CWMT 机器翻译评测回顾与展望

吕雅娟

中国科学院计算技术研究所

自然语言处理研究组

主要内容

- CWMT机器翻译评测回顾
- 2013年机器翻译评测设想与规划
- 2013年机器翻译评测讨论

全国机器翻译研讨会

- 统计机器翻译研讨会（SSMT，2005 — 2007）
- 全国机器翻译研讨会（CWMT，2008 — 2012）

	时间	承办单位和地点	评测
第一届	2005	厦门大学（厦门）	
第二届	2006	中科院计算所（北京）	
第三届	2007	哈尔滨工业大学（哈尔滨）	✓
第四届	2008	中科院自动化所（北京）	✓
第五届	2009	南京大学（南京）	✓
第六届	2010	中科院软件所（北京）	
第七届	2011	厦门大学（厦门）	✓
第八届	2012	西安理工大学（西安）	

历届机器翻译研讨会回顾

CWMT机器翻译评测概况

- 2007年首次评测，已举办4届
  - 2007， 2008， 2009， 2011
- 评测的主办单位
  - 中国中文信息学会
- 评测的组织单位
  - 中国科学院计算技术研究所
- 目标
  - 推动研究单位间的实质性交流
  - 推进机器翻译技术的发展

CWMT机器翻译评测概况

- 评测组织方式
  - 评测组织方提供训练、开发和测试数据
  - 参评单位在给定时间内返回翻译结果并提交技术报告
- 项目设置
  - 参考当时的研究热点和需求
- 评价方法
  - 自动评价
  - 人工评价（2009年汉蒙机器翻译评测）
- 历届评测资料：
  - http://10.28.0.169/mteval/CWMT/index.php

历届机器翻译评测回顾-2007

- 项目设置

序号	评测项目名称	语种	领域	备注
1	汉英机器翻译	汉语→英语	新闻领域	区分受限/不受限
2	英汉机器翻译	英语→汉语	新闻领域	区分受限/不受限
3	词语对齐	汉语→英语	新闻领域	语料受限

- 评价方法：自动评价
  - 机器翻译
    - BLEU4, NIST5, GTM，mWER, mPER, ICT
  - 词语对齐
    - 对齐错误率（AER），准确率，召回率，F1值

历届机器翻译评测回顾-2008

- 项目设置

序号	评测项目名称	语种	领域
1	汉英新闻领域单一系统	汉语→英语	新闻领域
2	汉英新闻领域系统融合	汉语→英语	新闻领域
3	英汉新闻领域机器翻译	英语→汉语	新闻领域
4	英汉科技领域机器翻译	英语→汉语	科技领域

- 评价方法：自动评价
  - BLEU、**BLEU-SBP**、NIST、GTM、mWER、mPER、ICT
  - **WoodPecker (检测点评价，微软亚洲研究院提供)**
  - 符号检验（显著性检验）

历届机器翻译评测回顾-2009

- 项目设置

序号	评测项目名称	语种	领域
1	汉英新闻领域单一系统	汉语→英语	新闻领域
2	汉英新闻领域系统融合	汉语→英语	新闻领域
3	英汉新闻领域机器翻译	英语→汉语	新闻领域
4	英汉科技领域机器翻译	英语→汉语	科技领域
5	汉蒙日常用语机器翻译	汉语→蒙语	日常用语

- 评价方法
  - 自动评测： **BLEU-SBP（主要指标）**，BLEU、NIST 、GTM 、mWER 、mPER、ICT、WoodPecker
  - **WoodPecker：** 加入了大量的人工干预，以减少词语对齐和句法分析带来的噪音
  - 人工评测：仅用于汉蒙评测方向
  - 开始考虑 **Progress评测**： 汉英和英汉新闻评测数据不公布参考答案，连续使用

历届机器翻译评测回顾-2011

- 项目设置

序号	评测项目名称	语种	领域
1	汉英新闻领域机器翻译	汉语→英语	新闻领域
2	英汉新闻领域机器翻译	英语→汉语	新闻领域
3	英汉科技领域机器翻译	英语→汉语	科技领域
4	日汉新闻领域机器翻译	日语→汉语	新闻领域
5	蒙汉日常用语机器翻译	蒙古语→汉语	日常用语
6	藏汉政府文献机器翻译	藏语→汉语	政府文献
7	维汉新闻领域机器翻译	维吾尔语→汉语	新闻领域
8	哈汉新闻领域机器翻译	哈萨克语→汉语	新闻领域
9	柯汉新闻领域机器翻译	柯尔克孜语→汉语	新闻领域

- 特点
  - 包括了**5种民族语言**到汉语的翻译评测
  - 重点关注**各种语言到汉语**的翻译
  - 汉英双向新闻**进展评测，提供了在线打分网站**

