机器翻译调研报告

### 问题定义

（介绍问题的应用场景、定义）

**定义：**

机器翻译（Machine Translation, MT）是用计算机把一种语言（源语言，source language）翻译成另一种语言（目标语言，target language）的一门学科和技术。

**应用场景[1]：**

1）在线翻译：谷歌、微软、百度、有道、搜狗等。

2）翻译机：从出国旅行，到国际文化交流，再到对外贸易，语言障碍是一个天然痛点，目前很多翻译类的产品将机器翻译和OCR（光学字符识别）技术以及语音识别技术进行结合，可以实时的通过摄像头来翻译外文指示牌、菜单、说明书等，也可以结合语音技术进行对话翻译，从而实现不同语种的无障碍交流。

3）语音同传技术：同声传译广泛应用于国际会议等多语言交流的场景，但是人工同传受限于记忆、听说速度、费用偏高等因素门槛较高，搜狗推出的机器同传技术逐步开始在会议场景出现，演讲者的语音实时转成文本，并且进行同步翻译，低延迟显示翻译结果，希望能够取代人工同传，实现不同语言人们低成本的有效交流。

4）跨语言检索：中文资讯只占世界信息的10%，而跨语言检索需求逐年增加，搜狗推出的海外搜索将机器翻译和信息检索技术进行了结合，不论用户输入中文还是英文，系统都会从海量优质的英文网页中选出用户想要的搜索结果，并应用国际领先的机器翻译自动将其进行翻译，为用户提供英文原文、中文译文、中英双语三个页面的搜索结果。

5）AI助力翻译行业升级：机器翻译和传统翻译行业相结合，利用机器翻译提升传统翻译行业的效率，提升商业价值。

### 评价方式

（介绍衡量算法好坏的常用评价指标）

**主观评测：**

流畅度、充分性、语义保持性

**客观评测：**

1）句子错误率：错误句子占全部译文的比率。译文与参考答案不完全相同的句子为错误句子。

2）单词错误率：分别计算译文与每个参考译文的编辑距离，以最短的为评分依据，进行归一化处理。

3）与位置无关的单词错误率：不考虑单词在句子中的顺序。

4）METEOR评测方法：对候选译文与参考译文进行词对齐，计算词汇完全匹配、词干匹配、同义词匹配等各种情况的准确率P、召回率R和F平均值。

5）BLEU评价方法[2]：统计同时出现在系统译文和参考译文中的n元词的个数，最后把匹配到的n元词的数目除以系统译文的n元词数目，得到评测结果。

修正的n元语法精度计算：Countclip = min(Count, Max\_Ref\_Count)

考虑句子长度对评分的影响：

6）NIST评测方法

### 常用方法

#### 传统做法

* + 1. **基于规则的机器翻译[3]（Rule-based machine translation: RBMT）**

**1）直接翻译法**

从源语言句子的表层出发，将单词、短语或句子直接置换成目标语言译文，必要时进行简单的词序调整。对原文句子的分析仅满足于特定译文生成的需要。这类翻译系统一般针对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。

**2）转换翻译法：**

1）对源语言句子进行词法分析；2）对源语言句子进行句法/语义分析；3）源语言句子结构到译文结构的转换；4）译文句法结构生成；5）源语言词汇到译文词汇的转换；6）译文词法选择与生成。

