文主要记录一些跨语言词向量的相关论文。

2 单语言词向量

常用的单语词向量有 Word2Vec，GloVe，fastText 等。下面主要介绍一下 Word2Vec[Mikolov et al., 2013c,a]，Word2Vec 基于分布式假设（Distributional hypothesis）：拥有相似上下文（context）的词语通常拥有相似的含义。其算法分为 Skip-gram 和 Continuous Bag of Words（CBOW）。Skip-gram 根据中心词预测周围的词，CBOW 根据周围的词预测中心的词语，如图1。

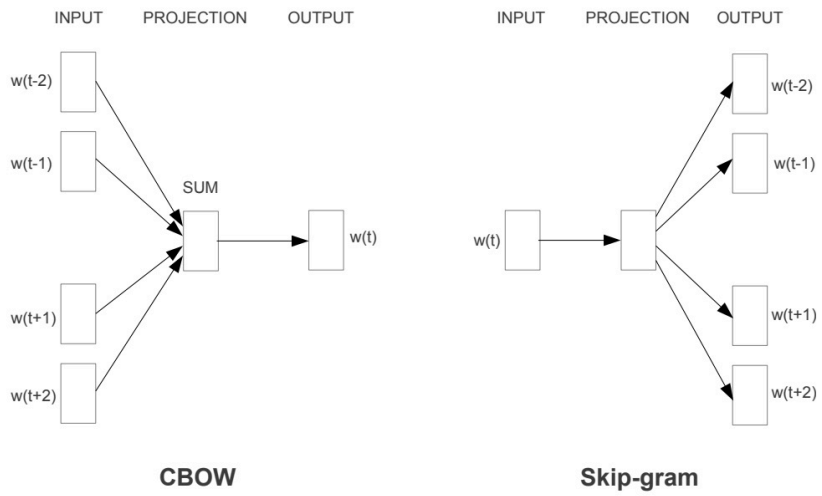

图 1: Word2Vec

一种常见的方法为 Skip-gram + Negative Sampling。简单来说，该算法构造两个向量矩阵，一个 Embedding 矩阵，一个 Context 矩阵。利用 Skip-gram 来构建训练正例，使用 Negative sampling 来构建负例，如图2。训练完成以后（教程可参考The Illustrated Word2vec，Vector Semantics），每个词语对应两个向量，一个 Embedding 矩阵中的表示，一个 Context 矩阵中的表示。最终表示可以直接使用 Embedding 矩阵作为词向量，或者可以将两个矩阵相加得到词向量，或者可以将两个

矩阵拼接得到词向量。

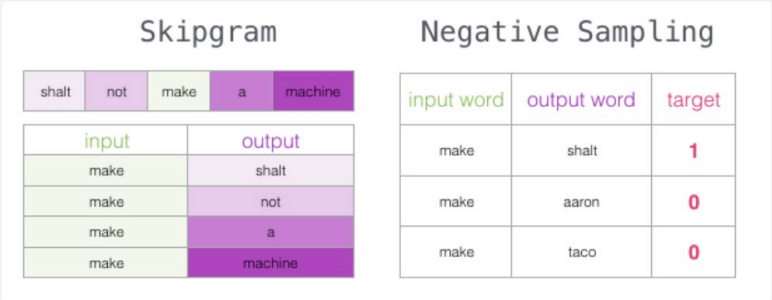

图 2: Skip gram + Negative Sampling，图片来源于The Illustrated Word2vec

3 基于词语映射的方法

[Ruder et al., 2017] 将基于词映射的方法根据映射方法（mapping method）、种子词语的选择（seed lexicon）、映射的改进（refinement）、最近邻词语的检索方法（retrieval）进行了分类。下面简单介绍其中的一些经典工作。