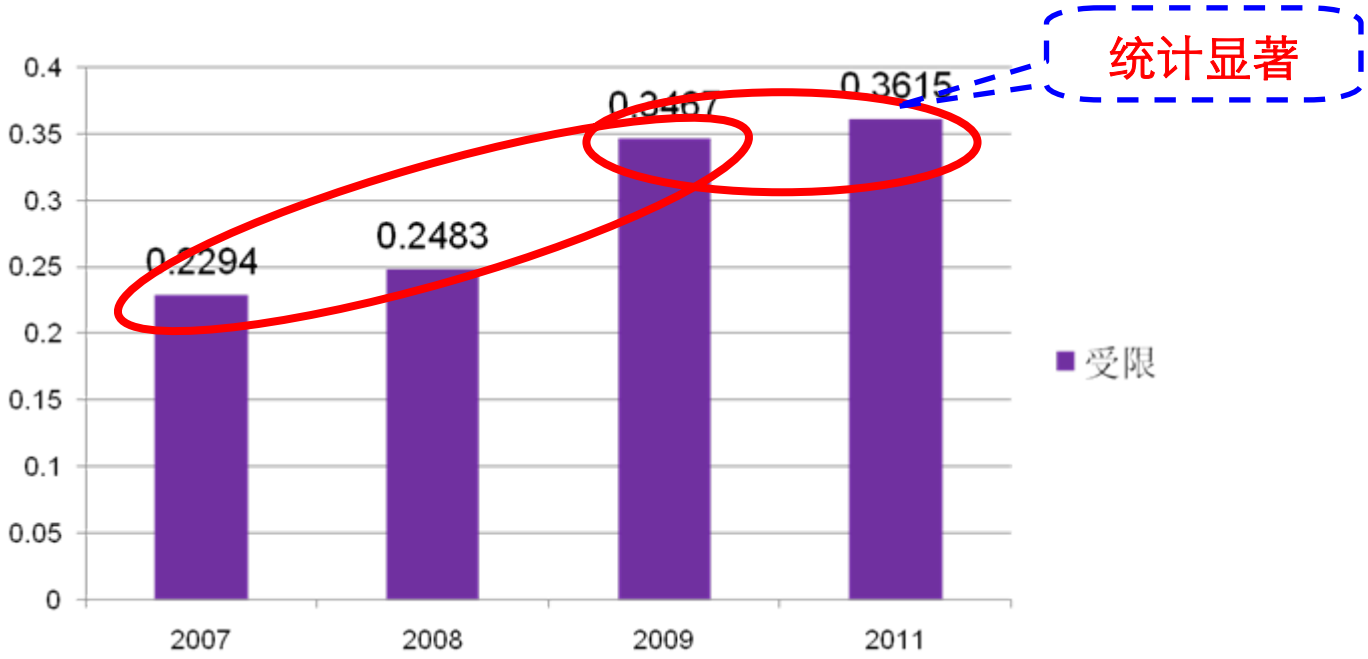
CWMT机器翻译评测小结

- 项目设置和参评情况：逐年增加

年度	项目数	语种数	领域数	参评单位数(国外)	参评系统数
2007	3	2	1	11 (2)	35
2008	4	2	2	15 (2)	73
2009	5	3	3	18 (4)	102
2011	9	7	4	19 (4)	165

