MINDS Workshops
Machine Translation Working Group
Final Report

报告人：刘群

2011-8-20 于 CNCCL2011

说明

- 本报告是对“ MINDS Workshops Machine Translation Working Group Final Report”一文的讲解（原文作者是 Alon Lavie ， David Yarowsky ， Kevin Knight ， Chris Callison-Burch ， Nizar Habash ， Teruko Mitamura ）
- 原文内容较旧，有些地方不太符合目前的进展
- 讲解中主要参考了英文报告原文以及董振东老师提供给我的由蔡东风老师组织人进行翻译的译文，特此感谢
- 本报告依据原文， 并参照蔡东风老师提供的译文，根据报告人自己的理解进行了阐释和补充
- 本报告翻译或理解错误之处由本人负责，并请方家指正。

目录

历史回顾

- 范式的转换
  - 1950s-1980s ：基于规则的知识丰富的系统
  - 1980s 末开始：以计算搜索为核心的系统
    - IBM 模型
    - 基于短语的模型
- 新范式渗透到了机器翻译的各个方面
    - 基于实例的方法
    - 基于转换的方法
    - 基于上下文的机器翻译方法
    - 多引擎机器翻译方法

历史回顾

- 产生新范式的原因
  - 大规模语料库资源
  - 快速计算资源的普及（低成本化）
  - 机器学习技术的进步
  - 机器翻译质量自动评价技术
    - 解决了机器翻译评价对人类专家的依赖
    - 为机器翻译的参数优化提供了目标函数

目录

最新机器翻译方法的局限

- 距离“面向多语言的全自动、广覆盖度、高品质的机器翻译”的最终目标仍有距离
  - 只有少量的语言对之间有大规模双语语料库，大量的语言对仍然无法达到较好的翻译质量
  - 大规模双语语料库只集中在少数的领域和体裁，脱离这些领域和体裁翻译质量迅速下降
  - 刘群补充：对于结构相差较大的语言对，即使语料很大，领域受限，机器翻译效果依然不能令人满意

最新机器翻译方法的弱点

模型能力弱

搜索中的区分能力弱

模型能力弱

根据模型无法产生正确的译文，在搜索空间中不存在正确的结果

- 现状：目前的模型主要通过语料来学习互译的短语
- 优点：模型简单，无需标注
- 缺点：不能刻划各个语言层次之间的关联
- 解决办法：对句法和语义之间的映射直接建模
- 问题：如何获得这样的模型？
- 探索：基于句法的模型、因素化模型

搜索中区分能力弱

即使搜索空间中有正确的译文，但搜索算法的评分函数无法有效区分正确的译文和错误的译文，导致无法搜索到正确的译文

- 语言模型是解码器搜索算法的知识主要来源
- 目前主要使用 trigram 模型
- 研究界在尝试使用更长的 N-gram 模型，有效果但不能根本解决问题
- 迫切需要针对机器翻译的新的语言模型建模方法

目录

对于新的重大研究课题的建议

机器翻译的巨大挑战

具体的研究课题建议

机器翻译的巨大挑战

众多语言机器翻译所需要的技术进展

领域和体裁的鲁棒性所需要的技术进展

人类水平的机器翻译所需要的技术进展

这几点与前述局限性对应

众多语言机器翻译所需要的技术进展

可用有限数据资源进行训练的机器翻译模型

适用于复杂形态语言的机器翻译模型

适用于众多语言的有效的语言模型

可用有限数据资源进行训练的机器翻译模型

- 在可以预见的未来，多数语言对之间的语料库的规模仍是有限的
- 如今使用有限数量的数据训练的翻译模型所表现出来的翻译效果仍然没有达到令人满意的水平
- 急需研究一个能够更好地利用有限的训练数据进行泛化的新模型，以在更多的语言对之间搭建机器翻译系统

