基于语义语法分析的中文语句困惑度评价

作者	何天文，王红
机构	山东师范大学 信息科学与工程学院；山东省分布式计算软件新技术重点实验室
发表期刊	《计算机应用研究》
预排期卷	2017 年第 34 卷第 12 期
访问地址	http://www.arocmag.com/article/02-2017-12-060.html
发布日期	2017-01-23 15:09:11
引用格式	何天文，王红. 基于语义语法分析的中文语句困惑度评价[J/OL]. [2017-01-23]. http://www.arocmag.com/article/02-2017-12-060.html.
摘要	目前用来评价机器翻译系统译文质量的方法主要是由 IBM 提出的 BLEU、TER 和 METEOR 等方法，他们分别以词汇的重现率、译文与参考译文之间的编辑距离和语言学知识等特征作为评价依据，在判定中文句子的困惑度方面具有一定局限性。所以提出在依存语法分析的基础之上，通过对中文句子及其句子主干的语法和语义两方面进行分析得出中文句子的困惑度。实验证明这种方法比通过译文加权改进后的 BLEU 方法准确率高出 4%。
关键词	困惑度，病句，语法，语义，机器翻译
中图分类号	TP391
基金项目	国家自然科学基金资助项目（61672329，61373149，61472233，61572300，81273704）；山东省科技计划资助项目（2014GGX101026）；山东省教育科学规划资助项目（ZK1437B010）；山东省泰山学者基金资助项目(TSHW201502038, 20110819)；山东省精品课程资助项目(2012BK294，2013BK399，2013BK402)

基于语义语法分析的中文语句困惑度评价 *

何天文[1,2]，王 红[1,2]

(1. 山东师范大学 信息科学与工程学院，济南 250358；2. 山东省分布式计算软件新技术重点实验室，济南 250014)

摘 要：目前用来评价机器翻译系统译文质量的方法主要是由 IBM 提出的 BLEU、TER 和 METEOR 等方法，他们分别以词汇的重现率、译文与参考译文之间的编辑距离和语言学知识等特征作为评价依据，在判定中文句子的困惑度方面具有一定局限性。所以提出在依存语法分析的基础之上，通过对中文句子及其句子主干的语法和语义两方面进行分析得出中文句子的困惑度。实验证明这种方法比通过译文加权改进后的 BLEU 方法准确率高出 4%。

关键词：困惑度；病句；语法；语义；机器翻译

中图分类号：TP391

Evaluating perplexity of Chinese sentences based on grammar & semantics analysis

He Tianwen[1,2], Wang Hong[1,2]

(1. School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China; 2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Ji nan 250014, China)

Abstract: Methods currently used to evaluate the translation quality of Machine Translation system are BLEU introduced by IBM, TER, METEOR. These methods do evaluation according to the recurrence rate, the edit distance and the linguistic knowledge, respectively. Although they are superior to other methods, they have some defects in judging the perplexity of Chinese sentences. So we propose a method for evaluating the perplexity of Chinese sentences. Specifically, we analyze the grammar and the semantics together for Chinese sentences and their trunks, based on the parsing of dependency grammar. In the comparison experiments, our method outperforms the mainstream evaluating ones, with accuracy 4 present higher than improved weighting BLEU method.

Key Words: perplexity; wrong sentences; grammar; semantics; machine translation

0 引言

随着计算机技术的快速发展，自然语言处理进入快速发展阶段，并取得巨大成功。在众多的自然语言研究领域中，机器翻译和人机对话无疑是巨大的研究热点，同时也存在着严峻的挑战。

首先，人工翻译面对大数据时代的海量数据，翻译过程中难免会显得力不从心。为了应对这一问题，现在很多自动翻译系统逐渐流行起来[1]。

其次，因为机器翻译的译文评价是一个主观性非常强的工作，评价的高低完全取决于评论者对翻译结果的认同程度，这就导致不同评价者对同一译文的评价不同。自动评价因其快速、廉价、客观的特点吸引了众多的研究，尤其是在机器翻译研究蓬勃发展的今天，需要快速发现译文中的错误、调节翻译系统的参数、评价系统性能、进行不同系统的比较等，使得译文质量的自动评价也成为研究热点。

在人机对话领域中，聊天机器人是一个非常重要的应用，有些也被称为会话代理或对话系统，这项应用的发展几乎代表着人工智能的最前沿的研究进度，要应对图灵测试的挑战[2]，是一个非常热门的研究方向。目前存在的会话代理或对话系统模型可以分为两大类，即：基于检索式模型[3]和基于生成式模型[4]。基于检索式模型根据用户的输入，结合知识库中收录的知识在一些策略指导下组织答案[5]。而生成式模型没有知识库，模型只能根据输入的信息从零开始组织需要返回的答案。生成式模型更加复杂，实现起来也更加困难。这两种模型各有利弊，其中生成模式是模型自己组织句子，语法错误是不可避免的。怎么判断句子是否通顺及句子的组织是否合理[6]，成了一个非常重要的研究课题。