**评价：**

优点：可以较好地保持源文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

弱点：规则一般由人工编写，工作量大，主观性强，一致性难以保障，不利于系统扩充，对非规范语言现象缺乏相应的处理能力。

1. **中间语言法**

这种方法将源语言转变成统一的一种中间语言表达形式，然后，中间语言可以转换成任何目标语言。

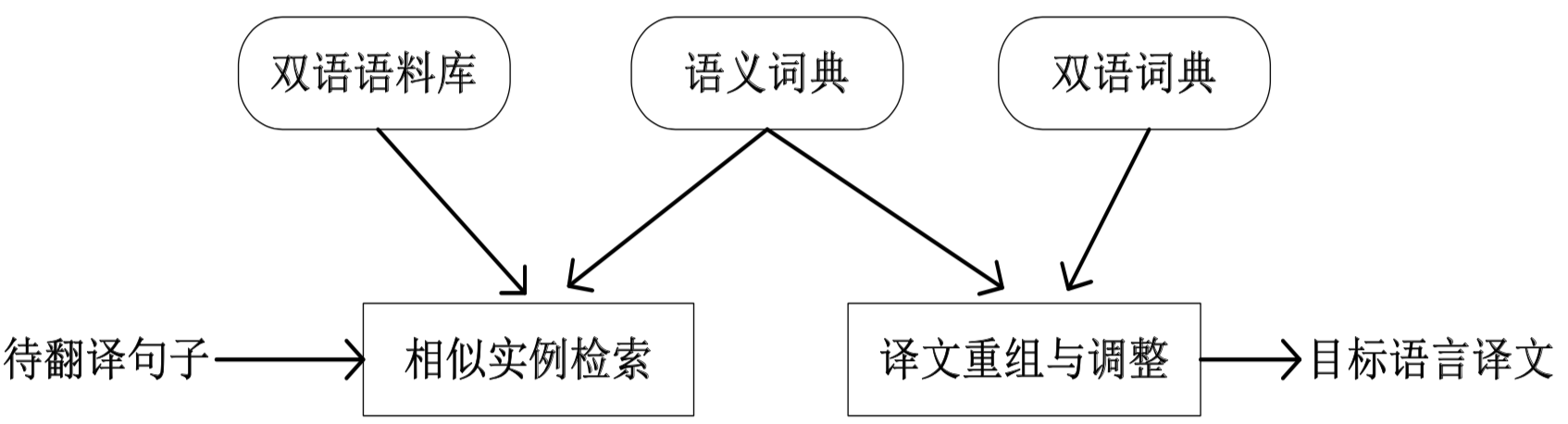
**评价：**

优点：中间语言法的语言规则针对每种语言和中间语言，而与互译的语言对无关。因此，该方法尤其适合多语言之间的互译。

弱点：如何定义和设计中间语言的表达方式，以及如何维护并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面，都面临若干困难。

* + 1. **基于实例的机器翻译（Example-based machine translation: EBMT）**

输入语句→与实例相似度比较→翻译结果，找出两个句子的不同之处，然后翻译有差异的单词，但不破坏句子的结构。



**评价：**

优点：不要求源语言句子必须符合语法规定，翻译机制一般不需要对源语言句子做深入分析。

弱点：两个不同的句子之间的相似性（包括结构相似性和语义相似性）往往难以把握，尤其在口语中，句子结构一般比较松散，成分冗余和成分省略都较严重，这更增加了分析句子与实例句子的比较难度。另外，系统往往难以处理实例库中没有记录的陌生的语言现象，而且当实例库达到一定规模时，其实例检索的效率较低。

* + 1. **统计机器翻译[3, 4, 5]（Statistic machine translation, SMT）**

**1）基于单词的SMT[6, 7]（生成模型）**

**基本原理：**

源语言句子：，目标语言句子：，贝叶斯公式：，。机器翻译=语言模型✖️翻译模型

1. IBM翻译模型1：仅考虑词与词的互译概率，缺乏对单词顺序的理解。
2. IBM翻译模型2：在句子中考虑单词的顺序，它记住了单词在输出句子中通常所处的位置，并在中间步骤重新排列单词。
3. IBM翻译模型3：考虑单词的繁衍能力，如果机器觉得有必要添加新词的时候，先插入NULL标记，然后为每个标记词选择合适的语法助词或单词。
4. IBM翻译模型4：考虑单词的相对顺序，即该模型学习了两个单词是否经常互换位置。
5. IBM翻译模型5：在功能上没有创新，只是为学习增加了更多参数，并修正了单词位置冲突的问题。

**2）基于最大熵的方法[8]（判别模型）**

对于一组特征，使得统计模型在这一组特征上的模型分布与样例中的经验分布完全一致，同时不对未知事件做任何假设，即保证这个模型尽可能的“均匀”（模型的熵值达到最大）。

**3）基于短语的SMT[9]**

在基于短语的模型中，直接将繁衍率信息、上下文信息以及局部对位调序信息记录在翻译规则中。

机器翻译=短语划分模型✖️短语翻译模型✖️短语调序模型✖️目标语言模型

**4）基于层次化短语的SMT[10]**

不破坏基于短语的翻译方法的优势，而是利用这些优势（短语有益于实现局部语义消歧和局部语序调整），并且对翻译规则进行泛化，解决数据稀疏、短语调序以及非连续短语翻译问题。

