目　录

第六章　翻译技巧的自动识别

在本章中，基于在 TED Talks 英法平行语料库中人工标注的实例，我们试验了几种自动分类翻译技巧的方法。尽管此数据集规模较小，但是实验结果验证了我们的工作假设：翻译技巧的自动识别是有可能的。这些实验也为我们指明了未来可继续深入研究的方向。我们的研究重点是基于人工翻译实例的标注，在短语层面对翻译技巧进行自动分类。这种分类可用于评估语言粒度为短语级别的机器译文质量。我们的长远目标是利用这种分类技术，在从双语平行语料库中提取短语级别复述时能够更好地控制语义。此研究也可用于帮助外语学习者了解平行语料库中包含的翻译技巧，详见本书第七章。

6.1　文献综述

首先，我们将详细介绍一些有关人工翻译，尤其是意译的研究。为了找到一种方法来描述为什么翻译是具有难度且非确定性的，Carl and Schaeffer（2017）提出了一个“绝对直译”的假设性定义，指译文在语法和语义层面与源语言完全对等的情况。他们开发了一个计算框架来衡量真实译文的“非直译程度”。理论和经验证据表明，直译更容易，且生成速度更快。在从零翻译或译后编辑的过程中，当译文越偏离“字面直译”的标准时，译文的产出就越困难和耗时。Carl and Schaeffer（2017）提出了一种语料库驱动的经验主义方法来衡量跨语言的语法和语义相似度。他们的多语言语料库翻译过程研究数据库（Translation Process Research DataBase，TPR-DB）包含大量的多版本译文。语法和语义层面的“字面直译”指标指的是在词序和词汇选择上的变化。他们关注翻译所用的时

间长度，假设较短的翻译时间意味着较少的认知投入，并认为翻译中的句式变化与词汇选择的变化相关联，反之亦然。

Tan and Bond（2011）在南洋理工大学多语平行语料库中发现很多运用中文成语来翻译英语非固定表达的情况。为识别中文译文中的成语，Ho *et al.*（2014）创建了一个包含 4000 个四字成语的词库，以标记南洋理工大学多语平行语料库中出现的所有成语。该词表的扩展是通过人工修订实现的，他们的最终目标是将其纳入中文 Wordnet 资源中。

Poirier（2014）提出了一种用于检测内容层面翻译错误和翻译偏离的半自动算法。他认为“翻译偏离”的产生原因是因为译文需要符合目标语言规范，由此译者可以使用不同数量的实词和不同结构来表达同一语义。我们可以发现，如果源语的结构在译文中被原封不动地保留，那译文有时就会不规范或不流畅。Poirier（2014）提出的算法需要一对句子作为输入，并根据为每种语言预先制定的列表删除语法虚词，算法将比较两种语言余下的实词数量。随后，作者手动定位翻译错误或翻译偏离，并为词或片段手动关联语义单元。这项研究可以向译者指出翻译中的潜在错误或偏离，方便译者进行更仔细的检查。

质量评估（Quality Estimation，QE）任务的目的是在无法进行自动或人工评估时，预测机器翻译的质量。该任务通常在翻译系统的运行过程中执行，在不依靠参考译文的情况下，在各种翻译单位（单词、短语、句子、文档）上进行评估（Specia *et al.*，2010）。质量评估的潜在应用是多方面的，例如，预测职业译员的译后编辑工作量（Specia, 2011），使终端用户了解翻译质量，从多个系统生成的选项中选择最佳译文，等等。

关于在句子层面开展翻译质量评估，Blatz *et al.*（2004）首先将其视为一个二元任务。Specia *et al.*（2009）提出了一个用于评估连续数值的回归任务，其结果更适用于评估译后编辑所需的时间。在第一个关于质量评估的共享任务中，Callison-Burch *et al.*（2012）提出了两项子任务：对基于短语的统计机器翻译系统生成的译文进行排序和为它们分配分数。Wisniewski *et al.*（2013）对该实验的数据集和结果进行了详细分析。围绕质量评估的后续共享任务增加了评估由神经机器翻译系统生成的译文，研究的粒度为单词、短语、句子和文档（Specia *et al.*，2018）。

对于在句子层面的翻译质量评估，以 2019 年的共享任务为例，系统需要根据译后编辑的工作量，通过必须进行编辑的人工翻译编辑率（Human-Targeted Translation Edit Rate，HTER）对句子进行评分。在单词层面上，参与者需要识别每个词元的翻译错误以及未被翻译的词元。在短语（或单词）层面，我们发现对非直译质量的评估还没有被考虑在内。