基金项目：国家自然科学基金资助项目（61672329，61373149，61472233，61572300，81273704）；山东省科技计划资助项目（2014GGX101026）；山东省教育科学规划资助项目（ZK1437B010）；山东省泰山学者基金资助项目（TSHW201502038，20110819）；山东省精品课程资助项目（2012BK294，2013BK399，2013BK402）.

作者简介：何天文（1992-），男，山东济南人，硕士研究生，主要研究方向为自然语言处理、机器学习、数据挖掘（sdsfhtw@163.com）；王红（1966-），女（通信作者），教授，博士，主要研究方向为移动社会软件、复杂网络、工作流.

在文献[7]中，提到了用困惑度（perplexity，PPL）判断系统给出的回答是否能作为对人给出的上一句对话的回应，将其作为评价指标来比较各个模型的性能。本文后面内容中将重新定义困惑度，并将其作为评价指标，判断一个句子是否通顺，句子搭配和语法的使用与人们正常使用的情况是否一致。

1 相关工作

目前对翻译质量自动评价的主要方法有 IBM 公司提出的 BLEU（bilingual evaluation understudy）、TER 和 METEOR[8]，还有用来衡量句子合理性程度的困惑度。

BLEU[9]是机器翻译领域的一个非常重要的评价方法，是 IBM 公司在 2001 年提出的一种计算机器翻译与人工翻译的接近程度的方法。它通过计算译文中的 N 元组（N-gram）参考译文中的重现率，得出与参照译文的相似度。针对计算句子不同粒度的词的重现率，公式(1)可以纠正一元组统计词频可能出现的误差[10]。

$$Count_{clip} = \min(Count, Max - Ref - Count) \quad (1)$$

式(1)中，$Count$ 表示某 N 元词在被测译文中的出现次数；$Max - Ref - Count$ 是该 N 元词在一个参考译文中最大的出现次数。公式(2)可以计算整个句子的 BLEU 得分。

$$BLEU = \sum Count_{clip}(N-gram) - c \quad (2)$$

式(2)中，c 为被测译文的长度值。BLEU 值越高则说明翻译的质量越好。

基于编辑距离的自动评测指标有 WER、PER 和 TER 等[11]。其中，TER 是通过计算译文与参考译文之间的编辑距离来评价译文的翻译质量的。TER 的公式可表示为式(3)所示。

$$\text{TER（翻译错误率）} = \frac{\text{编辑次数}}{\text{所有参考译文单词数的平均值}} \quad (3)$$

编辑距离越小表示译文与参考译文的相似程度越高，译文质量越好。

困惑度[12]是统计机器翻译系统中的一种评价指标，采用 N-gram 语言模型，判断机器翻译给出的译文是否是一个合理的句子。一个含有 K 个单词的句子 e，计算其 N 元片段的概率连乘，再对所得的概率取几何平均数，平均数的倒数就是句子的困惑度，其计算公式如(4)所示。

$$PP(e) = \left(\prod_{i=1}^{K} P_{LM}(e_i \mid e_{i-N+1}, \ldots, e_{i-1}) \right)^{-\frac{1}{K}} \quad (4)$$

上式中，困惑度是对概率平均数取倒数得到的，所以越合理的句子的困惑度越小。而且困惑度的计算结果与句子的长度 K 无关，因此困惑度可以用来计算任意长度句子的合理性。

贺敏等[13]使用了有“意义串”相关方法处理微博数据，包括微博中的关键字符串或者有意义的短语，或者是未登录的新词和命名实体。因为这些字符串包含具体语义，是灵活独立的语言单元，能在多种不同语境中使用，故称为“有意义串”。文中将有意义串作为表示微博关键信息的新特征。虽然不能凭借这些小的部分判断句子整体的困惑度，但是至少这个有“意义串”却是完全符合逻辑的，这些被发现的串的困惑度为 0，即有“意义串”是完全合理的。

常若愚 [14]介绍了语义组块分析技术。它是自然语言处理中浅层语义分析和句法分析的代表，旨在解释自然语言中语法和语义之间的关联。组块的概念最早由 Abney（1991 年）提出[15]，组块是比句子更短，比词更长的一个单位。在此基础上进行句子分析能够提高分析正确率，其处理结果对自然语言的分析研究是极具价值的。