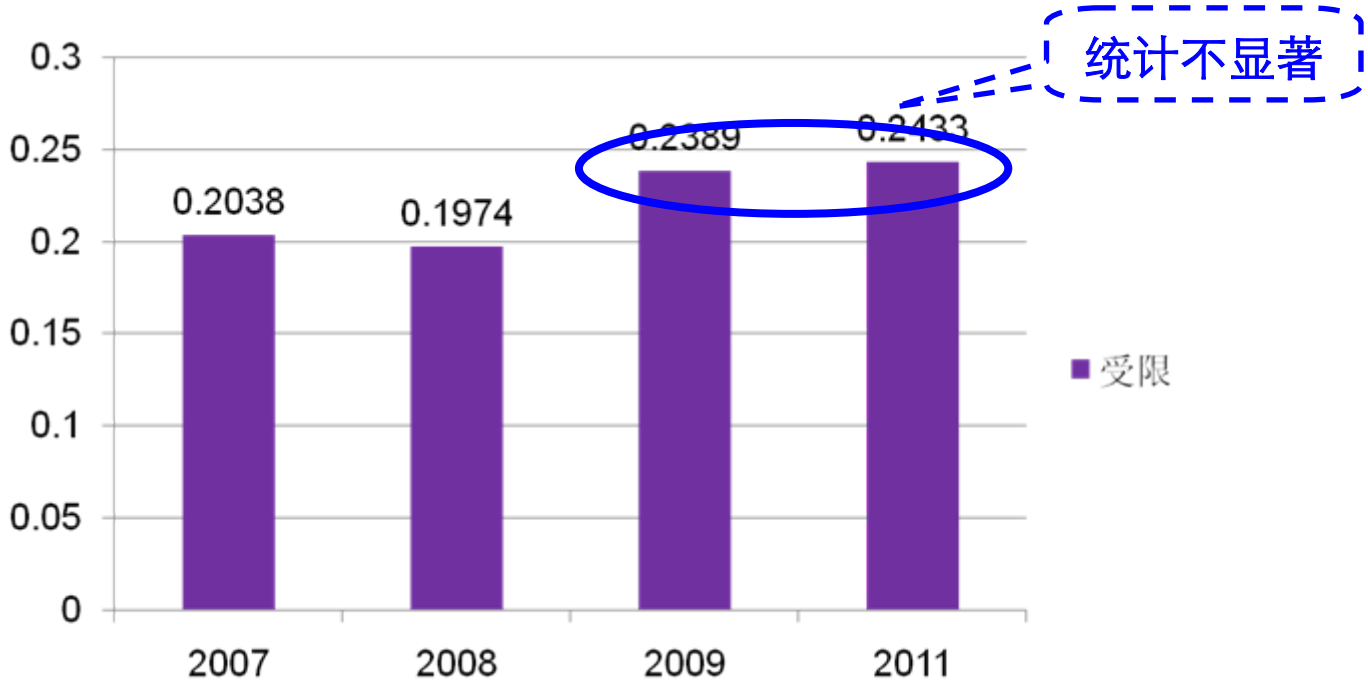
CWMT机器翻译评测小结

- 翻译质量：提高显著 → 提高不显著
  - 英汉新闻评测历届BLEU5前5名平均值对比



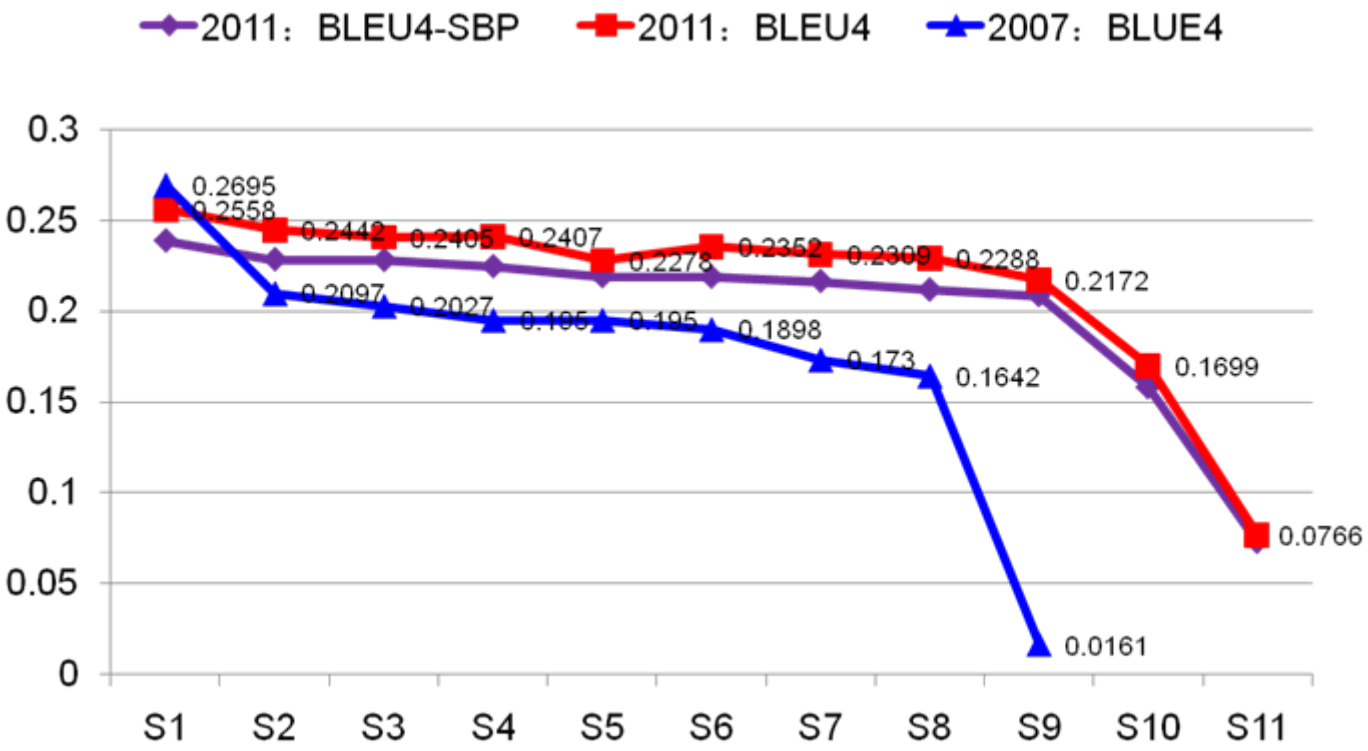
CWMT机器翻译评测小结

- 翻译质量：提高显著 → 提高不显著
  - 汉英新闻评测历届BLEU5前5名平均值对比



CWMT机器翻译评测小结

- 各参评单位差距逐渐缩小



2011年汉英新闻（Progress） 符号检验结果

INSTITUTE OF COMPUTING TECHNOLOGY

	S18	S6	S5	S1	S8	S2	S7◆	S15	S9	S14	S12
S18	—	●	●	●	●	●	●	●	●	●	●
S6	●	—	○	●	●	●	●	●	●	●	●
S5	●	○	—	○	●	●	●	●	●	●	●
S1	●	●	○	—	○	○	○	●	●	●	●
S8	●	●	●	○	—	○	○	○	●	●	●
S2	●	●	●	○	○	—	○	●	●	●	●