[Mikolov et al., 2013b] 观察发现，不同语言的词向量在向量空间中有着相似的几何排列。如图3。

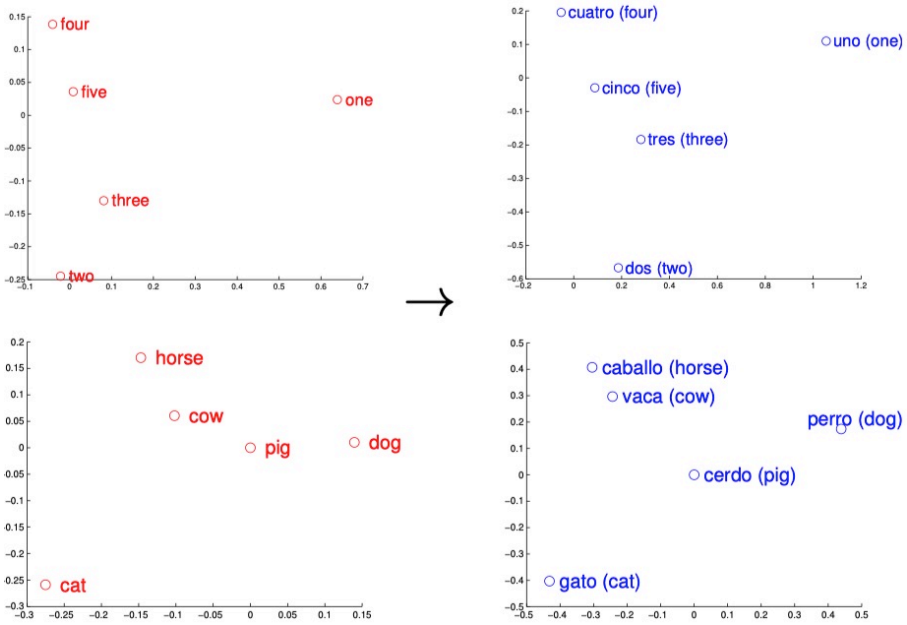

图 3: 英语、西班牙语词向量表示

左图为英语，右图为西班牙语（利用 PCA 进行词向量的降维）。发现，不论是数字还是动物，英语和西班牙语词向量的分布非常相似。基于这一观察，提出了一种简单地线性映射的方法来完成源语言向量空间到目标语言向量空间的转换。该方法的目标在于学习一个从源语言到目标语言的线性映射矩阵（linear transformation matrix）$\mathbf{W}^{s\to t}$，首先从源语言中选择 $n = 5000$ 个频率最高的词语 $w_1^s, \ldots, w_n^s$ 以及其对应的翻译 $w_1^t, \ldots, w_n^t$ 作为种子词语，用于学习线性映射。使用随机梯度下降来最小化均方误差（mean squared error, MSE）$\sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i^s - \mathbf{x}_i^t\|^2$。学习好映射矩阵之后，将

源语言映射到目标语言空间，根据 cosine similarity 来寻找翻译。

[Xing et al., 2015] 发现上述方法有几处不一致。词向量学习的时候使用的是内积（inner product），但是在选择词语的时候却是根据 cosine similarity，学习映射矩阵时，使用的是均方误差（mean square error），这些导致了几处不匹配。因此首先将词向量的长度限制为单位长度。这样相当于所有的向量都会在高维空间落在一个超球面上，如图4。这样就使得两个向量的内积和 cosine similarity 是一致的。然后将目标函数从以均方误差为目标修改为以 cosine similarity 为目标：$\max_W \sum_i (Wx_i)^T z_i$。之前的方法对映射矩阵是没有限制的，这里将映射矩阵限制为正交矩阵（Orthogonal transform），使得其满足 $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$，其实际求解是使用奇异值分解（SVD）来完成，$\mathbf{X}^{t\top}\mathbf{X}^s = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$，$\mathbf{W} = \mathbf{V}\mathbf{U}^\top$。其中 $\mathbf{X}^s$ 为源语言向量矩阵，$\mathbf{X}^t$ 为目标语言向量矩阵。实验证明，该方法的实际效果更好。[Xing et al., 2015, Ruder et al., 2017]。

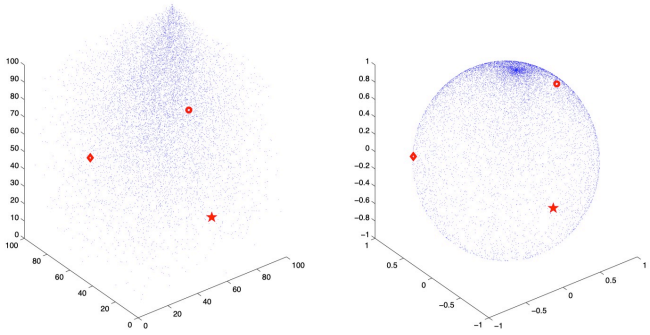

图 4: 未归一化向量和归一化向量

4 基于无监督的方法

之前的方法都是依赖于平行语料的，接下来主要介绍一些无监督的工作，也是当前比较热门的方向。

[Conneau et al., 2017] 提出了一种完全无监督的词级别的翻译（对齐）方法，首先使用对抗训练将两种语义空间对齐，然后使用迭代的方式来一步步更新学习到的映射矩阵，并提出了一种 CSLS 方法来检索最近的翻译词语。如图5。

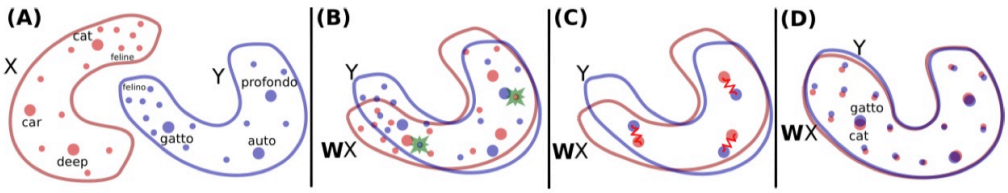

图 5: 无监督词翻译

由于没有对齐信号，所以有一个基本的前提条件是两种语言的词汇处于同一内容空间（碎碎念：FAIR 的无监督机器翻译），这样两种语言的向量空间几何排列才是相似的，才有可能通过映射完成两个空间的对齐，不然是完全没有任何对齐信号的。首先使用对抗训练的方式使得判别器无法区分映射之后的源语言向量和目标语言向量，相当于要求将源语言映射到目标语言语义空间下。判别器的学习目标为尽可能区分映射后的源语言与目标语言：$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{ source } = 1|Wx_i) -$

$\frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(\text{source}=0|y_i)$，映射矩阵的目标为尽可能使得判别器区分错误：$\mathcal{L}_W(W|\theta_D)=-\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(\text{source}=0|Wx_i)-\frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(\text{source}=1|y_i)$。在得到映射矩阵以后，有一个迭代调整的过程，根据学习到的映射，选择互为最近邻的词语作为词典来学习映射，可以迭代这个过程。作者还提出了一种新的相似性度量方式，因为在高维空间中存在一种现象叫做 Hubness，即向量空间中存在密集区域，其中的一些点会是很多点的最近邻。之前的方式采用 cosine similarity 来选择最近邻，作者设计了一种 Cross-Domain Similarity Local Scaling(CSLS) 的度量方式：$\text{CSLS}(Wx_s,y_t)=2\cos(Wx_s,y_t)-r_{\text{T}}(Wx_s)-r_{\text{S}}(y_t)$，其中 $r_{\text{T}}(Wx_s)$，为 Wx_s 和其 K 个目标语言最近邻的平均余弦距离。