Hao Wang[16]研究关于基于抽取的答案生成技术。他们在判断答案是否与问题相匹配时，采取了以下策略：语义相关性判定，逻辑一致性判定，语句行为对齐。这些判定策略提供了在问答系统中判断上下文是否匹配的方案，借鉴文中三种语义相关性判定等的评价，在一个句子的困惑度判定工作中，也可以分别从句法词法、语法语义等方面进行评判。

2 困惑度判定模型

2.1 句子的困惑度

本文提出的困惑度评价方法所关心的句子特征与 BLEU 等计算的 N-gram 的重现率不同。在 Yao K 等的工作中，困惑度在对话系统中用来评价系统生成的语句能否作为对用户上一句的应答，分析的是两个语句之间的相关性。根据这种思路本文试图分析句子内部词语的相关性及其搭配的合理性，并融合自然语言处理中 “有意义串”以及组块与句子规则的思想，提出将语句中不符合中文表达的词语搭配在整个句子中所占比重作为句子的困惑度，此外还提出对句子的词语搭配进行语法和语义两方面的分析（Evaluation of Chinese sentence perplexity based on semantic and syntax analysis， EPBSS），以提高分析的准确性。本文把句子中相邻或者相关联的词语对（WordPair(h,t)，简称 WP(h,t) ）在语法和语义两方面不合理程度的综合度量视作整个句子的困惑度。困惑度高的句子人们读起来会觉得不合逻辑或者是不容易理解。造成这种问题的原因可能是语法的乱用，让人们难以理解句子的结构，如句子“由于她这样好的成绩，得到了老师和同学们的赞扬”；或者是句子的词语搭配有歧义，让人无法理解其想表达的内容，如句子“新生市场苦熬淡季”，歧义：“新生”（1）新学生的市场（2）新产生的市场，还有其他诸多语言使用错误的原因。句子的词语对 WP(h,t)是指在句子的依存语法 (Dependency Parsing，简称 DP)分析结果中，前后直接相连的词语，如句子“我喜欢这个季节”的依存语法分析结果如表 1 所示。词语对 WP 的集合中包含两种词语对：一种是在句子的逻辑结构上前后紧密相邻的，如词语对“喜欢”和“季节”；另一种就是在语法结构上前后相邻的，如词语对“我”和“喜欢”。

2.2 基本思想

EPBSS 方法将整个评价过程分为句子主干评价和细节评价两个步骤，最终依据分析结果将两个评价综合之后给出一个最

终的困惑度得分，具体流程算法如下。

算法 1. EPBSS 核心算法

Input: sentence

Output: evaluation

1: EPBSS(sentence)

return evaluation

2: WordPairsSet[] ← DP(sentence)

3: K_k← GetMainSentence(WordPairsSet[])

4:ms_WordPairsSet ← DP(K_k)

5: $P_{k\text{-}g}$←GrammarAnalysis(ms_WordPairsSet)

6: $P_{k\text{-}s}$←SemanticsAnalysis(ms_WordPairsSet)

7: P_g←GrammarAnalysis(WordPairsSet[])

8: P_s←SemanticsAnalysis(WordPairsSet[])

9:evaluation←GetEvaluation($P_{k\text{-}g}$, $P_{k\text{-}s}$, P_g, P_s)

10:return evaluation

在对句子的困惑度评判中，先对句子做一个句子主干的评价。算法 1 中第 2 行~第 3 行通过依存语法分析抽取句子的主干部分重新组成一句话 K_k，这句话最能表达原句子的核心语义。第 4 行~第 6 行对这句话进行语法语义分析，得出句子 K 的主干句子在语法与语义两方面的困惑度 $P_{k\text{-}g}$ 和 $P_{k\text{-}s}$。这样可以判断出句子想要表达的中心思想的正确性，如果一个句子主干的表达都有问题，那么就可能出现句式杂糅等语病，就很难被人理解，这就会导致整个句子有很高的困惑度。第 7 行~第 8 行进一步从细节上进行分析：细节的分析也是分为语法和语义两方面，是对句子 K 的全部内容进行分析。通过依存语法分析得到所有有关联的词语对的集合，分别判断各个词语对在语法和语义上是否合理，得出困惑度 P_g 和 P_s。第 9 行用这四个困惑度计算整个句子的困惑度。

表 1 依存语法分析结果举例

词语 1	词语 2	依存关系
我_0	喜欢_1	SBV
喜欢_1	-1	HED
这个_2	季节_3	ATT
季节_3	喜欢_1	VOB

2.3 句子主干评价

句子主干评价要判断句子的主干部分有没有语法或者语义搭配的错误。获取句子主干 K_k，具体算法如下。

算法 2. 从完整句子中提取句子主干

Input: WordPairSet/*句子词语对集合*/

Output: mainsentence/*分析得出的句子主干*/

1:GetMainSentence(WordPairSet[])

return (mainsentence)

