自动化测试-工具复现

Structure-Invariant Testing for Machine Translation

软件学院 余欣然 191250187

目录

Part 01
研究背景
Background

Part 02
实现思路
Idea

Part 03
结果评估
Result

Part 04
研究总结
Summit

PART 01

研究背景

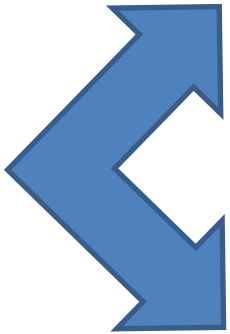
机器翻译：大多基于NMT（neural machine translation）

困难：①难驾驭 ②测试困难：输出复杂+参数众多+缺乏真实测试用例+通常只能找到输入错误

Wrong Translation Example：

Kinder bis 15 Jahre erhalten an ihrem Geburtstag gegen Vorweisen eines gültigen Ausweises den Zooeintritt geschenkt.

origin



Children up to the age of 15 are given free admission to the zoo on presentation of a valid ID.

Google

Children up to the age of 15 are given free admission to the zoo *on their birthday* on presentation of a valid ID.

True

PART 02

实现思路

实现思路

核心观点

相似的语句具有相似的句子结构

核心步骤

文本扩增

语句翻译

句子结构差异分析

实现思路

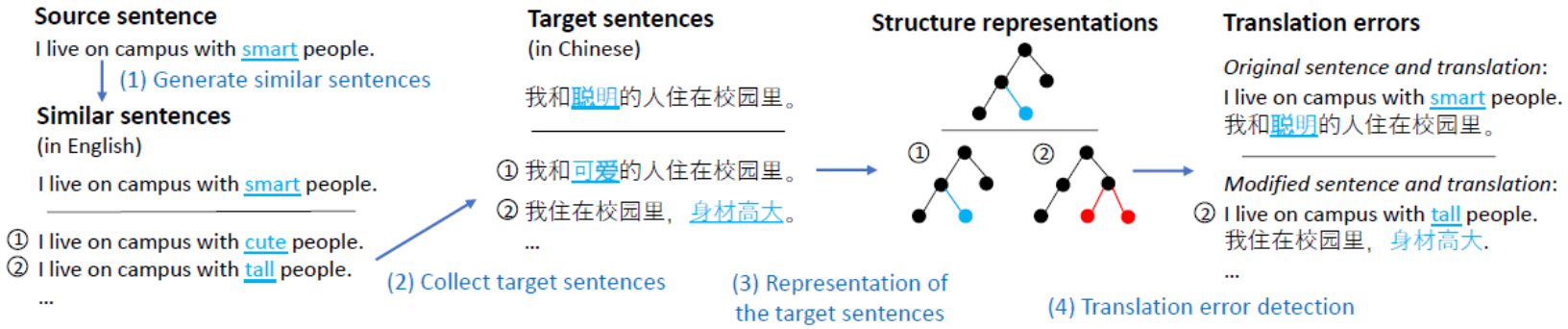


Figure 3: Overview of SIT.

- 数据集：business/politics From CNN
- 翻译工具：Google/Bing/(Baidu/Youdao)
- 文本扩增：bert词嵌入，替换noun/adj.为相同词性的近义词
- 语句依赖树结构分析：Stanford-coreNLP库

实现思路

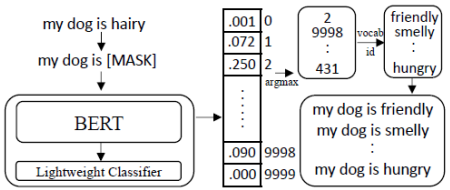


Figure 4: Similar sentence generation process.

① 相似句子生成策略

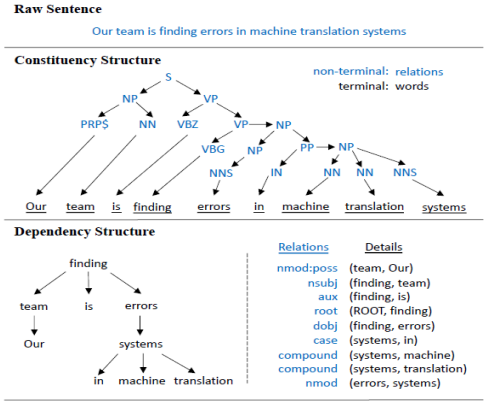


Figure 5: Representing sentence structures; both dependency & constituency relations can be displayed as trees.

