北京大学学报(自然科学版), 第 49 卷, 第 1 期, 2013 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 49, No. 1 (Jan. 2013)

基于特征比较和最大熵模型的统计机器翻译错误检测

杜金华† 王莎

西安理工大学自动化与信息工程学院, 西安 710048; †E-mail: jhdu@xaut.edu.cn

摘要 首先介绍 3 种典型的用于翻译错误检测和分类的单词后验概率特征, 即基于固定位置的词后验概率、基于滑动窗的词后验概率和基于词对齐的词后验概率, 分析其对错误检测性能的影响; 然后, 将其分别与语言学特征如词性、词及由 LG 句法分析器抽取的句法特征等进行组合, 利用最大熵分类器预测翻译错误, 并在汉英 NIST 数据集上进行实验验证和比较。实验结果表明, 不同的单词后验概率对分类错误率的影响是显著的, 并且在词后验概率基础上加入语言学特征的组合特征可以显著降低分类错误率, 提高译文错误预测性能。

关键词 错误检测; 词后验概率; 语言学特征; 最大熵分类器

中图分类号 TP391

Error Detection for Statistical Machine Translation Based on Feature Comparison and Maximum Entropy Model Classifier

DU Jinhua†, WANG Sha

Faculty of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048; † E-mail: jhdu@xaut.edu.cn

Abstract The authors firstly introduce three typical word posterior probabilities (WPP) for error detection and classification, which are fixed position WPP, sliding window WPP, and alignment-based WPP, and analyzes their impact on the detection performance. Then each WPP feature is combined with three linguistic features (Word, POS and LG Parsing knowledge) over the maximum entropy classifier to predict the translation errors. Experimental results on Chinese-to-English NIST datasets show that the influences of different WPP features on the classification error rate (CER) are significant, and the combination of WPP with linguistic features can significantly reduce the CER and improve the prediction capability of the classifier.

Key words error detection; word posterior probability; linguistic features; maximum entropy classifier

近年来, 随着基于统计方法的机器翻译(SMT)的发展, 涌现出多种不同类型的机器翻译(MT)系统, 如基于短语、基于层次短语及基于句法的机器翻译系统等等, 并且翻译性能得到了显著提高[1−3]。译文质量自动评价是统计机器翻译研究的一个热点, 可以分为有参自动评价和无参自动评价两种。在软件本地化领域(如大型跨国软件公司为适应中国客户需求, 对英文软件进行本地化中文显示), 后者是指在没有参考答案的情况下, 自动对译文质量给出置信得分或对译文中的翻译错误进行识别和分类, 以帮助译文编辑人员快速定位翻译错误位置, 提高工作效率。

为了提高机器翻译译文质量, 自动错误检测与分类在 MT 输出后处理中起着至关重要的作用, 一方面可以帮助后编辑人员提高工作效率, 另一方面可根据翻译错误分析对应源语言端的翻译难度, 从

国家自然科学基金(61100085)、陕西省教育厅专项科研计划项目(11JK1029)和西安理工大学青年科技研究计划项目(105211017)资助
收稿日期: 2012−05−29; 修回日期: 2012−08−17; 网络出版日期: 2012−10−26
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.001.html

而通过变换源语言端输入进行重解码, 进而提高翻译性能。因此, 译文错误判断、分类和分析是 SMT 技术发展和应用的重要研究内容之一。

目前, 译文错误检测方法多采用系统特征如单词后验概率(word posterior probability, WPP)进行译文置信度估计, 并采用浅层句法或语义知识作为辅助, 以降低分类错误率, 提高错误预测性能。在系统特征方面, 多种根据 *N*-best 或词图计算用于置信度估计的 WPP 方法被提出, 并且在机器翻译译文错误检测中已经得到了广泛应用[4−7]。之后, 一些研究人员尝试利用其他知识源作为特征, 并与 WPP 置信特征相结合来进行错误检测, 例如深层次句法或深层次语义特征等[8−10]。然而, 由于多数情况下外部特征的复杂性以及与具体语言相关等原因, 有效的并且通用的特征并不容易被抽取。因此, 目前来看, 单词后验概率、词以及词性等词汇化特征仍然起着主要作用。