1. **树翻译模型**

树到串的翻译模型[11]：首先将源语言句子分析为一棵句法结构树，然后递归地将源语言句子的句法结构树转换为目标语言句子。

树到树的翻译模型[12]：首先将源语言句子分析为一棵句法结构树，然后递归地将源语言句子的句法结构树转换为目标语言句子的句法结构树，最后拼接叶节点得到译文。

串到树的翻译模型：利用串到树转换规则，将源语言句子分析为一棵目标语言句法结构树，然后拼接叶节点得到译文。

#### State-of-art （深度学习相关方法）

**神经机器翻译[13]（Neural machine translation, NMT）**

1. **神经机器翻译的诞生**

2013年，Nal Kalchbrenner和Phil Blunsom提出了一种用于机器翻译的新型端到端编码器-解码器结构[14]。该模型可以使用卷积神经网络（CNN）将给定的一段源文本编码成一个连续的向量，然后再使用循环神经网络（RNN）作为解码器将该状态向量转换成目标语言。

但是梯度爆炸/消失问题让RNN实际上难以处理长距依存问题。

1. **用于长距问题的记忆**

2014年，Sutskever et al.和Cho et al.开发了一种名叫seq2seq学习的方法，可以将RNN既用于编码器也用于解码器，并且还为NMT引入了长短时记忆（LSTM）[15, 16]。在门机制的帮助下（允许在LSTM中删除和更新明确的记忆），梯度爆炸/消失的问题得到了控制，从而让模型可以更好地获取句子中的长距依存。

LSTM的引入解决了长距离重新排序的问题，但是将NMT的主要难题变成了固定长度向量的问题：不管源句子的长度多长，这个神经网络都需要将其压缩成一个固定长度的向量，这会在解码过程中带来更大的复杂性和不确定性，尤其是当源句子很长时。

1. **注意力机制**

自2014年Yoshua Bengio的团队为NMT引入注意力机制后[17]，固定长度向量的问题也开始得到解决。根据源句子动态地生成一个加权语境向量，网络根据语境向量而不是某个固定长度的向量来预测词。

1. **产业界**

2016年，谷歌神经机器翻译（GNMT）使用NMT替代基于短语的机器翻译[18]。

2017年，Facebook人工智能研究院使用CNN实现NMT的方法，其可以实现与基于RNN的NMT近似的表现水平，但速度却快9倍[19, 20]。

2017年，谷歌发布了一个完全基于注意力机制的NMT模型[21]。

2017年，亚马逊发布他们使用MXNet的NMT实现[22]。

2018年，谷歌发布Universal Transformer[23]，弥补了在大规模语言理解任务上具有竞争力的实际序列模型与计算通用模型之间的差距。

### Benchmark

#### 常用数据集[24]

1. LDC（https://www.ldc.upenn.edu/）

LDC（Linguistic Data Consortium，语言数据联盟）提供了大量的平行数据，大多数都是“阿-英”、“汉-英”和“法-英”的语料资源。

1. Europal（http://www.statmt.org/europarl/）

Europal语料是欧洲议会会议记录文本集，这些会议记录被译成11种语言，这些语料是由每种语言大约4000万的单词组成的。语料库中没有中文语料。

1. OPUS（http://opus.nlpl.eu/）

OPUS收集了各种平行语料，包括公开资源软件本地化的语料和文档。包含中文的一些比较大的数据集主要有MultiUN（http://opus.nlpl.eu/MultiUN.php）和OpenSubtitles2016（http://opus.nlpl.eu/OpenSubtitles2016.php），均有200万左右的句对。

1. Acquis Communautaire

（https://wt-public.emm4u.eu/Acquis/JRC-Acquis.2.2/doc/README\_Acquis-Communautaire-corpus\_JRC.html）