{ WordPairSet---句子经过依存语法分析得到的词语对，mainsentence---得到的句子主干}

2:**for all** WordPairSet do

3: keyword ← FindHED (WordPairSet[])

4:end for

5:**for all** WordPairSet do

6: Verbs ← FindCOO(keyword)

7:end for

8:**for all** Verbs do

9: **for all** WordPairSet do

10: Objects ← FindObjects(keyword, WordPairSet[])

11: end for

12: Objects ← FindATT(Objects)

13: **for all** WordPairSet do

14: Subjects ← FindSubject(keyword, WordPairSet[])

15: end for

16: Subjects ← FindATT(Subjects)

17:end for

18: mainsentence ← Merge(Objects, Verbs, Subjects)

19:**return** (mainsentence)

算法 2 中第 2-7 行，从句子依存语法分析结果中找出句子的关键词，第 9-12 行在与这个词相关联的词语集合中找到与它是主谓关系的词，如果有并列词也一同找出，找出的这一组词语集合就是整个句子的主语部分。为了保证句子主语的完整性，需要将与主语成定中关系的词与主语拼接起来一同作为主语。还需要检查有没有与主干句子的谓词成并列关系的谓语，如果存在这样的谓语，那么这就是一个有多个谓语的并列句。第 13-16 行是检索与前面找出的主干谓词有动宾关系的词，这些词都是主干句子的宾语，而这些宾语都不需要带有修饰或者是与其成连接关系的复杂成分。第 18 行将找到的所有部分顺序连接得到主干句子。按照上面的方法对前面做了依存语法分析的句子做主干句子提取，得到的结果是“农夫山泉坚持理念，确保品质”。

提取句子主干之后，分别用后文 3.4.1 和 3.4.2 中提到的语法和语义评价方法对句子主干进行评价，得到主干句子的语义困惑度 $P_{k\text{-}s}$ 和语法 $P_{k\text{-}g}$ 困惑度。

对原句子分析过程中，通过依存句法分析解析句子各个成分之间的句法关系，从而得到句子的词语对 WP 集合，下面是对“农夫山泉始终坚持水源地建厂的理念，以确保天然品质”的句法分析数据，如表 2 所示，相对应的依存句法分析的图示，如图 2 所示。

表 2 依存语法分析数据举例

词语 1	词语 2	依赖关系
农夫_0	山泉_1	ATT
始终_2	坚持_3	ADV
理念_8	坚持_3	VOB
，_9	坚持_3	WP
品质_13	确保_11	VOB

表 2 中词语 1 与词语 2 构成一个词语对，词语后面的数字是词语在句子分词后的序列里的位置序号。依赖关系解释的是

词语之间的语法关系。

依存语法分析 API 由哈尔滨工业大学开发的语言技术平台(language technology platform，简称 LTP)[17]提供，它定义了 24 中依存关系如表 3 所示。

上表提供的这些关系可以覆盖句子中词语间的大部分关系，利用这些关系可以对句子结构进行解析。

图 1 描述了一句中文语句进行依存句法分析之后词语之间的依赖关系，每一条弧线连接的两个词都能构成一个词语对。从图中，可以看出句子中的“坚持”是中心谓词，向前是句子主语部分，向后是句子宾语以及其他并列谓词。对依存分析结果的分析表明，针对句法的依存语法分析能把句子中关键的主谓关系，动宾关系和谓语的并列关系都表示出来，如：与 Root 相连的动词是整个句子 K 的核心。

表 3 LTP 平台定义的依存关系关系名及其含义

关系名	关系含义	关系名	关系含义	关系名	关系含义
ATT	定中关系	DE	“的”字结构	VV	连动结构
QUN	数量关系	DI	“地”字结构	DC	依存分句
COO	并列关系	DEI	“得”字结构	HED	核心
APP	同位关系	BA	“把”字结构	IC	独立分句
RAD	后附加关系	ADV	状中结构	CNJ	关联结构
VOB	动宾关系	MT	语态结构	SBV	主谓关系
POB	介宾关系	CMP	动补结构	IS	独立结构