本文首先研究和分析 3 种基于不同方法的单词后验概率特征对统计机器翻译错误检测的影响, 然后分别将其与语言学特征如词性、词及由 LG 句法分析器抽取的句法特征等进行组合, 利用最大熵分类器预测翻译错误, 并在汉英 NIST 数据集上进行实验验证和比较。实验结果表明: 1) 不同的单词后验概率对分类错误率的影响是显著的; 2) 在词后验概率基础上加入语言学特征的组合特征可以显著降低分类错误率, 提高译文错误预测性能。

1 相关研究工作

国内外众多研究人员对译文置信度估计问题做了广泛而深入的研究。2004 年 Blatz 等[6]对机器翻译的置信度估计基本方法进行了改进, 将基于神经网络的特征和贝叶斯分类器相结合, 预测基于句子级和词级的翻译错误。该方法所用到的置信估计特征主要有: 根据 *N*-best 列表计算得到的单词后验概率, 基于 SMT 翻译模型的系统特征, 从 WordNet 中抽取的语义特征, 以及浅层的句法特征, 等等。实验结果表明单词后验概率的泛化能力比基本语言学特征的表现更好。

Ueffing 等[5,7]对单词后验概率特征进行了更深入的分析和研究, 分别采用目标位置窗、相对频率、系统模型等方法计算后验概率。实验结果验证了单词后验概率在译文置信估计中的有效性。Ueffing 等的工作大大推动了统计机器翻译译文置信估计的研究与应用。

2009 年, Specia 等[8−9]在面向本地化领域的计算机辅助翻译的译文质量置信估计方面做了大量的研究, 为译文置信估计方法向实用化发展做出了贡献。通过对后编辑翻译工作中的机器翻译片段进行句子级的二值评分, 将句子分为好与差两类。采用的置信度特征为“黑盒子特征”, 即在只给定输入(源语言句子)和翻译结果输出(目标语言句子)条件下, 如何从任意 MT 系统中得到更为通用和泛化的特征, 如源语言和目标语言句子的长度及其之间的比例关系, 以及源语言输入句子和语料中用于训练 SMT 系统的句子之间的编辑距离的关系特征等。利用这些特征, 根据预测的理想置信度得分控制分类阈值, 实验结果大大提高了基于句子级的机器翻译译文质量的置信估计。

2010 年 Xiong 等[10]将基于 LG 句法分析而抽取的句法特征与词汇化特征、词后验概率特征相组合, 利用基于最大熵模型的二值分类器来预测机器翻译假设中每个单词的类别, 即正确(correct)或不正确(incorrect)。实验结果表明在译文翻译错误检测任务中, 有效的语言学特征不仅可以显著降低分类错误率, 并且其性能表现要优于词后验概率特征。另外, 语言学特征与单词后验概率特征的组合特征比单个特征的优势更明显, 可以显著降低分类错误。

2011 年 Bach 等[11]采用 Goodness 方法来衡量机器翻译译文的置信度。针对之前研究的不足之处, 如源端信息不充分、复杂特征提取困难等, Bach 等通过提取更加丰富的源端信息特征集合, 细化翻译错误类型(如将翻译假设中的单词分成 4 类), 结合基于句子和词级的特征对译文质量进行置信估计。通过预测带有置信度得分的 SMT 输出中每个单词的错误类型, 进而扩展到句子级, 最终应用于 *N*-best 列表重排序任务中, 以提高 MT 翻译质量。

以上研究在不同程度上对单词后验概率和浅层语言学特征进行了研究和验证, 表明了其有效性。但在面向更实用的翻译错误检测与标注任务中(如计算机辅助翻译任务中的后编辑工作), 需要机器翻译系统不仅能够输出高质量译文, 并且能够有效地判断翻译错误的位置, 以方便后编辑人员快速定位, 提高工作效率。因此, 本文主要进行两方面的研究工作: 1) 单词后验概率特征有较强的通用和泛化能力, 其计算和抽取过程与具体语言无关, 具有更广泛的应用前景, 因此, 本文分析和验证基于不