基于上述工作，[Lample et al., 2017] 在没有对齐语料的情况下，仅使用单语语料来完成无监督机器翻译。该方法可以很好地泛化到其他语言，并且为有监督的方法提供了性能下限。其 baseline 模型如 [Johnson et al., 2017]。首先使用上述无监督方法得到的翻译词典来初始化翻译模型。接着使用降噪自编码器训练，跨领域训练和对抗训练得到最终模型，如图6。

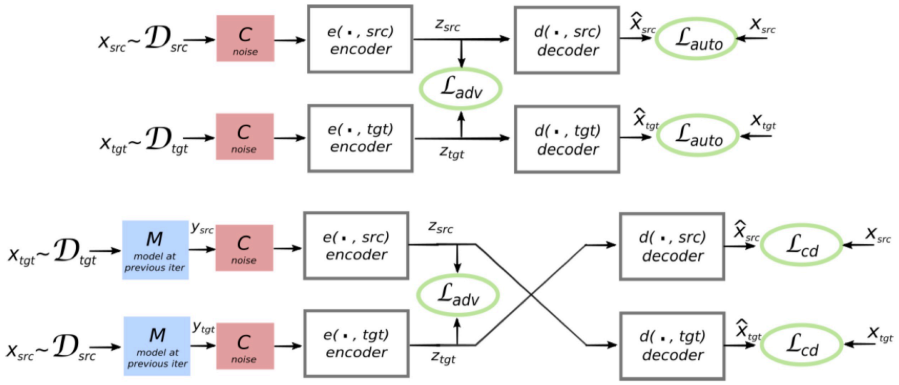

图 6: 无监督机器翻译

降噪自编码器部分，首先从数据集中采样一条数据 x，然后给输入数据引入噪音 $C(x)$，使用编码器对该噪音输入进行编码 $e(C(x),\ell)$，接着使用解码器进行解码 $\hat{x}\sim d(e(C(x),\ell),\ell)$ 得到输出。其损失函数为：$\mathcal{L}_{auto}(\theta_{\text{enc}},\theta_{\text{dec}},\mathcal{Z},\ell)=\mathbb{E}_{x\sim\mathcal{D}_\ell,\hat{x}\sim d(e(C(x),\ell),\ell)}[\Delta(\hat{x},x)]$，其中 Δ 为交叉熵损失。其中噪音模型 $C(x)$ 有两种方式，一种是以一定的概率丢弃每个词语。第二种是打乱输入，但是在文中限制了新的位置距离原本的位置不能超过 k，如图7。

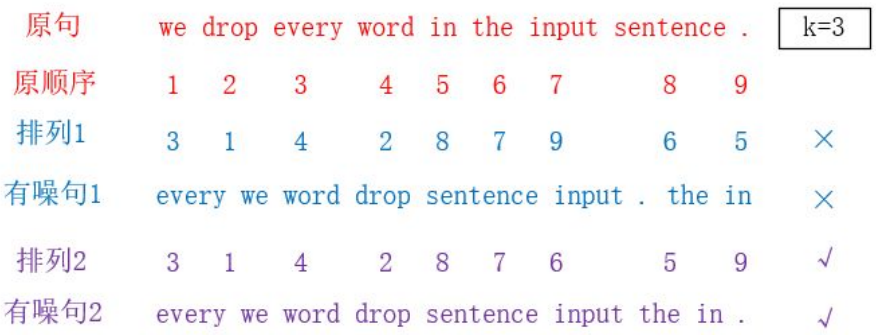

图 7: 轻微打乱句子，图片来源于《Unsupervised Machine Translation Using Monolingual Corpora Only》阅读笔记

第二部分是跨领域训练，这部分是得到翻译模型的关键。利用到了 back translation，首先从 l_1 语

言中采样一个句子，使用当前翻译模型翻译到 l_2 语言下，然后给加噪声 $C(M(x))$，使用 $(C(M(x)), x)$ 作为训练对来训练模型，其损失函数为：$\mathcal{L}_{cd}\left(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \ell_1, \ell_2\right) = \mathbb{E}_{x\sim\mathcal{D}_{\ell_1}, \hat{x}\sim d(e(C(M(x)),\ell_2),\ell_1)}[\Delta(\hat{x}, x)]$。

第三部分为对抗训练部分，希望编码器可以将表示编码到一个语言无关的空间下，其中有一个判别器目前是区分两种语言：$\mathcal{L}_{\mathcal{D}}\left(\theta_D|\theta, \mathcal{Z}\right) = -\mathbb{E}_{(x_i,\ell_i)}\left[\log p_D\left(\ell_i|e\left(x_i, \ell_i\right)\right)\right]$，这部分要更新的参数是：$\theta_D$，编码器的目标是尽可能使得判别器无法区分：$\mathcal{L}_{adv}\left(\theta_{\text{enc}}, \mathcal{Z}|\theta_D\right) = -\mathbb{E}_{(x_i,\ell_i)}\left[\log p_D\left(\ell_j|e\left(x_i, \ell_i\right)\right)\right]$，这部分要更新的参数是 $\theta_{\text{enc}}, \mathcal{Z}$，如图8。对于选择模型的的超参，论文提出了代理准则（surrogate criterion），如公式1，即输入和重构的输入之间的 BLEU 分数。还有一些细节【decoder 如何判断当前生成的语种？在多语言翻译中，通常通过在解码端添加翻译方向的标志位来控制解码方向。但是在本文的假设中，只有非此即彼的两个语种，并且 encoder 对它们一视同仁的。因此，作者只是将两者的解码起始符 <s> 加以区分，各自维护一个。两个训练过程是如何共享同一套 Seq2Seq 框架的？作者所谓的“同一个 encoder 和 decoder”，其实是针对隐层部分而言的。每个语种有自己的 embedding 层和 pre-softmax 层，在模型训练中进行 look-up 来获取各自的参数矩阵。此外，分成“源语言”和“目标语言“是为了便于描述，实际上两者并不区别。最终训练得到的模型，可以在这两种语言中做任意方向的翻译。（碎碎念：FAIR 的无监督机器翻译）】