S7◆	●	●	●	○	○	○	—	○	●	●	●
S15	●	●	●	●	○	●	○	—	○	●	●
S9	●	●	●	●	●	●	●	○	—	●	●
S14	●	●	●	●	●	●	●	●	●	—	●
S12	●	●	●	●	●	●	●	●	●	●	—

- Sign test (Collins et al., 2005) for significance test with BLEU5-SBP for primary systems
- Significant difference marked with ● and no significant difference marked with ○, $p<0.05$

CWMT机器翻译评测小结

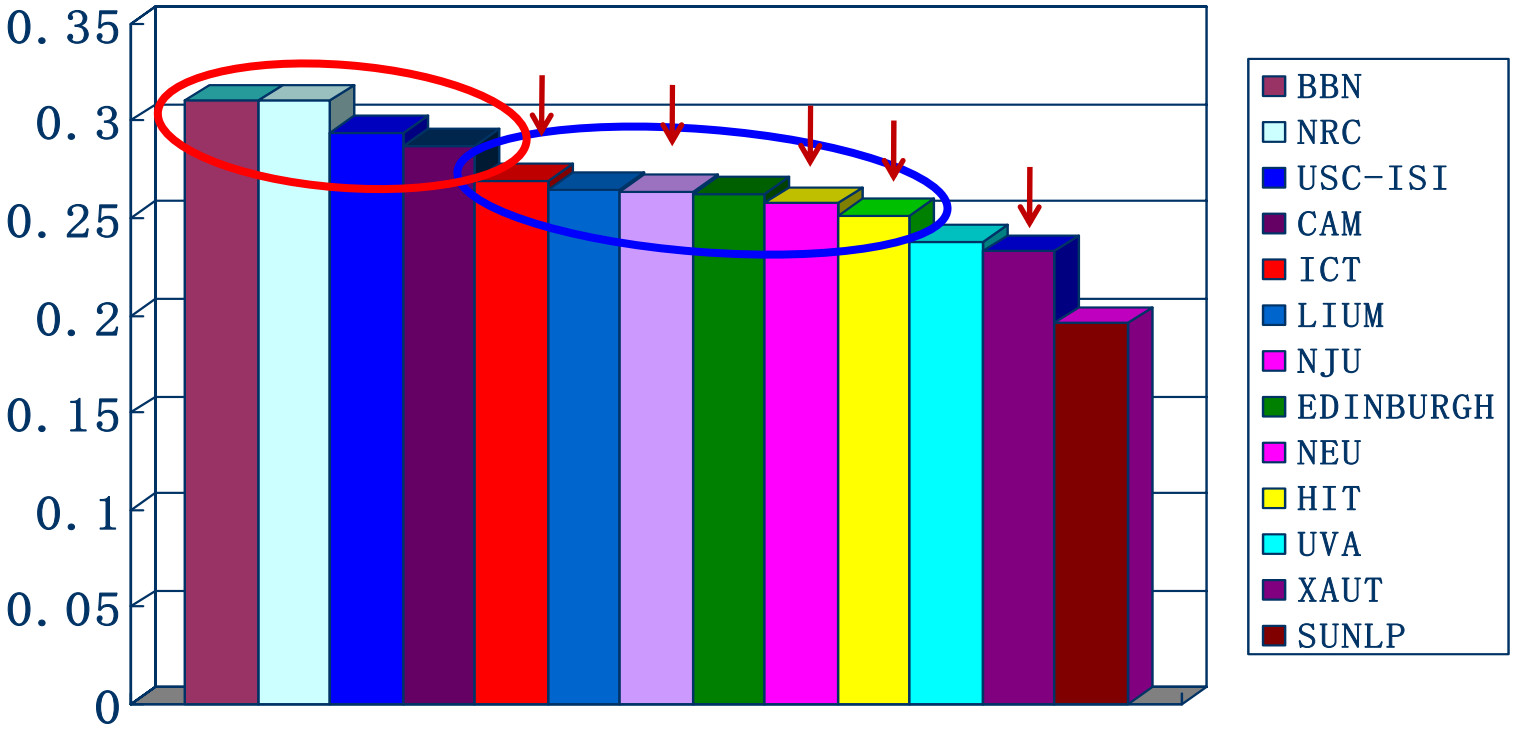
- 为一些单位开展机器翻译研究提供了平台
  - 增加了很多新的参评单位
    - 北京交通大学
    - 西安理工大学
    - 中科院新疆理化所
    - 北京航空航天大学
    - 中科院合肥智能机械研究所
    - 内蒙古师范大学
    - ……