近年来，人们提出了不同的模型来自动检测平行语料中的翻译差异，其目的是自动筛选掉语义不对等的句子对（如包含翻译错误、句子对齐问题、漏译源语片段、在译文中采用明晰化翻译技巧等），以提高训练机器翻译系统的语料质量。Carpuat *et al.*（2017）提出了一个使用词对齐信息和句子长度特征、基于支持向量机的跨语言差异检测器。Vyas *et al.*（2018）提出了一种基于深度神经网络的方法，其训练不需要人工标注数据。Pham *et al.*（2018）以无监督的方式，根据单词间相似度生成句子向量，然后衡量句子间的语义对等度，以指导筛选平行句对。

与我们的研究比较接近的一项任务是识别文本语义蕴含（Dagan *et al.*，2005）。文本蕴含的定义是：对于自然语言中的两个句子（一个前提 P 和一个假设 H），如果某人阅读 P 后可以推断出 H 可能是真的，那么这两者之间存在文本蕴含关系。

在单语文本蕴含关系识别任务（Recognizing Textual Entailment，RTE）中使用的特征已被用于改善机器翻译质量评估（Padó *et al.*，2009a，2009b）。通过建模机器译文和人工参考译文之间的（非）对应关系，Padó 等能够区分保留语义的不同形式的译文与不正确的译文。

Mehdad *et al.*（2010）提出将单语 RTE 框架扩展为一项跨语言文本蕴含关系任务（Cross-Lingual Textual Entailment, CLTE）。他们组织了两项运用 CLTE 进行多语言内容自动同步的共享任务（Negri *et al.*，2012，2013）。对于给定的跨语言句对，参与者需要识别出多方向的蕴含关系（前向、后向、双向、无蕴含关系）。对于两个用不同语言编写的关于同一主题的文档，该任务的应用是自动检测和解决它们所含信息的差异，以生成信息对齐或相互补充的文档。

Upadhyay *et al.*（2018）提出了运用无监督方法识别跨语言上位词，这是一种细粒度的不对称蕴含关系。他们的分布式方法使用了一个在小型双语词典上习得的跨语言词向量模型，以及从依存句法分析语料库中

提取出的多样化单语句法语境。不过，他们通过众包方法构建的跨语言数据集是由无上下文语境的词对组成的。

在本书中，我们在单词或短语层面开展研究。我们选择这种粒度而不是整个句子，是因为它们在语料库中更容易被重复识别和使用。我们的自动分类依赖于人工标注的数据，而不是像已有研究一样多采用人工合成的数据。

与在句子层面作出二元决定（好的翻译或误译）的研究相比，我们感兴趣的是更细致的短语层面，以及使用不同翻译技巧所生成的优质翻译（尤其是非直译）。我们开展了二元和多元分类实验。有些翻译技巧会带来语义或句法变化，但并不构成对翻译质量产生负面影响的翻译偏离。

6.2　数据集

表 6.1 展示了本实验中每个类别下的实例数量。在语料库中，我们标注了“调适 + 置换”这一类别，因为它代表两种翻译技巧相结合的现象（两种技巧的重要性不相上下）。然而，该类别在标注语料中只出现了 53 例。因此，我们在实验中将“置换”和“调适 + 置换”合并为一个名为“包含置换”的类别，即淡化其中的“调适”技巧。此外，“修辞翻译”技巧在本语料库中出现次数很少，于是我们在实验中先暂时将其忽略。由于数据集的各类别分布不均衡，我们将在不同的实验条件下评估分类器。

表 6.1　每类翻译技巧所含实例数量

直译	3771	直译（3771）
对等	289	非直译（1127）
包含置换（置换、调适 + 置换）	342	
调适	195	
宽泛化	86	
具体化	215	

对于英语和法语语料库，我们使用 Stanford Tokenizer 工具进行词切分。该工具也在开始标注之前被用于预处理原始语料。我们对所有单词进行了小写化处理。英语词形还原使用 Stanford CoreNLP 工具（Manning *et al.*，2014），法语词形还原使用 Tree Tagger 工具（Schmid, 1995）。在词形还原过程中，我们保留了 Stanford Tokenizer 的词切分结果，以方便后续提取特征。

我们在一个简化后的场景下进行实验，即已知双语片段边界，只预测使用的翻译技巧。例如，给定 deceptive → une illusion，我们的目标是预测“包含置换”标签。