③ 目标句结构建模

② 收集翻译语句结果

字符编辑距离

它比较两个字符串，并通过计算将一个字符串转换为另一个字符串所需的最小字符编辑(删除、插入和替换)数量来确定它们之间的匹配程度。

成分集差异

句子的成分应该在两个句子之间保持不变，其中只有相同词性的单个词不同，因此计算两个成分语法列表之间的距离为每个短语类型的绝对差异的总和

依存关系集差异

为了计算两个依赖关系列表之间的距离，我们将每种依赖关系数量的绝对差相加。

④ 比较检查翻译错误

PART 03

结果评估

关于数据集

$$\text{Accuracy}_k = \frac{\sum_{i \in I} \mathbb{1}\{buggy(i,k)\}}{|I|},$$

分母为SIT返回的问题个数

Table 1: Statistics of input sentences for evaluation. Each corpus contains 100 sentences.

Corpus	# of Words/ Sentence	Average # of Words/Sentence	# of Words Total	# of Words Distinct
Politics	4~32	19.2	1,918	933
Business	4~33	19.5	1,949	944

Table 2: Top-k accuracy of SIT.

Google Translate	Top-1 (#buggy issues)	Top-2 (#buggy issues)	Top-3 (#buggy issues)
SIT (Raw)	55.0% (55)	63.0% (63)	66.0% (66)
SIT (Constituency)	61.3% (62)	66.3% (67)	68.3% (69)
SIT (Dependency)	**69.5% (64)**	**71.7% (66)**	**73.9% (68)**
Bing Microsoft Translator	**Top-1 (#buggy issues)**	**Top-2 (#buggy issues)**	**Top-3 (#buggy issues)**
SIT (Raw)	58.8% (60)	69.6% (71)	71.5% (73)
SIT (Constituency)	67.0% (67)	**71.0% (71)**	74.0% (74)
SIT (Dependency)	**70.0% (70)**	**71.0% (71)**	**78.0% (78)**

翻译错误类型

Table 4: Number of sentences that have specific errors in each category SIT (Dep).

Google \ Bing	Under translation	Over translation	Incorrect modification	Word/phrase mistranslation	Unclear logic
Top-1	35 \ 17	9 \ 8	4 \ 2	44 \ 54	27 \ 31
Top-2	48 \ 23	12 \ 15	6 \ 3	59 \ 60	44 \ 41
Top-3	61 \ 35	15 \ 21	10 \ 4	75 \ 93	53 \ 59

Source	After pleading guilty in the Manhattan probe, Cohen also later pleaded guilty to lying to Congress in a case brought by Mueller's website.
Target	在曼哈顿调查中认罪后，科恩后来还对穆勒网站提起的一起案件中的撒谎罪供认不讳。(by Bing)
Target meaning	After pleading guilty in the Manhattan probe, Cohen also later pleaded guilty to lying in a case brought by Mueller's website.

Figure 6: Example of under-translation errors detected.

① 部分词汇未翻译

Source	The investigators were right that the airplane itself was safe.
Target	调查人员认为飞机本身是安全的。(by Google)
Target meaning	The investigators thought that the airplane itself was safe.

Figure 7: Example of over-translation errors detected.

② 部分词汇重复/过度翻译

Source	The South has emerged as a hub of new auto manufacturing by foreign makers thanks to lower manufacturing costs and less powerful businesses.
Target	由于制造成本降低和业务不那么强大，南方已成为外国制造商新的汽车制造中心。(by Google)
Target meaning	The South has emerged as a new hub of auto manufacturing by foreign makers thanks to the reducing manufacturing costs and less powerful businesses.

Figure 8: Example of incorrect modification errors detected.

③ 匹配错误/结构理解错误

Source	The most elite public universities admit a considerably larger percentage of students from lower income backgrounds than do the elite private schools.
Target	最精英的公立大学承认，与精英私立学校相比，低收入学生的比例要高得多。(by Google)
Target meaning	The most elite public universities agree unwillingly that considerably larger percentage of students from lower income backgrounds than do the elite private schools.
Source	The South has emerged as a hub of new auto manufacturing by foreign makers thanks to lower manufacturing costs and less powerful unions.
Target	由于制造成本较低，工会实力较弱，韩国已成为外国制造商新汽车制造业的枢纽。(by Bing)
Target meaning	The South Korea has emerged as a hub of new auto manufacturing by foreign makers thanks to lower manufacturing costs and less powerful unions.