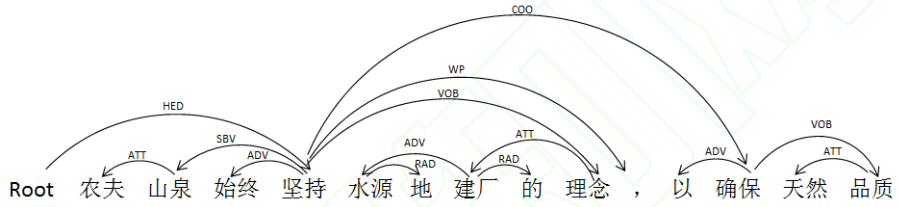

图 1 句子依存句法分析图

2.4 细节评价

在细节评价部分，要对原完整句子进行依存语法分析得到词语对集合 S_{pair}，并将集合 S_{pair} 中的所有词语对都进行语法与语义方面的分析。由于依存语法分析所得的集合 S_{pair} 中存在句子主干分析中已经分析过的词语对 WP_K，在细节分析中应该减小这些词语对的错误对细节评价困惑度的影响，因此，在对句子细节分析时，如果遇到词语对中的两个词语 WP_s 在句子中不是直接相邻的，则为其添加惩罚项 α，细节分析整句困惑度的计算公式如公式(5)所示。

$$P_k=\sum Score(WP_k)+\alpha\sum Score(WP_s) \quad (5)$$

其中，Score(.)为评价函数，WP_k 为在句子中位置直接相连的句子主干中的词语对，WP_s 为不直接相连的词语对。

2.4.1 语法困惑度计算

如果一个句子的语法错误都非常明显，非常多，那么这句话从语法角度就会有较高的困惑度，也就是即便语义组合和搭配都是正确的，乱用的语法或者一些语病就足以让人无法读懂这句话。中文语法的使用非常复杂，使用自己抽取语言语法规则来创建的自动机[18]，利用这种自动机来对中文进行处理，语法分析过程中本文也借鉴这种自动化处理的思想，具体算法如下。

算法 3. 中文语句的语法分析

Input: Sentence,Rules[]

Output: P_g /*句子的语法困惑度得分*/

1:SemanticsAnalysis(Sentence, Rules[])

return P_g

2:C_{pos}← Rules[]

3:S_p← DP (Sentence)

4:S_{pos}← POS (Sentence)

5:$S_{p\text{-}pos}$←Link(S_p, S_{pos})

6:P_g←Match(C_{pos}, $S_{p\text{-}pos}$)

7:return P_g

算法 3 第 2 行首先获取文本语料的词性标注集合 C_{pos}。第 3 行~第 4 行把需要分析的句子先进行依存语法分析得到词语对集合 S_p，再获取其词性标注集合 S_{pos}。

句子“我喜欢这个季节”的词性标注结果如表 4 所示。

第 5 行将两者进行整合，用词性标注所得的词的类型，与依存语法分析的词语对中相应的词语结合，结果如表 5 所示。

第 6 行将这些词性对与文本的词性标注集合 C_{pos} 中通过对正确句子做词性标注得到的语法规则集作匹配就能得出词语对的语法是否被人们在组织正确的句子时使用过，是否是人们惯用的符合人们逻辑的语法。这样将所有词性对分析完成之后，将匹配到的词语对数与所有词语对个数的比值，1 减去这个比值的结果就得到句子在语法层面的困惑度 P_g，计算如公式(6)所示。

$$P_g=1-\frac{RightPairCount}{TotalPairCount} \tag{6}$$

其中，RightPairCount 是正确匹配到的词语对数，TotalPairCount 是所有词语对的个数。

表 4 词性标注结果举例

词语	词性	词语	词性
我	pronoun	这个	pronoun
喜欢	verb	季节	noun

表 5 词性标注与依存语法分析结果整合

词语 1	词语 2	关系
我_pronoun	喜欢_verb	SBV
喜欢_verb	-1	HED
这个_pronoun	季节_noun	ATT
季节_noun	喜欢_verb	VOB

2.4.2 语义困惑度计算

如 2.4.1 中所述，基于抽取规则的方法难以应对语义复杂多变的问题。本文采用语义匹配的方法处理语义问题，具体方法是将待测试句子通过依存语法分析进行拆分得到词语对，再将词语对与语料中的句子进行匹配，如果词语对能在参考语料的同一个句子中匹配到，则认为这种词语搭配方法是在正确的句子中出现过的用法，是可以接受的；而在语料中匹配不到的，也就是正确句子中没有出现过这种使用方法，则认为词语搭配是不合理的，理解起来有困惑的。将一个句子的所有词语对组合全部匹配完之后，就得到了这个句子的正确使用的词语对的数量，再除以这句话分解的词语对的总数，再通过计算就可以得到一个困惑度 P_s，计算过程如公式(7)所示。