6.3 传统机器学习分类器

由于能用作交叉验证的实例数量比较有限，我们主要致力于研究设计和抽取特征，并使用 Scikit-Learn 工具包（Pedregosa *et al.*，2011）训练不同的统计学习分类器，同时也测试了基于神经网络的分类器（见本书 6.4 小节）。

我们首先研究了与语境无关的特征。在分析语料特点之后，我们针对英语–法语语料提出了在复杂程度和所需资源方面各有不同的五组特征：文字表面、词形–句法分析、句法分析、使用外部资源、自动词对齐。

此外，英法两种语言用于词形–句法分析、成分句法分析和依存句法分析的标签集被转换为三个统一的、紧凑的标签集（Petrov *et al.*，2012；Leung *et al.*，2016）。

6.3.1 与语境无关的特征

6.3.1.1 文字表面

我们计算出英语和法语的词元数量（分别记为 le 和 lf）、这些数量的比值（le /lf , lf /le),以及两个片段之间字符的 Levenshtein 距离（Levenshtein, 1966）。由于法语和英语的一些同源词拼写很相似（Hammer and Monod, 1976），一对互为直译的单词或片段之间可能有一个很小的 Levenshtein 距离。

6.3.1.2　词形–句法分析

我们使用 Stanford CoreNLP 工具对英语、法语两种语言进行词形–句法分析。对于每种语言，每个词性标签出现的次数被计入一个向量中（总共有 17 个两语种通用的词性标签），这两个向量被用作特征（见表 6.2）。我们计算向量之间的余弦相似度，计算情况分两种：一是保留所有单词；二是只保留句中实词[1]。

表 6.2　两个向量分别记录每个语种中每个词性标签出现的次数

语种	ADJ	DET	NOUN	...	INTJ
英语	1	0	0	0	0
法语	0	1	2	0	0

针对“包含置换”类别的实例，我们手动构建了大约 50 个词性序列变化模板（Hunston and Francis, 2000），来验证一个新的实例是否匹配其中一个模板（我们在语料中发现，“包含置换”类别中有 48% 的实例匹配这些模板）。例如：

methodologically → de façon méthodologique (ADV → ADP NOUN ADJ)

vision impairment → l’altération visuelle (NOUN NOUN → DET NOUN ADJ)

are increasing rapidly → est en augmentation rapide (VERB VERB ADV → VERB ADP NOUN ADJ)

6.3.1.3　句法分析

法语和英语成分句法分析分别是通过 Bonsai（Candito *et al.*，2010）和 Stanford CoreNLP 工具完成的[2]。这是一个二元特征，包括三种情况：第一，如果是一对单词，我们比较它们的词性标签；如果标签相同，该

1　实词标签包括 ADJ、ADV、NOUN、PROPN、VERB。如果某片段内不包含任何实词，那我们使用该原始片段。

2　在法语成分句法分析方面，Bonsai 比 Stanford CoreNLP 效果更好，而且在非终端节点出现的明显错误更少。

特征取值为 0，否则为 1。第二，如果是一对短语，我们比较它们的非终端节点标签；如果标签相同，该特征取值为 0，否则为 1。第三，如果是一个词被译为一个短语，或者反之，我们比较其标签的类别。例如，如果一个形容词被译为一个动词性短语，特征取值为 1；如果标签类别相同，特征取值为 0。

为共享同一标签集，英法两种语言的依存句法分析都使用 Stanford CoreNLP 工具。在片段内部，我们计算每种依存关系出现的次数（见图 6.1 和表 6.3）。

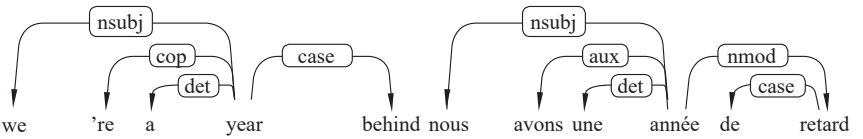

图 6.1　片段内部的依存语法分析

表 6.3　计数两种语言中的依存语法关系标签

语种	amod	det	nmod	case	...	nsubj
英语	0	1	0	1	...	1
法语	0	1	1	1	...	1

在图 6.2 中，英语动词 adapt 被译为法语名词形式 adaptation，同时在短语外部，ability to 和 capacité d’ 与上述两词之间具有依存关系。在通过依存关系链接的外部词中，我们只保留标注者手动对齐过的单词。由此我们计算如图 6.2 所示的依存语法关系。