Acquis Communautaire语料是由欧盟成员国签订的法律文献的集合，有超过20多种语言，语料库中没有中文语料。

1. UM-Corpus（http://nlp2ct.cis.umac.mo/um-corpus/index.html）

UM-Corpus是由自然语言处理与中葡机器翻译实验室整理的语料，大约200万中英对齐语料，涵盖教育、法律、微博、新闻、科学、演讲、字幕和论文等8个主题。

1. WMT数据集

#### 算法在数据集上的性能

|  |  |  |
| --- | --- | --- |
| Model | BLEU | |
| WMT14 En-De | WMT14 En-Fr |
| Universal Transformer(base)[23] | **28.9** |  |
| Transformer(big)[21] | 28.4 | **41.8** |
| Transformer(base)[21] | 27.3 | 38.1 |
| Deep-Att + PosUnk Ensemble[21] |  | 40.4 |
| MoE[21] | 26.03 | 40.56 |
| Deep-Att + PosUnk[21] |  | 39.2 |
| ByteNet[21] | 23.75 |  |
| ConvS2S(BPE 40K)[20] | 25.16 | 40.51 |
| GNMT(Word pieces) [20] | 24.61 | 38.95 |
| GNMT(Word 80K) [20] | 23.12 | 37.90 |
| GNMT(Word pieces)+RL[20, 21] | 24.6 | 39.92 |
| ConvS2S Ensemble[20] | 26.43 | 41.44 |
| GNMT Ensemble[20] | 26.20 | 40.35 |
| GNMT + RL Ensemble[20] | 26.30 | 41.16 |
| PBMT[25] | 20.67 | 37.03 |
| RNNSearch[25] | 16.46 | 29.97 |
| RNNSearch-LV[25] | 16.95 | 32.68 |
| LSTM[16] |  | 30.59 |
| LSTM Ensemble[16] |  | 34.81 |

### 总结

在高速发展和高度竞争的环境中，神经机器翻译技术正在取得显著的进展，同时NMT还将在许多方面得到不断完善，其中包括：罕见词问题、单语言数据使用、多语言翻译/多语言NMT、记忆机制、语言融合、覆盖问题、训练过程、先验知识融合、多模态翻译等。因此，我们有足够的理由相信NMT还将取得更大的突破，还将替代SMT逐渐发展成主流的机器翻译技术，并在不久的将来让全社会受益。

### 引用

[1] 王砚峰. https://www.zhihu.com/question/59955860/answer/175412172

[2] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.

[3] 从冷战到深度学习，一文看懂机器翻译发展史. https://blog.csdn.net/csdnnews/article/details/79662331

[4] Koehn P. Statistical machine translation[M]. Cambridge University Press, 2009.

[5] Brown P F, Cocke J, Pietra S A D, et al. A statistical approach to machine translation[J]. Computational linguistics, 1990, 16(2): 79-85.

[6] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.

[7] Knight K. A statistical MT tutorial workbook[C]//Prepared for the 1999 JHU Summer Workshop. 1999.

[8] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 295-302.

[9] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.

[10] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 263-270.

[11] Liu Y, Liu Q, Lin S. Tree-to-string alignment template for statistical machine translation[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 609-616.

[12] Zhang M, Jiang H, Aw A T, et al. A tree-to-tree alignment-based model for statistical machine translation[J]. MT-Summit-07, 2007: 535-542.

[13] 机器之心. 神奇的神经机器翻译：从发展脉络到未来前景. https://www.jiqizhixin.com/articles/2017-08-22-6

[14] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.

[15] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[16] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[17] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[18] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[19] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation[J]. arXiv preprint arXiv:1611.02344, 2016.

[20] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[22] Train Neural Machine Translation Models with Sockeye | Amazon Web Services(2017). Amazon Web Services. Retrieved 26 July 2017, from https://aws.amazon.com/blogs/ai/train-neural-machine-translation-models-with-sockeye/

[23] Dehghani M, Gouws S, Vinyals O, et al. Universal Transformers[J]. arXiv preprint arXiv:1807.03819, 2018.

[24] 老顽童2007. 神经机器翻译（NMT）相关资料整理. http://www.cnblogs.com/zhbzz2007/p/6276712.html

[25] Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation[J]. arXiv preprint arXiv:1412.2007, 2014.