$$P_s=1-\frac{RightPairCount}{TotalPairCount} \tag{7}$$

其中，RightPairCount 是语义匹配正确的词语对的个数，TotalPairCount 是所有词语对的数量。

2.5 句子最终困惑度的计算

通过对主干句子和全句分别进行语法分析和语义分析，最终得到 4 个困惑度的评价：主干句子的语法分析困惑度，主干句子的语义分析困惑度，完整句子的语法困惑度和完整句子的语义困惑度。实验数据验证，用计算出来的 4 个困惑度值的平均数可得最终的困惑度 PPL，计算公式如公式(8)所示。

$$PPL=\frac{P_{k-g}+P_{k-s}+P_s+P_g}{4} \tag{8}$$

其中 $P_{k\text{-}s}$ 和 $P_{k\text{-}g}$ 分别是句子主干的语义困惑度和语法困惑度，P_S 和 P_g 分别表示细节分析得到的语义困惑度和语法困惑度。

3 实验与结果分析

本文做了判断机器翻译与人工翻译的相似程度和判断句子的合理程度两个实验。实验中用于句子依存语法分析的平台是哈工大的 TLP。平台包含一系列的汉语语言处理模块，其中包括分词、词性标注、命名实体识别、依存句法分析和语义角色标注等功能。

3.1 机器翻译评价

实验中选用译文加权改进的 BLEU、TER 作为对比模型。唐承亮等[19]的工作中使用了东北大学自然语言处理实验室整理发布的机器翻译测试集数据。该数据集提供的数据中含有中文语句及相应的英语译文。本文将数据集中提供的 1000 句英语句子作为原始数据，将这些句子依次输入到互联网上的翻译系统中进行翻译得到测试数据。用数据集中的中文语句作为模型评价过程中需要参考的文本数据。

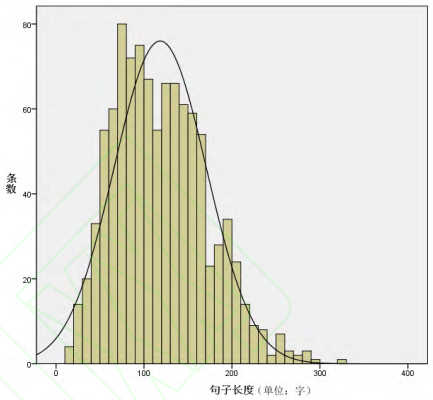

图 2　句子长度分布图

图 2 展示了数据集中句子的长度统计结果，句子的平均长度为 118.39 字，最长句子 326 字，最短句子为 14 字。

为了了解每个机器翻译的句子的真实的困惑度，本文组织了 6 名研究生和两名本科生参照标准译文对机器翻译结果的每个句子的翻译效果进行人工评价，打分采用的是满分十分的评价方法。为了检验 8 人评分的一致性，本文汇总评分数据后计算其 kappa 系数[20]，最终 kappa 系数计算结果为 0.69，根据 kappa 系数分级，实验人员的评价结果具有较高一致性，可以用于实验分析。

通过 SPSS 数据分析软件对实验数据进行回归分析，得出了 4 个困惑度影响 PPL 的权重系数都是 0.25，常数项的值小于 0.0001，所以在计算句子的困惑度时取四个分析得到的困惑度的平均值。

实验过程中，两个模型分别对测试句子进行评价，将评价结果与人工评价结果进行对比得出方法评价的准确率，准确率比较如图 4 所示。

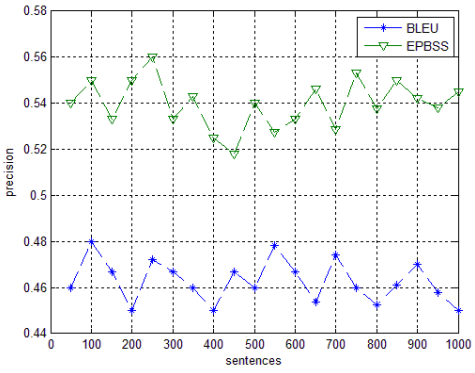

图 3 准确率对比图

从图 3 可以看出 EPBSS 方法准确率要高于基于译文加权改进后的 BLEU 准确率，尤其是数据量增加到 700 句之后，准确率开始趋于稳定。当测试句子数达到 1000 句时，EPBSS 能够判别对 544 句，也就是判别的这些句子比改进后的 BLEU 方法的得分与人工标注的得分更相近。在实验过程中，EPBSS 直接以所有的参考句子作为匹配对象，无论是在语义分析还是语法分析，EPBSS 能够借助更多的参考数据进行分析，同时对句子的依存语法分析和后续的匹配过程，能从待分析的句子中获得更多信息。分析出来的句子主干也为困惑度的评价提供子的另一部分信息。在实验过程中，本文发现主干句子的语法分析比较容易出错，而且相对全句的语法分析，得出的困惑度偏高，这可能与主干句子提取的准确度有一定关系，下一步还要尽量提高提取准确率。EPBSS 的平均准确度达到 55%，比译文加权改进后的 BLEU 平均准确度高出 4%。图中两个方法的线条都不是很平滑，可见两者的鲁棒性都不是很好，相对而言 EPBSS 在语法分析方面受参考语料影响更大。