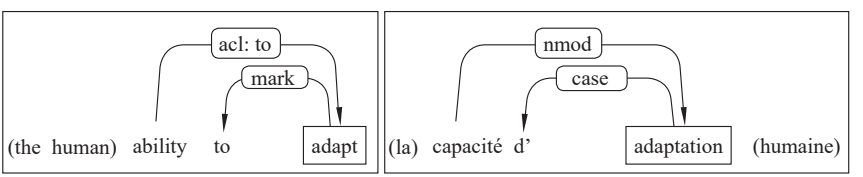

图 6.2　片段外部的依存语法分析

6.3.1.4　使用外部资源

我们利用了多语言资源 ConceptNet，这是一个可开放访问的语义网络，旨在让机器理解单词的含义。该资源的构成包括：从众包渠道获取的知识、电子词典、严肃游戏以及链接开放数据（Linked Open Data）。基于 ConceptNet 的系统在 SemEval2017 组织的“多语言和跨语言词汇语义相似性计算”任务中获得第一名（Camacho-Collados *et al.*，2017；Speer and Lowry-Duda, 2017）。

为计算双语向量之间的余弦相似度，我们使用了 ConceptNet Numberbatch 资源（Speer *et al.*，2017）。该资源中的向量是基于 ConceptNet 数据和 Word2Vec（Mikolov *et al.*，2013）以及 Glove（Global vectors for word representation）（Pennington *et al.*，2014）构建的。Faruqui *et al.*（2015）利用语义词典中的关系信息调整了该向量空间内的表征，使词典中语义相关联的词拥有近似的向量表示。

我们可以在这个资源里找到某些多词固定表达式的整体向量。如果某表达式未被收录进该资源，我们只计算该表达式中实词的向量平均值。在经过词形还原后的表达式上，我们也再次如此计算。

ConceptNet 资源还提供三元组形式数据，即一对由某种关系连接起来的单词或词组。基于这些数据[1]，我们设计了一个特征（取值为整数）来验证一对英语–法语表达是否属于如下情况：1）存在直接关联；2）通过另一个法语单词或短语存在间接联系（如 complete ← complet /entier → total）；3）完全没有关联。

我们选取了原始形式、经过词形还原后的形式以及经过词形还原及词汇筛选后的形式[2]三种文本形式进行测试。例如：

> 原始形式：increasing rapidly → en augmentation rapide
>
> 经过词形还原后的形式：increase rapidly → en augmentation rapide
>
> 经过词形还原及词汇筛选后的形式：increase rapidly → augmentation rapide

1　本研究使用了英语–法语和法语–法语的三元组数据。

2　我们根据一个手动制作的列表来筛选单词，其中包括无实际语义的动词、冠词、代词等。

在上述第三种文本形式上，我们计算出在 ConceptNet 资源中间接关联的双语词元数量，然后将其除以两种语言的词元总数。例如，deceptive 和 illusion 在资源中不存在直接关联，但两者都与 illusoire 一词关联，因此两词之间存在间接联系。这种情况经常发生在运用置换技巧的译文中。

6.3.1.5 自动词对齐

对于这组特征，我们使用了 Berkeley Word Aligner 工具提供的词汇翻译概率表（Liang *et al.*，2006）。为此，我们将本研究构建的全部 TED Talks 语料（163,092 行）和一部分 Paracrawl 语料合并为一个英法平行语料库（总计 180 万对句子，4100 万个英语词元），在此基础上训练 Berkeley Word Aligner 工具。

词汇翻译概率分布的熵是按照如下公式计算的（Gray，1990；Carl and Schaeffer，2017），这些值由 Berkeley Word Aligner 直接提供：

$$H(x) = \sum_{i} P(x_i)I(x_i) = -\sum_{i} P(x_i)\log_e^{P(x_i)}$$

我们为实词计算平均熵。熵值越高，说明该词的含义越宽泛或者为多义词。我们为经过词形还原的实词计算了同样的特征。同时，我们根据 Koehn *et al.*（2003）提出的公式为实词计算了双向词汇权重。由于我们为非直译片段进行了多对多词对齐，我们假设两个片段（e 和 f）之间词对齐的形式为每个源语单词与所有目标语单词都呈对齐关系，并将所有的对齐称为集合 *A*，计算公式为：

$$lex(e|f, A) = \prod_{i=1}^{length(e)} \frac{1}{|\{j|(i, j) \in A\}} \sum_{\forall(i, j)\in A} w(e_i|f_j)$$