CWMT机器翻译研讨会和评测小结

- 全国机器翻译研讨会和评测为国内统计机器翻译研究的普及和发展发挥了重要作用
  - 早期参加的单位已经在国际评测中崭露头角
    - 计算所在NIST2006、2009成绩优异、IWSLT2010评测6项第一
    - 自动化所IWSLT2009评测取得多项第一名
  - 后来加入的单位发展迅速， 各参评差距逐渐缩小
    - 东北大学、南京大学
  - 统计翻译技术已经得到普及和关注

存在的问题

- 传统项目近两次评测进展不显著
  - 进入平台期?与国外顶级单位其实还有不小的差距！



NIST 2012, Task: MT12 Current 结果

存在的问题

- 传统项目近两次评测进展不显著
  - 进入平台期?与国外顶级单位其实还有不小的差距！
- 各单位技术交流不够深入
  - 技术报告过于笼统
  - 好的系统搞不清好在哪里
- 很多单位利用开源系统，缺少技术特点
  - 能否尽量减少低水平的重复？
- 自动评测指标BLEU有时不能很好的评价翻译结果
  - Round Robin Scoring 结果

Round Robin Scoring

任务代号	人的平均分	最好系统的平均分
汉英新闻（Progress）	0.3864	0.2122
英汉新闻（Progress）	0.4678	0.3195
英汉新闻（Current）	0.5131	0.3075
英汉科技	0.7289	0.4531
日汉新闻	0.5901	0.4593
蒙汉日常用语	0.7815	0.6074
藏汉政府文献	0.5942	0.1858
维汉新闻	0.3852	0.4606
哈汉新闻	0.6254	0.3981
柯汉新闻	0.7036	0.4229

维汉结果分析

问题	为何机器翻译的BLEU值高于人的翻译？
src	يۇقىرى پەللىگە چىقىش ، ساياھەت نۇقتىلىرى ۋە ساياھەتچى كۆپ بولۇشتەك ئۇچ چوڭ ئالاھىدىلىك كۆرۈلدى بۇ يىل ئۆكتەبىر بايرىمىق دەم ئېلىشتىكى ساياھەتتە يولۇچىلار ئېقىمى بالدۇرراق .
以下译文中匹配的为彩色部分，圈出来的部分为不匹配的部分	
H-MT	今年 10月假期的旅游客流高峰 ， 就、旅游和游客的三大特点。
Ref1	今年十一长假旅游，呈现高峰早、热点多、增幅大三大特点。
Ref2	今年的十一长假体现了游客人流提前到达顶峰，景点和游客增多的三个特点。
Ref3	今年的十一长假出现了出游人流高峰出现早，游点多和人数多的三个特点。
Ref4	今年10月节假日旅游量很早就到了高峰期，旅游点多、旅游人数多这样的三大特点。
结论	因为机器翻译有更多和参考答案完全相同的翻译片段，而人的翻译形式更加灵活多样，可见BLEU值高并不一定代表翻译得好。