另外还与 TER 做了对比，各个评分的原始平均得分数据如表 6 所示。表 6 中数据表示的是各个评分标准在不同句子数时得到的平均得分。因为各个标准的评分不尽相同，实验对评分进行标准化处理之后得到评分对比图。

表 6　各个方法的平均评分

句子数	BLEU	TER	EPBSS	人工评价
200	7.117	0.402	5.750	6.5
400	7.252	0.385	5.875	7.0
600	7.278	0.354	5.917	6.7
800	7.082	0.377	5.313	6.0
1000	7.082	0.390	5.350	5.9

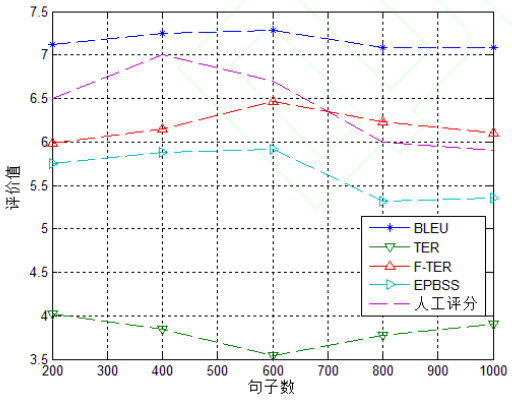

图 4　各方法评分对比图

图 4 给出了不同评价标准的得分的平均值随着测试句子数的变化。其中 F-TER 是根据 TER 得分进行标准化之后得到的评分数据，它只能反映变化趋势并不能反映真实的得分。改进后的 BLEU 的得分相对于人工评价而言则明显偏高。从图上可以看出，EPBSS 的评分与人工评分更为接近，而且随着测试句子数量的增加两者的变化趋势也是一样的，这说明 EPBSS 的评价方法更接近人工评价，体现了该方法的有效性。

3.2　病句判别

因 TER 等方法只是以编辑距离作为评价标准的，所以不适合作为对比模型。实验中，选择 BLEU 用来做对比模型。数据集是本文从历年中高考语文试题中摘录的 400 个有成分残缺语病的句子，从《人民日报》上摘取了 400 个表达正确的句子，将这两部分句子作为测试数据，同时本文还根据修改病句的答案，将修改后的句子，作为模型在进行语义计算和语法计算过程中需要参照的正确的句子。EPBSS 模型在病句和正确句子上的准确率对比如图 5 所示。

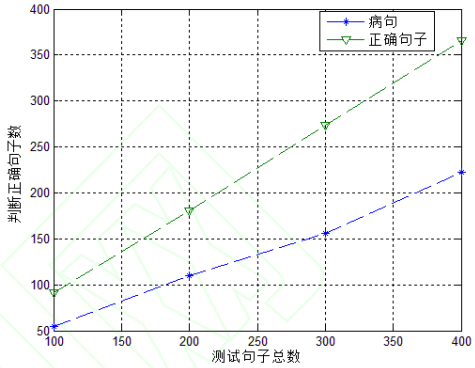

图 5　准确率对比图

图 5 表示了 EPBSS 模型在正确句子和病句两组句子中分类的准确率情况，从图上可以看出 EPBSS 在正确句子上有较高的准确率，而在判断病句的准确率则较低，通过对病句数据的进一步分析，本文认为是 EPBSS 模型中对提取出来的主句的成分残缺更加敏感，而对于句子中的修饰成分存在的问题判断准确率不是很高。

针对模型的准确率，本文还与 BLEU 模型对分类正确的句子数进行了统计，统计结果如图 6 所示。

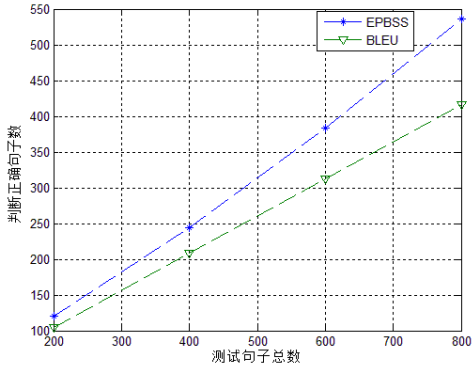

图 6　分类结果对比图

图 6 中统计的句子数是对句子是否是病句判断正确的句子总数，曲线的斜率代表模型的准确率。从图上可以看出，EPBSS 的准确率要更高一些，BLEU 主要是以词语重现率为评价标准，而 EPBSS 随着数据量的增加，能有更多可以参考的数据用来在语义分析和语法分析中做匹配，随着可参考语料的增加，能匹配到的语义搭配和语法都会增加。