为计算英译法方向的词汇权重 lex(e|f, A)，每个英语词元 e_i 被视为是与其对齐的法语词元 f_j 生成的，且词汇翻译概率为 $w(e_i|f_j)$。法译英方向词汇权重 lex(f|e, A) 的计算方式与之同理。在经过词形还原的实词上，我们计算了同样的特征。该特征值越大，双语片段之间词对齐的可信度就越高。

我们还计算了在统计上意义最可能的翻译（根据词汇翻译概率表得出）和语料库中人工翻译之间的词汇翻译概率差之和。对于每个源语词，我们在人工翻译中选取词对齐概率最高的目标语单词。例如，对于 alternatives → solutions de remplacement，最直白的法语翻译是单词 alternatives，在概率表中数值为 0.4。在译者给出的短语翻译 solutions de remplacement 中，solutions 与 alternatives 的对齐概率最高，但概率值仅为 0.07。我们为每个源语单词将这种概率差进行相加，将不具有最高概率的目标语单词视为未对齐词（比如这里的 de remplacement），并计算这些词分别占两侧词元总数的比例（这些特征是在双向翻译上进行计算的）。

6.3.2　实验设计和结果分析

6.3.2.1　实验设计

由于非直译翻译的实例数量（1127）不及直译实例数量（3771）的三分之一，我们在以下三种情况下对分类器进行评估。第一，六个类别（直译、对等、宽泛化、具体化、调适、包含置换），其中直译类数量分为两种情况：一包含所有实例，二随机抽取 200 个直译实例以使各类别达到近似均衡分布。第二，两个类别（直译和非直译），并通过随机抽取直译实例构成三种比例：3∶1，2∶1，1∶1。第三，五个类别，只有非直译实例。

我们训练了 RandomForest、Multi-Layer Perceptron、Logistic Regression、Support Vector Machine、K-nearest Neighbors、Decision Tree、Bernoulli Naive Bayes、Multinomial Naive Bayes 和 Gaussian Naive Bayes 等分类器，对它们的超参数进行了优化[1]。分类器的评估是通过五折交叉验证进行的（我们使用了 StratifiedKFold，其优点是在构建每折数据集时可以保留各类实例的原始比例）。我们采用的衡量标准包含平均准确度、微观平均和宏观平均 F1 值（Tsoumakas *et al.*，2011）。表 6.4 列出了不同实验条件下 RandomForest 的分类结果，其中 Dummy 分类器被用作基线，它总是预测数据集中占比最多的那个类别。实验结果显示，在所有的实验条件下，RandomForest 分类器的表现最佳。

1　为找到最佳超参数，10% 的数据被分离出来作为测试集，然后在剩余的训练数据上进行三折交叉验证。

表 6.4 使用所有特征在不同实验条件下获得的结果

类别分布	分类器	平均准确率	微观 F1 值	宏观 F1 值
六类				
六类，其中包含 3771 例直译	Dummy	76.99%	0.77	0.14
	RandomForest	**83.10%** ± 0.35%	**0.83**	**0.44**
六类，其中包含 200 例直译	Dummy	25.77%	0.26	0.07
	RandomForest	57.04% ± 1.47%	0.57	0.52
两类				
直译（3）：非直译（1）	Dummy	76.99%	0.77	0.43
	RandomForest	**90.16%** ± 0.98%	**0.90**	**0.86**
直译（2）：非直译（1）	Dummy	66.67%	0.67	0.40
	RandomForest	88.85% ± 0.71%	0.89	0.88
直译（1）：非直译（1）	Dummy	50.00%	0.50	0.33
	RandomForest	87.09% ± 2.50%	0.87	0.87
五类				
非直译	Dummy	30.35%	0.30	0.09
	RandomForest	55.10% ± 1.45%	0.55	0.47

6.3.2.2 结果分析

我们首先进行最为直接的六类别分类。由表 6.4 中可知，RandomForest 分类器的结果大幅度优于 Dummy 分类器。另外，当各类别的实例数量近似均衡时，多类别分类任务的难度体现了出来（类别数量多，但每类实例数量有限）。因此，我们接下来研究了二元分类（直译与非直译），以及在五个非直译类别间进行分类。

对于二元分类，两个表现最好的分类器是 RandomForest 和 Multilayer Perceptron。此外，RandomForest 比通过硬投票（hard voting）或软投票（soft voting）将这两种分类器结合起来的效果更好。在数据量均衡分布的情况下，最高准确率达到 87.09% ± 2.50%。从自然分布（3∶1）过渡到均衡分布（1∶1），非直译类别的平均 F1 值从 0.78 增加到