主要内容

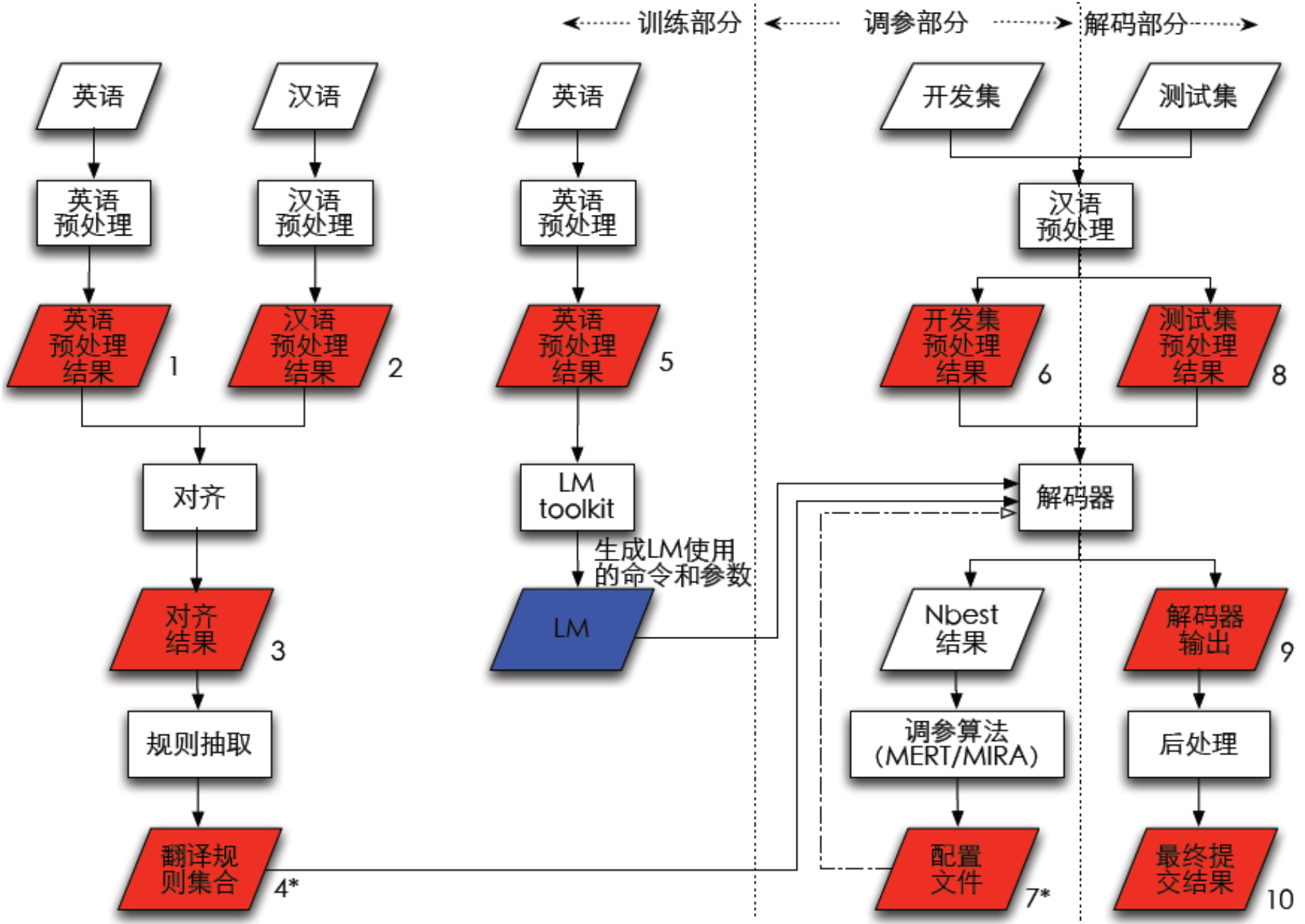
- CWMT机器翻译评测回顾
- 2013年机器翻译评测设想与规划
- 2013年机器翻译评测讨论

CWMT2013年机器翻译评测设想

- 促进参评单位间更深入的技术交流
  - 参评系统提交关键步骤中间结果
    - 黑箱 ->白箱（灰箱）
- 减少低水平的重复
  - 提供基线（**Baseline**）系统
    - 源代码和中间结果数据
- 探索更合理的评价方法
  - 引入人工评测
    - 采用“众包”的方式：参评单位参与人工评测
  - 尝试新的自动评测指标？

CWMT2013年机器翻译评测方案

- 参评系统提交关键步骤中间数据结果
  - 目的：使参评单位的处理流程相对透明
- 关键步骤的选择
  - 统计机器翻译处理流程分析



4*：使用开发集和测试集过滤后的翻译规则集合

7*：最优配置文件
即对测试集进行解码时使用的配置文件

参评系统需要提交的中间数据

- 训练
  1. 源语言预处理结果
  2. 目标语言预处理结果
  3. 词语对齐结果
  4. 翻译规则表
     （使用开发集和测试集过滤后的规则表）
  5. 单语预处理结果
  6. 开发集预处理结果
  7. 参数配置文件

- 解码
  8. 测试集预处理结果
  9. 解码器输出
  10. 最终翻译结果

- 规则翻译系统
  1. 测试集预处理结果
  2. 解码器输出
  3. 最终翻译结果
  4. 测试集每个句子翻译使用到的翻译规则？

基线 (Baseline) 系统实施方案

- 基线系统内容
  - 源代码（或已开源工具的执行命令行）
  - 训练过程的中间数据文件（1-7）
  - 开发集翻译结果
  - 测试集解码中间数据文件（8-10）[评测后提供]
- 基线系统选择
  - Moses （短语模型、层次短语模型）
  - NiuTrans（短语模型、层次短语模型）
  - 参评单位可以获得参评项目的基线系统