Figure 9: Examples of word/phrase mistranslation errors detected.

④ 词语翻译错误

Source	And attacking a dead man who spent five years as a prisoner of war and another three decades serving the country in elected office, is simply wrong.
Target	并且攻击一名死去的人，他在战争中担任战争囚犯五年，另外三十年担任民选职务的国家，这是完全错误的。(by Google)
Target meaning	And attacking a dead man who spent five years as a prisoner of war and another three decades serving in elected office as a country, is simply wrong.

Figure 10: Example of unclear logic errors detected.

⑤ 句子逻辑不清

效率评估与阈值选取

- 成分分析往往比依存分析更耗时
- 总体是均相对高效
- 阈值挑选时数值较大则报告问题较少，但相对精度更高

Table 5: Average running time of SIT on Politics and Business datasets.

Google \ Bing	Running time (sec)	Translation time (sec)	#Sentence translated	Time of other SIT steps (sec)
SIT (Raw)	1,469 \ 922	1,417 \ 870	2,012	52 \ 52
SIT (Constituency)	1,524 \ 981	1,417 \ 870	2,012	107 \ 110
SIT (Dependency)	1,488 \ 945	1,417 \ 870	2,012	71 \ 75

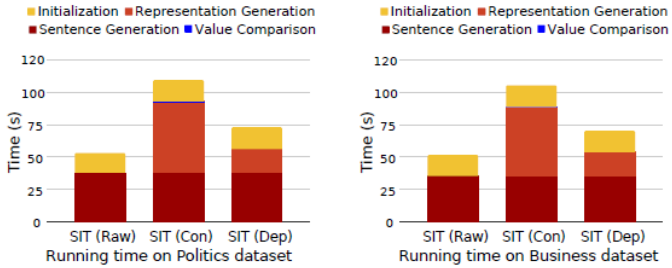


Figure 12: Running time details of SIT (excluding translation time) in testing Google Translate.

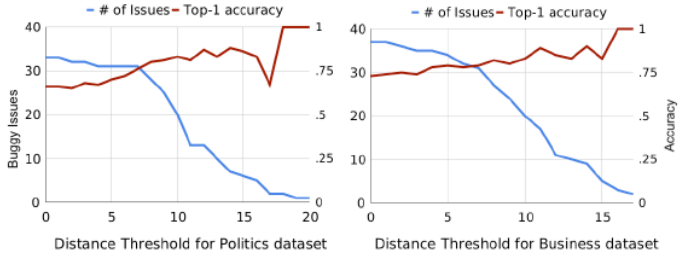


Figure 13: Impact of distance threshold when testing Bing Microsoft Translator.

PART 04

研究总结

主要贡献：

- 介绍了结构不变测试SIT,一种新颖的、广泛适用的机器翻译软件测试方法
- SIT的实现方法,使用BERT生成相似句,并用语法分析来表示句子结构
- 使用从网上抓取的200个句子对SIT进行评估,成功地找到了谷歌翻译和微软翻译的64个和70个错误翻译,具有较高的准确性
- 讨论了SIT发现的各种错误,包括欠翻译、翻译过度、修饰错误、单词/短语误译和逻辑不清,而这些错误都无法通过最先进的度量标准发现

个人想法：

- 在文本扩增中，原论文直接采用了bert预训练好的模型，这样在词嵌入时面对特殊符号如€，人名和缩写时可能会出现困难，相当于间接增大了数据集选择处理的难度。
- 在文本扩增词嵌入时，出现了将所有字符转为小写的操作，这一步在大多数时候能够提高处理的准确度，比如避免句子开头词汇首字母大写被误读，而也存在问题，即人名、公司名和国家地区名等被转换后往往会变为别的意思。
- 同样不可忽视的还有俚语、固定用法的影响，在词嵌入时，虽然已经确保了更改的词语词性相同且非标点符号，但俚语和固定用法的影响仍可能导致翻译后的语句结构完全不同。
- 对于翻译工具本身，不同的翻译工具往往有不同的文化背景基础，如在有道翻译的结果中发现Chan往往被直接译作成龙。除此之外，有的句子可能因为在末尾补上句点后就能出现完全不同结构的翻译结果，因而标点符号也不能简单去除忽略，目前暂时想到的方法仍旧是保留不处理，没有更好的结构性相关方法。

感谢观看