$$
\begin{aligned}
MS\left(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}\right) = & \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{src}}\left[\text{BLEU}\left(x, M_{src\to tgt}\circ M_{tgt\to src}(x)\right)\right] \\
& + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[\text{BLEU}\left(x, M_{tgt\to src}\circ M_{src\to tgt}(x)\right)\right]
\end{aligned} \tag{1}
$$

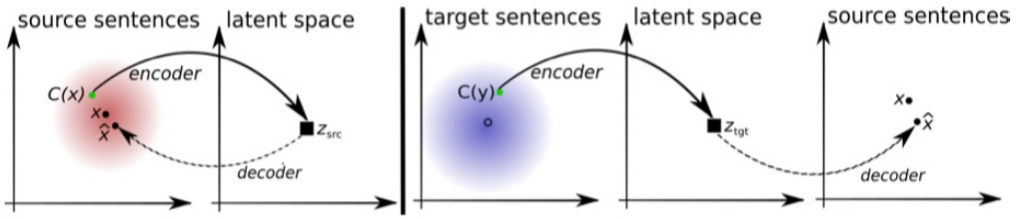

图 8: 降噪自编码器和跨领域训练

[Lample et al., 2018] 指出了 [Lample et al., 2017, Artetxe et al., 2017] 几点特点：使用无监督方法推理出来的词典来初始化系统，使用了基于 Seq2Seq 的降噪自编码器模型，使用 back translation 来将无监督问题转换为有监督问题。同时使用了对抗训练来将不同语言编码到同一空间。本文总结了无监督机器翻译的三个核心点。第一点，初始化，初始化可以帮助模型具有一定的先验知识。第二点，语言模型，根据大规模的单语语料可以学习到好的语言模型。第三点，迭代的反向翻译，该方法可以将无监督转换为有监督，可以完成翻译任务的学习。如图9。

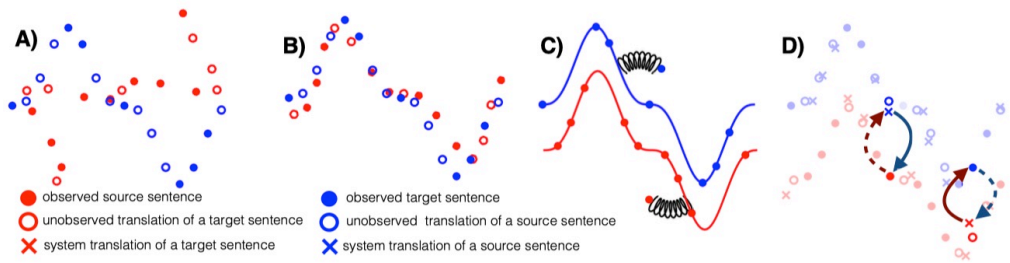

图 9: 初始化、语言模型、反向翻译

对于初始化，本文使用源语言和目标语言的单语语料来共同学习 BPE，学习完成以后用来初始化编码器和解码器的向量查找表。对于语言模型，使用降噪自编码器来学习语言模型。对于反向

翻译，使用迭代的反向翻译来完成翻译模型的学习。该模型同时共享了编码器和解码器的参数，期望学习到共享的语义空间表示。

5 基于虚拟双语语料库的方法

[Xiao and Guo, 2014] 利用 Wikitionary 作为两种语言之间的桥梁，构建了统一的双语词典。首先构建源语言词典，然后利用 Wikitionary 找到其所有的翻译。删除满足以下条件的翻译：一个源语言词语有多个目标语言翻译、一个目标语言词语有多个源语言翻译、源语言的目标语言翻译词语在目标语言数据集中没有出现。经过以上三步处理，可以得到一个一对一的双语词典。将源语言和目标语言建立统一的双语词表 V，利用构建好的双语词典，在词表 V 中属于词典映射关系的两个词语将会被映射到相同的词向量空间。然后利用神经网络来学习词向量表示。其任务是一个二分类问题，输入是一个子句，通过替换正例中的词语来构建负例。最终会学习到统一双语词典的向量表示，以此作为双语空间的桥梁。其模型如图10。这种方法对齐词语有同一表示。