基线 (Baseline) 系统实施方案

- 基线系统搭建
  - 每个项目实现一个或多个基线系统
  - 组织方将联合其他单位分别实现不同项目的基线系统

评测项目	基线系统	实现单位
汉英、英汉新闻	Moses	哈尔滨工业大学
汉英、英汉新闻	NiuTrans	东北大学
蒙汉日常用语	Moses	中科院计算所
藏汉政府文献	Moses	厦门大学
维汉新闻	Moses	中科院自动所

人工评价

- 评测组织方提供在线人工评价工具
- 参评单位参与进行人工评价
- 人工评价方法
  - 流利度、忠实度
  - 可接受度
  - 基于排序的方法

忠实度、流利度

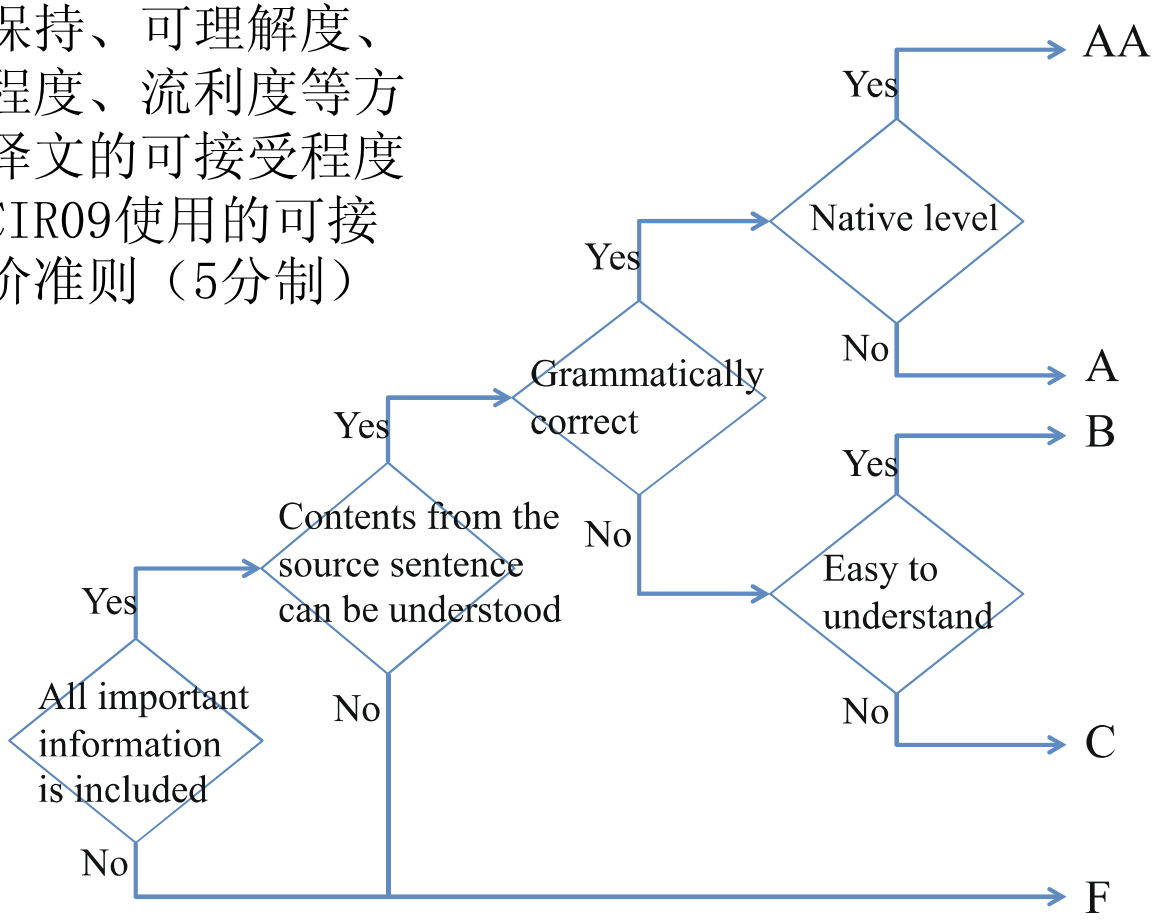
Source: IWSLT07_TEST_463\أين الحمّام؟

Reference: IWSLT07_TEST_463\1\Where is the restroom ?

Translation	Adequacy	Fluency
IWSLT07_TEST_463-spk24_3\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_TEST_463\Where's the toilet?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_AE_TEST_463\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
Annotator: cam **Task:** IWSLT07 Arabic English ASR NIST		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

可接受度

- 从信息保持、可理解度、合语法程度、流利度等方面评价译文的可接受程度
- 图为NTCIR09使用的可接受度评价准则（5分制）



基于排序的方法

Source: IWSLT07_CE_TEST_159\这个 电子 邮件 地址 能 收 邮件 吗 ？

Reference: IWSLT07_TEST_159\1\Can I receive email at this address ?