适用于复杂形态语言的机器翻译模型

- 目前的研究主要集中在少数语种（阿、汉→英）
- 目前的研究都不考虑词形变化
- 相当多语言具有复杂的形态变化
- 复杂形态变化会严重加剧数据稀疏问题
- 现有一些要素化模型和句法方法尝试解决这一问题，但远远不够令人满意
- 开发一个可以有效地解决复杂词形变化的新模型，无论以该语言作为源语言还是目标语言进行翻译

适用于众多语言的有效的语言模型

- 目前最有效的模型还是 N-gram 模型
- N-gram 模型对很多其他语言并不有效：
  - 不能处理复杂形态
  - 不能处理长距离形态标记依赖关系，如：
    - 动词和论元之间的一致性标记（如主谓语人称、数的一致性）
    - 动词和相应论元的格标记

机器翻译的巨大挑战

众多语言机器翻译所需要的技术进展

领域和体裁的鲁棒性所需要的技术进展

人类水平的机器翻译所需要的技术进展

领域和体裁的鲁棒性所需要的技术进展

可以很好泛化的机器翻译模型

到新体裁和领域的模型自适应

从平行数据中有目标的主动的学习

可以很好泛化的机器翻译模型

- 短语模型是完全词汇化的，泛化能力不足
- 导致机器翻译系统对于词汇和文本风格的变化显得极其脆弱
- 新的机器翻译模型应该能够从一个领域（或体裁）的语料库中学习到与领域（或体裁）无关的语言映射规律，使得从一个领域（或体裁）中学习到的知识可以很好地用于其它领域（或体裁）文本的反应

到新体裁和领域的模型自适应

- 不同领域（或体裁），对应的翻译模型的参数和概率空间有明显不同
- 当系统被应用于与训练数据不同的领域（或体裁）是，系统应该能自动调整模型的参数，以便适应新的领域（或体裁）
- 自适应过程中可以使用少量的新领域（或体裁）的数据

从平行数据中有目标的主动的学习

- 如何有效地利用有限的新数据来调整和拓展机器翻译系统，使之适应新的领域、体裁，将是一个主要的挑战
- 目前还很少有人研究如何识别不同的领域和文本体裁之间的差异，并利用这种差异信息来有针对性地学习一个新模型
- 此外，在某些情况下，可以主动创建少量的有针对性的新的训练数据，用于改进机器翻译系统的性能。例如机器学习技术中的“主动学习”就是探索和解决这类问题的一般框架

机器翻译的巨大挑战

众多语言机器翻译所需要的技术进展

领域和体裁的鲁棒性所需要的技术进展

人类水平的机器翻译所需要的技术进展

人类水平的机器翻译所需要的技术进展

基于高层的句法和语义表达的机器翻译模型

融入句间上下文的机器翻译模型

以语义为中介的基于统计中间语言的机器翻译方法

基于高层的句法和语义表达的机器翻译模型

- 研究人员都认识到，要想让机器翻译的水平接近人类的翻译水平，需要能够获取较高的语法和语义层次的表示及其在不同语言间映射的翻译模型
- 关键的技术挑战是要定义既丰富、强大，又简单到足以支持从训练数据中自动获取模型的表示形式
- 我们认为，最可行的方法是利用标注好句法和语义结构信息的句对齐平行语料库来训练基于句法和语义的翻译模型
- 一个更具挑战性的情况是使用某种平行语料库训练的模型，该平行语料库只有其中一种语言的句法和语义结构
- 开发这种带标注的语料库是一个关键步骤，如果没有这个步骤任何研究都无法开始

融入句间上下文的机器翻译模型

- 目前，我们所有的工作实际上都集中在翻译独立的句子上
- 但是，在所有语言中，人类书写的文本都有明显的篇章结构，用于表达各种思想，并且不同语言的篇章结构也不尽相同
- 进而，需要句子间上下文信息来分析指代关系，才能正确翻译代词指代对象和其他实体
- 因此，我们认为真正的人类水平的机器翻译将需要能够显式地使用句子间的信息，来解决特定类型的歧义的方法，以及能生成目标语言中连贯的多语句篇章结构的方法

