MI&T Lab
BRIDGE MULTILINGUAL COMMUNICATIONS

机器翻译自动评价十年

杨沐昀、赵铁军、朱俊国

哈尔滨工业大学计算机科学与技术学院

机器智能与翻译研究室

2012年9月，西安

前言

- 自动评价已成为机器翻译中的关键因素
  - 自动评价为导向的MT建模
    - 研究中采用多种自动评价指标验证成果
    - 评测中使用多种自动评价指标考量性能
    - 每年不断出现新的角度构建自动评价策略
  - 某些情况下自动评价仍无法正确区分翻译质量
    - 系统融合案例

提 纲

- 机器翻译自动评价研究概况
  - 字符串相似度方法、机器学习方法
  - 机器翻译自动评价的公开评测
  - 分析型评价方法的出现
- 机器翻译自动评价中的挑战
  - 语言学特征能否解决评价问题
  - 用户选择什么样的译文
- 小结

机器翻译自动评价研究概况

- 基于字符串相似度方法
- 基于机器学习的多特征融合方法
- 机器翻译自动评价的评测实践
- 分析型评价的出现

基于字符串相似度的方法

- “机器译文”的自动评价出发点：
  - 有多好　　(×)
  - 哪个更好　　(√)
- 评价性能度量：人工评价结果为标准
  - 准确率　　(×)
  - 结果一致程度　　(√)
    - 相关系数：Pearson, Spearman, Kendall’s Tau
    - 不关心人工评价和自动评价具体分数

基于字符串相似度的方法

- BLEU： (Papineni et al, ACL 2002/IBM TR 2001)
  - 基于n-gram精确率的相似度计算、简单、高效
  - 系统级评价与人高度一致
  - 句子级评价性能较差
    - 不区分词的差别
    - 不区分n-gram的差别
    - 未考虑召回率
    - 几何平均值
    - 参考译文不完备
    - ……

$$Pn = \frac{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count(n-gram)}$$

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

$$N=4，\ W_n=1/4$$

基于字符串相似度的方法

Metrics	Type of Gram	位置	Stem	Word Net	精确率	召回率	F值	模型
NIST	N-gram	gram内有序 gram间无序	—	—	✓	—	—	相似度
Rouge	Skip-bigram	gram内有序 gram间无序	—	—	—	✓	—	相似度
GTM	——	——	—	—	—	—	✓	相似度
WER	——	有序/编辑距离	—	—	—	—	—	错误率
PER	——	无序	—	—	—	—	—	错误率
TER	——	有序/编辑距离 /允许块移动	—	—	—	—	—	错误率
METEOR	unigram	对齐最小交叉	✓	✓	—	—	✓	相似度

基于机器学习的多特征融合方法

- 核心问题：机器译文与参考译文间的相似度计算
- 基于机器学习的多(语言)特征融合
  - QARLA, 22 features, Amigò et al. (ACL, 2005)
  - SVM-regression, 53 features, Albrecht and Hwa (ACL, 2007)
  - “Linear Combination”, 30 features, Giménez and Màrquez (IJCNLP, 2008)
  - SVM-ranking, 20 features, Duh (WMT, 2008)
  - Linear regression, 155 features Padó et al. (ACL, 2009)
- 各种机器学习方法都进行了尝试
  - 分类、回归、排序

基于机器学习的多特征融合方法

- 原有方法以及其中使用的信息均作为特征
  - BLEU1, BLEU2, BLEU-2…BLEU4, BLEU-4
  - NIST…..
  - Rouge……..
  - METEOR,GTM
- 各种语言学特征纷纷引入
  - 从句法到语义，一一出现并且实验证明有效
    - Text Entailment (Padó et al, ACL 2010)
    - Semantic Role Labeling (Lo et al, ACL2011/SSST 2012)）

机器翻译自动评价的评测实践

- 美国: NIST MetricsMATR
- 美国国家标准与技术局(NIST) 举办
  - 2008
  - 2010

	阿拉伯语-英语 (约8800词)
测试数据	汉语、阿拉伯语、波斯语-英语（约83,000词） 英语、阿拉伯语-法语（约92,000词） 法语、英语-阿拉伯语（10年新增）
人工评价标准	精确度（接受/不接受,1-4,1-5,1-7)、流利度(1-5) 、偏向性、适当度、基于人工编辑的翻译错误率
级别	句子/篇章/系统
参考译文数量	1 或 4
评价指标	Pearson, Spearman, Kendall’s Tau
评测条件	08年47(共计141项) 10年51个(共计153项，新增阿拉伯语)
最佳系统	TERp
参与单位	14家单位：CMU, Stanford, USC-ISI, IBM, Columbia Univ., UW, Maryland, RWTH等

机器翻译自动评价的评测实践

- 欧洲: WMT Shared Evaluation Task
  - 组织者：ACL/EACL Workshops on SMT
  - 举办时间：2007~2011