Translation	**Rank**
IWSLT07_CE_TEST_159\Can you mail this mail address ?	1 Worst 2 3 4 5 Best
Will this e mail mail ?	1 Worst 2 3 4 5 Best
IWSLT07_CE_TEST_159\Can you accept mail this mail address ?	1 Worst 2 3 4 5 Best
IWSLT07_TEST_159\This address E mail, will you accept mail.	1 Worst 2 3 4 5 Best
IWSLT07_CE_TEST_159\Can you mail this e-mail ?	1 Worst 2 3 4 5 Best
Annotator: cam **Task:** IWSLT07 Chinese English Clean	Annotate
Instructions: Rank each **whole sentence** translation from Best to Worst relative to the other choices (ties are allowed).	

CWMT2013年机器翻译评测方案

- 评测方式
  - 评测组织方提供训练、开发、测试数据（受限数据）
  - 评测组织方提供基线系统和训练中间结果数据
  - 参评单位在给定时间内返回翻译结果，提交技术报告和中间结果数据
- 项目设置
  - 汉英新闻
  - 英汉新闻
  - 维汉、蒙汉、藏汉

CWMT2013年机器翻译评测方案

- 评价方法
  - 自动评价
    - BLEU-SBP, BLEU, ……
    - 考虑新的评测指标，如METEOR, TER 等？
  - 人工评价 （部分评测项目）
    - 忠实度、流利度
    - 参评单位参与人工评价

CWMT2013年机器翻译评测方案

- 训练语料

语言对	领域	句对数
英-汉	新闻+科技	584＋90万
蒙-汉	日常用语	6.8万
藏-汉	政府文献	10万

特别感谢新疆大学、青海师范大学、内蒙古大学、厦门大学、西北民族大学为本次评测提供训练和测试语料

-

英-汉	新闻	1,002 (progress), 1,001 (current)
汉-英	新闻	1,003 (progress)
蒙-汉	日常用语	500 (progress)
藏-汉	政府文献	658 (progress)
维-汉	新闻	700 (progress)

CWMT2013年机器翻译评测方案

- 进度安排（关键时间点）

时　间	任　务
2012年9月-10月	**确定评测方案、Baseline系统提供单位，训练、测试数据提供单位**
2012年11月中	组织方提供Baseline系统数据接口说明
2012年11月-12月底	组织方收集整理训练、开发数据 给Baseline系统制作单位提供训练、开发数据
2012年2月1日	**评测大纲发布**
2013年4月15日	报名截止日期
2011年4月20日	评测组织方发放训练数据和开发数据和Baseline系统
2011年6月中旬	**进行评测，参评单位提交结果**
2011年7月初	人工评价打分，评测组织方通知评测结果
2011年7月底	参评单位提交评测技术报告
2011年9月中、下旬	在研讨会上进行研讨

主要内容

- CWMT机器翻译评测回顾
- 2013年机器翻译评测设想与方案
- 2013年机器翻译评测方案讨论

2013年机器翻译评测方案讨论

- 评测方式
  - 提供中间结果的方式是否可行？中间结果数据设置是否合理？
  - 对于规则翻译系统如何处理？
  - 数据受限，是否允许不受限？
- 项目设置
  - 项目设置是否合适？
- 评价方法
  - 人工评测：以流利度、忠实度为主？如何保证一致性？哪几个项目试点？
  - 自动评价：引入哪些新的评价指标？主指标？
- 评测语料
  - 训练语料、测试语料，是否还需增加？
  - 采用进展评测、测试语料是否公布？
- 基线系统
  - 基于短语、层次短语的系统
  - 基于句法的系统？
- 其他评测相关事项
  - 评测时间
  - 参评单位，是否邀请外国单位？
  - 评测项目报名费

感　　谢

感谢所有愿意承担基线系统实现的单位：

哈尔滨工业大学、东北大学、厦门大学

中科院自动化所、中科院计算所、

感谢所有为评测语料作出默默贡献的人们！他们来自：

中国科学技术信息研究所、新疆大学、青海师范大学、内蒙古大学、厦门大学、西北民族大学、自动化所、计算所、南京大学、大连理工大学、东北大学、点通数据有限公司、北京大学计算语言学研究所、哈工大信息检索实验室、哈工大机器智能与翻译研究室、微软亚洲研究院自然语言计算组、搜狗实验室、路透社

感谢计算所赵红梅、谢军、及历届评测小组成员的艰辛劳动！

感谢中文信息学会老师及所有为CWMT评测辛勤工作的人们！

请大家继续关注和支持CWMT机器翻译评测！

谢谢大家！

Thank you ！

INSTITUTE OF COMPUTING TECHNOLOGY