4 结束语

本文主要介绍了通过对给定句子的句子主干和完整的句子在依存语法分析的基础上进行的语法和语义两方面进行分析评价得出困惑度，文中的实验结果证实 EPBSS 是有效的。在机器翻译领域，句子困惑度的评价，能够用来评价机器翻译模型给出的翻译结果是不是符合人们正常表达逻辑的，机器翻译模型的翻译结果的困惑度越低，就意味着这个句子翻译得越好。在人机对话领域中，困惑度可以用来评价基于生成模型的对话系统给出的答案是否符合人的正常的理解逻辑。

EPBS 方法对句子的局部合理性判定还存在一些问题，在句子整体不通顺时，句子局部的合理性也可以降低句子整体的困惑度。当方法参照大规模语料判定一个句子的困惑度时，对语料所包含的内容有较强的依赖性，还需要进一步改进参照语料的组织形式。语法分析的准确度受词语对词性标注的准确度影响较大。目前 EPBS 仅仅分析了句子的语义语法信息而用于分析翻译译文还需要分析两种语言相应的语言学信息，如语言的贴合度、指代准确性以及成分的否定关系等。今后的工作将继续在以上几个方面进一步展开。

参考文献：

[1] 秦颖. 翻译质量自动评价研究综述[J]. 计算机应用研究, 2015, 32(2): 326-225.

[2] 黄炎孙. 人工智能的符号主义立场研究[D]. 北京: 化工大学, 2014.

[3] 钱宏泽. 基于中草药语义网的自动问答系统的研究与实现[D]. 杭州: 大学, 2016.

[4] 张辉, 刘奕群, 马少平. 文本情感分类中生成式情感模型的发展[J]. 计算机应用研究, 2014, 31(12).

[5] Denny Britz, Deep learning for chatbots, part 2 – implementing a retrieval-based model in Tensorflow[EB/OL], http: //www. wildml. com/2016/07/.

[6] 梁华参. 基于短语的统计机器翻译模型训练中若干关键问题的研究[D]. 哈尔滨: 哈尔滨工业大学, 2013.

[7] Yao K, Zweig G, Peng B. Attention with Intention for a Neural Network Conversation Model[J]. Computer Science, 2015.

[8] Papineni K, Roukos S, Ward T, *et al*. BLEU : a method for automatic evaluation of MT . IBM research division, RC22176(W0109-022)[R]. Watson Research Centre, 2001.

[9] 王茜. 基于 BLEU 的英语翻译自动评分研究[J]. 安徽电子信息职业技术学院学报, 2010, 09(4): 65-66.

[10] 张剑, 吴际, 周明. 机器翻译评测的新进展[J]. 中文信息学报, 2003, 17(6).

[11] 赵红梅, 刘群. 机器翻译及其评测技术简介[J]. 产品安全与召回, 2010(1): 36-41.

[12] 刘章. 面向服务机器人的口语对话系统和语言模型技术研究[D]. 中国科学技术大学, 2014.

[13] 贺敏, 王丽宏, 杜攀, 等. 基于有意义串聚类的微博热点话题发现方法[J]. 通信学报, 2013(S1): 256-262.

[14] 常若愚. 汉语语义组块识别研究[D]. 杭州电子科技大学, 2015.

[15] 李优, 黄德根. 一个基于规则的汉语句子组块识别系统[C]// 中国模糊逻辑与计算智能联合学术会议. 2005.

[16] HaoWang, Zhengdong Lu, Hang Li, *et al*. A Dataset for Research on Short-Text Conversation [C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 935–945. .

[17] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: A Chinese language technology platform[C]//Proc of the Coling Demonstrations. 2010: 13-16

[18] 孙连恒, 杨莹, 姚天顺. OpenE: 一种基于 n2gram 共现的自动机器翻译评测方法[J]. 中文信息学报, 2004, 18(2).

[19] 唐承亮, 肖海青, 向华政. 基于文字 RGB 颜色变化的脆弱型文本数字水印技术[J]. 计算机工程与应用, 2005, 41(36): 6-8.

[20] 华琳, 阎岩, 张建. 关于对诊断一致性 Kappa 系统的探讨[J]. 数理医药学杂志, 2006, 19(5): 518-520.