	WMT07	WMT08	WMT09	WMT10	WMT11
训练集	无	WMT07	WMT08	WMT08 WMT09	WMT08，WMT09， WMT10
测试集(句子数)	17,820	25,051	35,786	37,996	47,610
涉及语言	英语-法语、德语、西班牙语、捷克语 法语、德语、西班牙语、捷克语-英语				
参考译文数量	1个				
评价指标	Spearman	Searman，Kendall’s Tau			
人工评价标准	精确度， 流利度， 句对排序	句对排序			
级别	系统级	句子级/系统级			
评测条件	2个	4个			
参赛情况	11个系统	15个系统 （7家单位）	19个系统 （9家单位）	26个系统 （14家单位）	21个系统 （9家单位）
最佳系统	Semantic / BLEU	meteor-ranking （2个1st）	Terp （2个1st）	SVM_RANK (3个1st)	MTERATER-PLUS （2个1st）

机器翻译自动评价技术的评测

- 评测中的经验总结
  - 模型方面
    - 模型框架：排序学习
    - 特征工程：语言独立
    - 训练集选择： 尽量选择同一个系列的数据进行训练
  - 人工评价方面
    - “MT Results are equally good.”---NIST report
    - NIST倾向于提供人工评分（5分/7分）
      - 7分制人工评价结果一致率40% (NIST08).
    - WMT主要采用人工排序(Preference Judgment)

分析型评价的出现

- 基于语言学检测点的方法
  - (Yu， MT 1993)
- 基于词性的错误分析方法
  - (Popovic& Ney, WMT 2007)
  - 自动分析三种典型错误
  - 过程上配合使用WER和PER方法
- Wood Pecker 方法
  - (Zhou et al, Coling 2008)

机器翻译自动评价中的挑战

- 语言学特征能否解决评价问题
  - 持续引入语言学特征的性能上限？
- 用户选择什么样的译文
  - 自动评价能否体现用户的偏好
  - 用户的选择是与翻译质量一致

语言学特征能否解决评价问题

- 目前研究难点
  - 句子级高质量的评价方法
- 主要矛盾：语言独立性与高性能
  - 前者带来普遍的适用性、方便
  - 后者能提高性能，使用范围受限(语种、数据)
- 从解决问题角度：引入语言学特征是否持续有效？

语言学特征能否解决评价问题

实验 1

- 定义翻译角度55种语言学错误
  - 词汇层、短语层、句法层、篇章层
  - 2人手工标注、
- 英汉人工翻译152个样本
  - 12分制
- 汉英机器翻译随机324个样本
  - 5分制 [LDC2006T04]
- SVM回归模型进行特征融合

模型	英汉	汉英
BLEU4	0.64	0.35
NIST5	**0.77**	0.34
METEOR	0.69	**0.41**
ROUGE-S*	0.71	0.37
SVM(仅语言学特征)	0.78	0.38
SVM(全部特征)	0.84	0.44

语言学特征能否解决评价问题

实验 1

- 人工方式穷尽翻译错误能部分提升评价性能
- 字符串相似度与语言学特征形成互补
- 存在显著负相关的翻译错误
  - 译文缺失、句子乱译、介词短语错译、非谓语动词错译
  - 错的越多的分越高
  - 如何在自动评价模型中利用？

用户选择质什么样的译文

实验2

- 考察翻译用户的选择与自动评价的关系
  - 本领域文章标题：229条，获取5个网上汉英翻译结果
  - 任务：自主选择一个译文修改并完成人工翻译
  - 记录：用户编辑行为、最终人工翻译结果
  - 人员：31人/英语翻译硕士
  - 结果：4773个人工修改译文，平均154条/人
- 考察层次：系统级 | 句子级

用户选择质什么样的译文

实验2

- 系统级结果(BGBYS)

	系统1	系统2	系统3	系统4	系统5
BLEU	0.23	0.14	0.13	0.08	0.06
TER	0.73	0.81	0.81	0.94	0.95
NIST	5.20	4.28	4.09	3.47	2.93
METEOR	0.42	0.35	0.34	0.30	0.30
译文采用次数	1409	1112	1338	782	808
平均编辑时间	38.8	44.7	32.7	40.3	67.2
增加/删除操作	3.40	4.13	4.12	4.37	5.20

用户选择质什么样的译文

实验2： 自动评价的优化目标？

- 句子级一致性(Spearman相关)
  - 自动评价、人工评价、使用率、编辑量

	BLEU	TER	NIST	METROR	人工评价	使用率	编辑量
BLEU	1	.707	.759	.813	.087	.311	.316
TER		1	.594	.651	.112	.271	.257
NIST			1	.821	.129	.357	.339
METEOR				1	.100	.350	.385
人工评价					1	.414	.213
使用率						1	.380
编辑量							1

小结

- 翻译自动评价也许不存在唯一解
  - 继承翻译评价的中的所有难题
- 从翻译模型训练角度
  - 需要具有良好数学特性、满足不同学习算法
- 从机器翻译研究人员角度
  - 需要某种“白箱”评价技术
- 从翻译用户角度
  - 需要经济、可靠的、体现用户满意度的指标

谢谢！

For any problems, please contact
ymy {AT} mtlab.hit.edu.cn