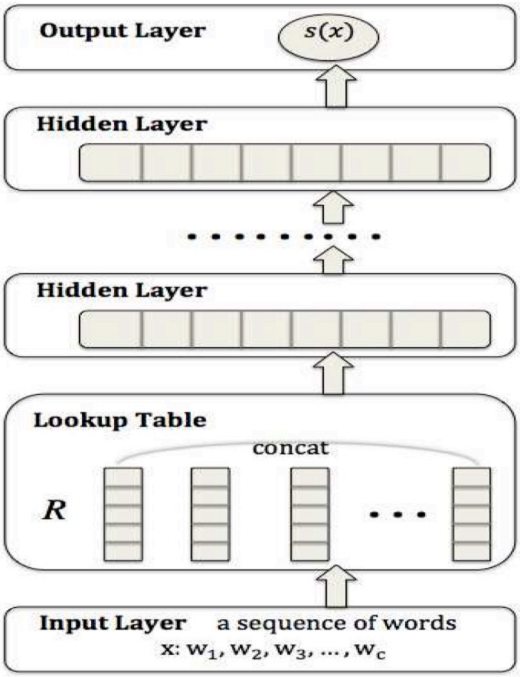

图 10: 神经网络模型学习跨语言表示

[Gouws and Søgaard, 2015] 构建了一种真实的虚拟双语语料库，混合了不同的语言。针对不同的任务可以定义不同的对应等价方法，例如根据翻译，可以定义英语 house 和法语 maison 是等价的，根据词性标注，可以定义英语 car 和法语 maison 都是名词是等价的。因此这里的对齐方式不一定是翻译，可以根据具体的任务来定义，然后利用这种对齐关系来构造双语伪语料。首先将源语言和目标语言数据混合打乱。对于统一语料库中一句话的每一个词语，如果存在于对齐关系中，以一定概率来替换为另一种语言的词语。通过该方法可以构建得到真实的双语语料库。例如根据翻译关系，原始句子 build the house 经过构建可以得到 build the maison，就是将 house 替换为了 maison。利用构建好的全部语料来使用 CBOW 算法学习词向量，由于替换以后的词语有相似的上下文，因此会得到相似的表示。对于那些没有对齐关系的词语，例如“我吃苹果”和“I eat apple”，吃和 eat 没有对齐关系，但如果我和 I、苹果和 apple 有对齐关系，根据构造出来的语料“I 吃 apple”也可以完成吃和 eat 的隐式对齐。这种方法对齐词语有相似表示。

[Ammar et al., 2016] 提出了一种将上述方法扩展到多种语言上的方法 multiCluster。借助双语词典，将词语划分为多个集合，每个集合中是相同语义的词语。然后将所有语言的单语语料库拼接，对于其中的一句话，如果词语在集合中，那就替换为集合中其他语言的词语。得到新的多语语料库以后，使用 skip-gram 来训练得到词向量表示。

[Duong et al., 2016] 提出的方法与上述方法类似，区别在于，只在使用 CBOW 算法学习词向量的时候替换目标词语。而非预先利用词典构造多语语料库。在学习的时候会同时预测源语言目标词语及其对应的替换后的目标词语作为联合训练目标。除此以外，之前的方法都没有处理一词多义的问题，例如 bank 可能有两种意思：river bank 或者 financial bank，对应在意大利语中的翻译就是 sponda 和 banca。因此作者利用上下文词汇表示结合中心词汇表示的方式来选择最合适的翻译词语。通常来说，在 CBOW 算法中，会有两个矩阵，一个 context 矩阵 V，一个 word 矩阵 U。作者指出，使用这种方式训练的词向量，V 矩阵更倾向于单语表示，U 矩阵更倾向于双语表示。其过程如图11。

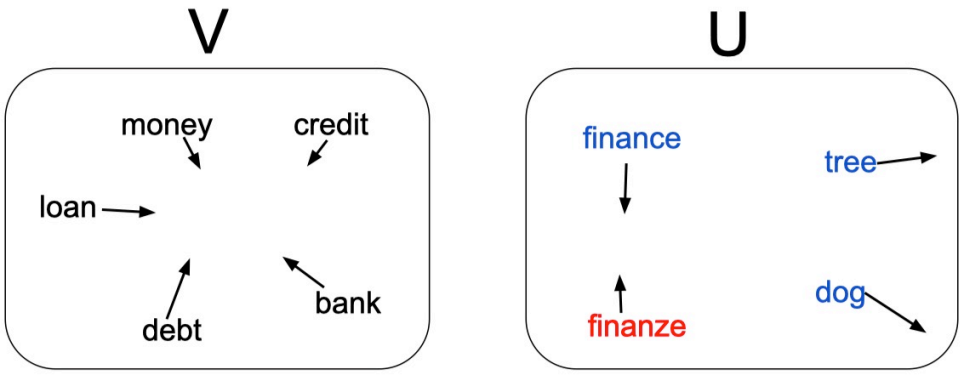

图 11: UV 矩阵，CBOW 目标为预测 finance 和 finanze，tree 和 dog 为负例

6 基于预训练的方法

[Devlin et al., 2018] 提出了 Multilingual BERT，与单语 BERT 结构一样，使用共享的 Word-piece 表示，使用了 104 中语言进行训练。训练时，无输入语言标记，也没有强制对齐的语料有相同的表示。[Pires et al., 2019] 分析了 Multilingual BERT 的多语言表征能力，得出了几点结论：Multilingual BERT 的多语言表征能力不仅仅依赖于共享的词表，对于没有重叠（overlap）词汇语言的 zero-shot 任务，也可以完成的很好；语言越相似，效果越好；对于语言顺序（主谓宾或者形容词名词）不同的语言，效果不是很好；Multilingual BERT 的表示同时包含了多种语言共有的表示，同时也包含了语言特定的表示，这一结论，[Wu and Dredze, 2019] 在语言分类任务中也指出，Multilingual BERT 由于需要完成语言模型任务，所以需要保持一定的语言特定的表示来在词表中选择特定语言词语。