以语义为中介的基于统计中间语言的机器翻译方法

- 基于深层语义分析和中间语言的机器翻译方法曾被广泛关注，但是在十五年前已经不再流行，因为该方法的弱点现在还无法克服。
- 尽管如此，我们认为在自动分析和生成中间语言的模型上的研究和努力是有价值的，该研究将最终导致真正的人类翻译水平的机器翻译系统的产生。

对于新的重大研究课题的建议

机器翻译的巨大挑战

具体的研究课题建议

具体的研究课题建议

有效的通用亚句子级机器翻译模型

从更少的数据学习更多的知识

从英语到其他语言的机器翻译

多引擎机器翻译

机器翻译评价

有效的通用亚句子级机器翻译模型

- 我们建议创建这样的研究方案，该方案明确地鼓励机器翻译的研究人员把工作重心放在开发一个能获得不同层次句法和语义关联的子句机器翻译模型
- 一个首要的关键步骤就是构建一个单语或双语都标注了准确句法和语义结构的句对齐的平行语料库
- 另一个思路是研究普遍适合于各种领域和体裁的机器翻译模型，这将有助于产生更通用的机器翻译系统
- 还有一个思路是研究一些特定的机器翻译子问题，如明确表现出语言差异性的句子的翻译。

从更少的数据学习更多的知识

- 目前的研究可以利用大规模数据进行有效训练
- 然而，我们认为，还需要从标注有较高表示层次的有限数据中学习尽可能多的知识，这就激励研究者们开发更加适合实际，并且当前只有有限可用数据的机器翻译模型。
- 这个挑战就是，通过给定有限的训练数据以及其他自然语言处理工具和资源，如何开发具有最佳性能的机器翻译系统。

从英语到其他语言的机器翻译

- 现有研究项目中机器翻译的研究主要集中在将少量语言翻译成英语
- 但是将英语翻译成其他语言也极为重要，因为它们形成了新的挑战，迫使机器翻译研究人员去解决之前没有引起足够重视的语言现象
- 上文所述的对这种复杂形态的恰当处理，是翻译成其他语言而不是英语的面临一个明显的挑战
- 这样的研究可以充分利用源语言（英语）端成熟的自然语言处理工具
- 我们认为其他语言 NLP 工具的缺乏也是导致目前利用源语言端 NLP 技术的机器翻译研究落后的原因

多引擎机器翻译

- 我们认为研究多引擎机器翻译是非常重要的，因为可以预见的未来，不太可能出现某种单一的机器翻译方法，能够在各种情况下全面超过其他方法。
- 此外，在评估科研小组的研究方案时，不仅考虑应其机器翻译系统独立使用时的性能，而且还需考虑该系统在多引擎机器翻译系统中的贡献，这便鼓励了不同机器翻译方法的研究。

机器翻译评价

- 对于当前和未来的机器翻译的研究而言，这些评价指标都是非常重要的，因为自动评价标准可以使研究小组根据常见的具体性能评价结果来指导其翻译系统的开发，而且自动机器翻译评价指标（例如BLEU）提供了一个可用于优化模型参数的“目标函数”，以达到最佳的翻译性能
- 然而，目前的自动机器翻译质量评价标准仍然非常粗糙，且与人类对翻译质量的评价的关联度较低。
- 随着机器翻译系统不断改进并实现较高翻译质量，自动评价指标变得非常重要，该评价指标应对句子层次的微小差异敏感，以至于能够检测出微小的改进，对具体的翻译错误进行分离和识别，并且优化系统参数以真正达到最好的翻译性能。

感想

- 看法基本一致
- 有些已经实现
  - 句法模型
  - 系统融合
- 有些还没有进展
  - 语义模型
  - 语言模型
  - 复杂形态处理
  - 主动学习
- 本人认为 有前景的方向：
  - 深层机器学习算法与深层语言知识的结合
  - 与社交网络的结合
-

谢谢