[Lample and Conneau, 2019] 提出了基于多种语言预训练的模型 XLMs，首先从单语语料库中采样一些句子，对于资源稀少的语言可以增加数量，对于资源丰富的语言可以减少数量，将所有语言使用统一 BPE 进行表示。使用三种语言模型目标来完成学习。前两个是基于单语语料库的，最后一个是基于双语对齐数据的。第一种是 Causal Language Modeling (CLM)，根据之前的词语预测下一个词语。第二个是 Masked Language Modeling (MLM)，和 BERT 类似，但是使用一个词语流，而非句子对。第三种是 Translation Language Modeling (TLM)，可以随机 mask 掉其中一些两种语言中的一些词语，然后进行预测。其模型如图12。

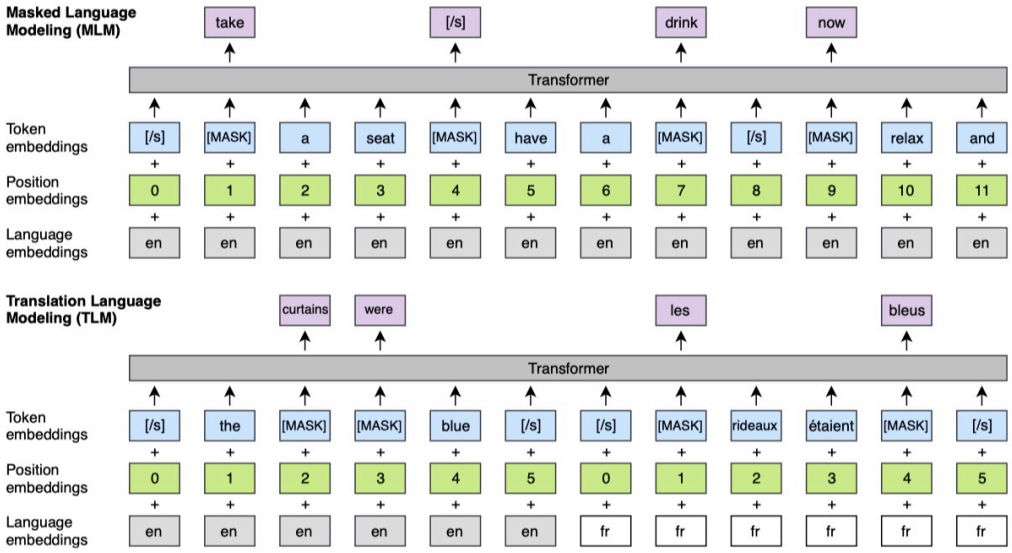

图 12: 跨语言语言模型预训练

7 多语言机器翻译

[Johnson et al., 2017] 使用一个模型来完成多种语言的机器翻译任务。唯一的不同是输入的开始需要拼接一个特殊的指示符，代表目标语言。例如 How are you? -> ¿Cómo estás? 需要修改为 <2es> How are you? -> ¿Cómo estás?，代表该句将被翻译为西班牙语。另一个核心点在于使用共享的 Wordpiece，利用 BPE 来完成。模型在训练的时候，一个 mini-batch 中混合多个语言的平行数据。该模型的优点在于：简单，只需要修改输入数据就可以；可以提升资源稀缺数据的翻译效果；支持直接的 zero-shot 翻译任务。

[Escolano et al., 2019] 利用不同语言之间共有的词表来作为知识迁移的桥梁，提出了两种方法，progAdapt 和 progGrow。第一种方法 progAdapt 将一种语言对的翻译任务迁移到另一种翻译任务上，保留词表中共享的部分，添加新任务的词语，词表大小保持不变，并使用新任务的数据。第二种方法 progGrow 利用递增的方式来学习一个多语言的机器翻译模型，将新语言的词表添加到旧词表上，并使用新旧任务一起的数据。如图13。

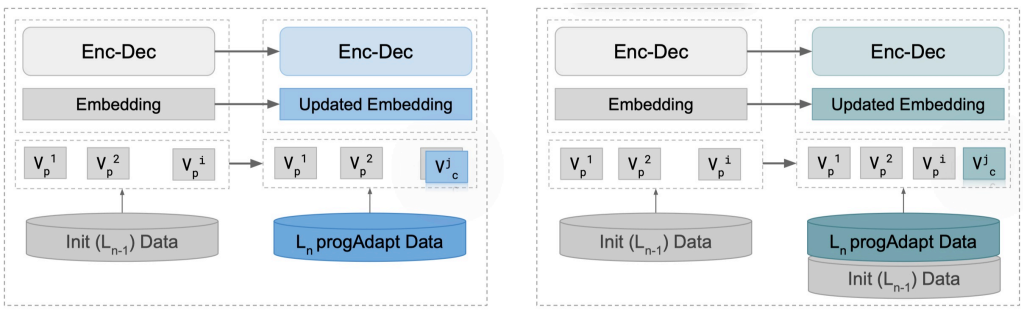

图 13: progAdapt 和 progGrow

[Pires et al., 2019] 指出 [Johnson et al., 2017, Escolano et al., 2019] 的问题在于当语言的词表有显著的不同时，例如中文，词表会变得很大。因此提出了一种方法，每一种语言有自己的特定的编码器和解码器，编码器和解码器之间不共享参数。对于一个翻译对 X-Y，会完成自编码任务（X-X，Y-Y）和翻译任务（X-Y，Y-X），同时会要求编码器得到的两种表示相近。新来一种语言以

后 Z，假设目前有 Z-X 的平行语料，只需要添加 Z 语言的编码器，然后固定住 X 语言的解码器参数来进行训练，这个过程只更新 Z 编码器的参数。如图14。

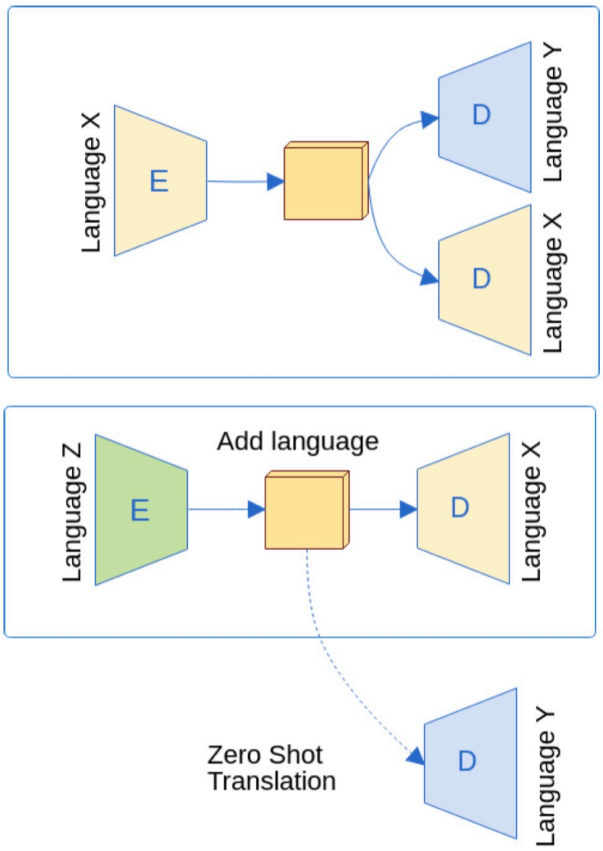

图 14: 多语言机器翻译递增训练

[Kim et al., 2019] 也认为，训练一个共享的多语言机器翻译模型一方面需要语言之间相关，以此来构建一个共享的词表，另一方面当增加一种语言时，如果该语言的词汇不在现有此表中，词表需要更新，模型需要重新训练。因此在多语言机器翻译或者迁移学习的设定下，距离较远的语言词表不匹配（vocabulary mismatch）是一个急需解决的问题。因此提出了一种在向量空间完成隐式翻译的方法，本质上是使用了跨语言词向量。当需要添加一种新的语言 t 时，首先训练语言 t 的单语词向量，然后将已经训练好的机器翻译模型的词向量参数矩阵取出，在两者之间学习一个线性映射 W，用于将新的语言 t 转换到模型的语义空间下，该方法不需要重新更新词表或者重新训练模型，由于在向量空间完成了隐式对齐，当新的语言句子输入以后，会首先通过 W 矩阵来把单语向量空间映射到模型的语义空间，然后接着训练。这种方法虽然确实没有显式的两个词表对齐、增加、替换的过程。但实际上在学习完映射矩阵 W 以后，将新语言的词向量经过映射替换到训练好的模型中，实际上已经隐式的完成了词表的替换，这个映射过后的向量参数矩阵也会随着训练来更新。除此以外，新的语言和原来的语言可能语序不同，因此在训练原机器翻译模型时，会在输入端通过随机插入、删除，交换来引入一些噪音。例如 Ich arbeite hier 通过交换以后变为 Ich hier arbeite。同时由于新语言往往是低资源语言，这里没有使用 back translation 来构建新的语料。而是原来语言数据和新语言数据词表重合的部分保留，其他替换为 unk 来构建伪语料。例如德语数据 Hallo,John! 会变为巴斯克语数据 <unk>,John! 保留了共有部分 John。

[Vázquez et al., 2019] 利用一个语言共享的自注意力机制（attention bridge）来将不同语言编码到同一空间。不同语言的编码器和解码器不共享参数，在使用 LSTM 得到特定语言的表示以后，

使用共享的 attention bridge 得到语言无关表示，用来初始化解码器的初始状态。

8 相关论文

[Liu et al., 2019] 利用一种共享-私有（Shared-Private）词向量来建模源语言词向量和目标语言词向量之间的关系，以及减少模型参数量。其核心想法在于，词向量的一部分是语言无关的，是共享的，另一部分是语言相关的，是私有的。并提出了三种共享关系，相似词语表示（lm）、相同词形（wf）、不相关（ur）。如图15。利用 fast-align 首先根据一定的阈值找到语义对齐的词语。具体实现时，拿源语言词向量矩阵 $\mathbf{E}^x$ 来举例，该矩阵由三个部分构成，$\mathbf{E}^x = \mathbf{E}^x_{\text{lm}} \oplus \mathbf{E}^x_{\text{wf}} \oplus \mathbf{E}^x_{\text{ur}}$，分别代表了三种共享关系词语的表示，每个词语只属于其中一种关系，并按照上述顺序的优先级来排序。其中每一种共享关系由共享部分和私有部分组成，例如 lm 部分，$\mathbf{E}^x_{\text{lm}} = \mathbf{S}_{\text{lm}} \tilde{\oplus} \mathbf{P}^x_{\text{lm}}$，其中 $\mathbf{S}_{\text{lm}}$ 代表语言和目标语言共有的，$\mathbf{P}^x_{\text{lm}}$ 代表源语言私有的。整个实现由矩阵拼接完成。

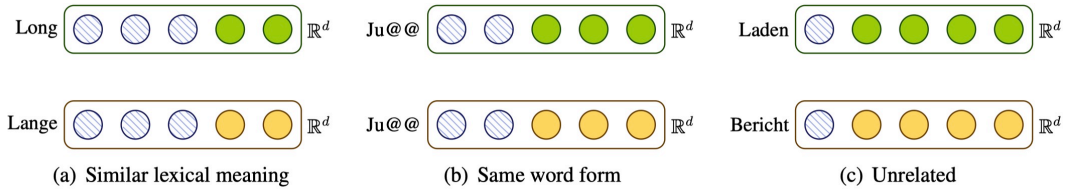

图 15: 共享私有词向量

[Kumar et al., 2019] 利用资源丰富的语言来辅助资源稀少语言的问题生成任务，该任务输入句子，输出问题。并构建了一个新的印度语的问题生成数据集 HiQuAD。其具体做法为：首先使用降噪自编码器（DAE）和反向翻译（back translation）来完成模型的预训练，然后在监督学习部分，分别使用各自数据进行训练。其模型在编码器部分和解码器部分会共享部分参数。其模型如图16。

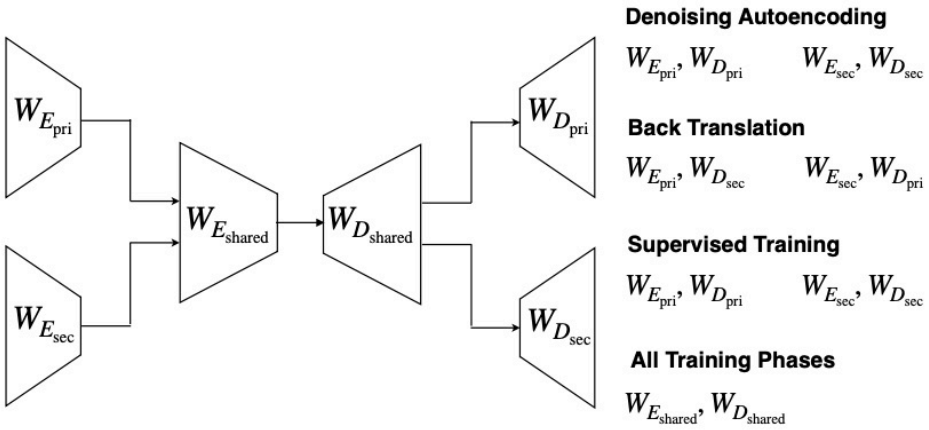

图 16: 问题生成模型

[Duan et al., 2019, Shen et al., 2018] 利用知识蒸馏结合机器翻译来完成跨语言句子摘要任务。其核心想法为使用现有句子摘要数据集训练教师模型，为跨语言句子摘要模型提供监督信号。同时还利用目标输入句作为中间桥梁，来利用两个方向的注意力权重来指导生成。其基本执行流程如图17。

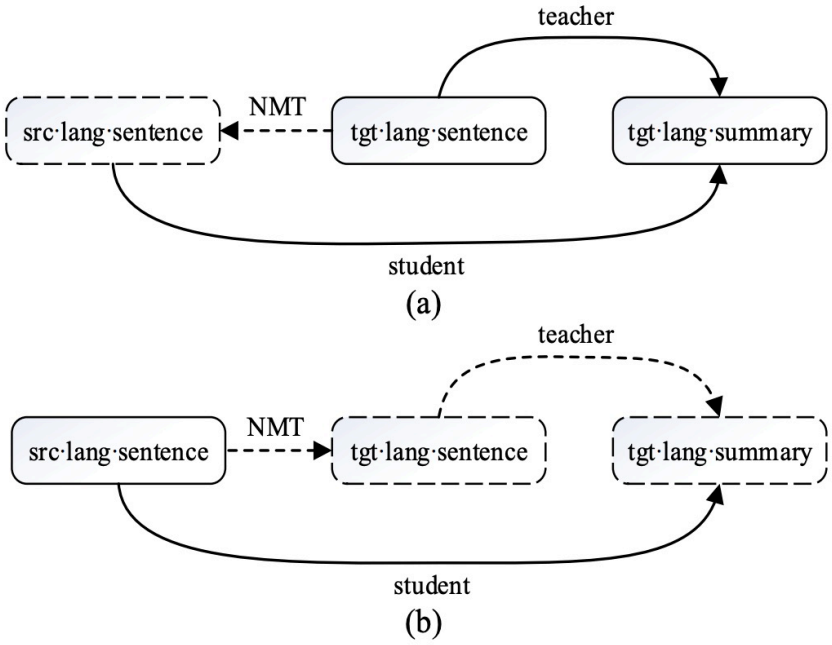

图 17: 句子摘要模型

参考文献

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1305.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*, 2016.

Carlos Escolano, Marta R Costa-Jussà, and José AR Fonollosa. From bilingual to multilingual neural machine translation by incremental training. *arXiv preprint arXiv:1907.00735*, 2019.

Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, 2015.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google' s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Yunsu Kim, Yingbo Gao, and Hermann Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1120.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. Cross-lingual training for automatic question generation. *arXiv preprint arXiv:1906.02525*, 2019.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

Xuebo Liu, Derek F. Wong, Yang Liu, Lidia S. Chao, Tong Xiao, and Jingbo Zhu. Shared-private bilingual word embeddings for neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3613–3622, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1352.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013c.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1493.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*, 2017.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(12):2319–2327, 2018.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4305.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019.

Min Xiao and Yuhong Guo. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, 2014.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015.