同方法的单词后验概率对翻译错误检测性能的影响; 2) 针对具体语言的翻译错误检测任务, 语言学特征可以有效提高预测能力, 降低分类错误率, 因此, 本文将不同的词后验概率特征与语言学特征相结合, 采用 NIST 机器翻译评测任务 2005 年、2006 年、2008 年中到英的翻译数据集进行对比验证, 以表明语言学特征的有效性及词后验概率在不同错误检测任务中的一致性。

2　词后验概率特征

从语音识别到目前的统计机器翻译译文置信估计, 词后验概率一直作为一个主要且有效的置信估计特征。就统计机器翻译而言, 给定源语言输入, 一个词的后验概率指的是该词在目标语言句子即机器翻译输出结果发生的概率。从一般意义上来讲, 其基本思想是给定源语言输入, 如果目标语言的翻译假设中一个词出现的概率(或频次)很高, 则其为正确翻译的可能性就很大。但实际情况表明, 单纯基于数学意义上的词后验概率计算, 有时并不能真实反映目标词是否一定正确, 这是因为影响目标词后验概率准确程度的因素主要有: 1) 训练语料或翻译模型对源语言输入的覆盖率(coverage)。过低的覆盖率会导致大量的集外词出现在翻译假设中, 从而使得根据 N-best 列表所计算的集外词的后验概率较高; 2) 训练语料与源语言输入的领域适应度(adaptation)。数据领域的差异会造成翻译假设与源语言输入在意义表达上的差异较大, 但意义表达不准确的目标词在 *N*-best 列表中出现的概率可能会很高, 从而导致翻译假设的可懂度较低。因此, 本文选取 3 种典型的词后验概率计算方法进行分析和对比, 以验证其对错误检测的影响。

词后验概率的一般数学描述为: 对于统计机器翻译系统 S, 给定源语言输入句子 f_1^J, 其 *N*-best 输出记为 $e_{n,1}^{n,I_n}$, 其中 $n=1,...,N$; e_n 表示 *N*-best 列表中第 n 个翻译假设, 每个翻译假设的翻译概率记为 $p(f_1^J, e_{n,1}^{n,I_n})$, 则面向译文错误检测任务的单词后验概率可表示为计算 *N*-best 列表中翻译概率最高的翻译假设即 1-best 结果中每个目标词的后验概率, 记为 $p(e_1^i / f_1^J)$。

如前所述, 在基于 *N*-best 列表计算词后验概率的方法中, 词的后验概率可以从 *N*-best 列表中翻译假设的后验概率得到。在已知句子的后验概率情况下, 词后验概率可以通过多种形式对包含目标单词的所有句子的概率求和得到。在 *N*-best 列表中由于各种编辑操作如删除、插入、调序等原因, 单词 e 出现在任意翻译假设中的位置不是固定不变的, 并且其出现频率也是不同的。这样, 多种不同的词后验概率计算方法被提出。

基于以上分析, 本节描述 3 种典型的根据目标语言的 *N*-best 列表计算词后验概率的方法, 试图从词对齐、目标词出现的位置、上下文关系等方面对后验概率计算及其对翻译错误检测的影响进行深入分析, 从而为单词后验概率在置信估计的实际应用提供依据。本文研究的词后验概率全部从统计机器翻译系统输出的 *N*-best 列表中产生。

2.1　基于固定位置的单词后验概率

基于固定位置的单词后验概率计算方法的基本思想是: 给定源语言输入, 则位置 i 的目标单词 e 的后验概率由 *N*-best 列表中其他翻译假设在对应位置 i 出现 e 的概率求和得到, 如下式所示:

$$p_i(e/f_1^J)=\frac{\sum_{n=1}^{N}\delta(e_{n,i},e)\cdot p(f_1^J,e_{n,1}^{n,I_n})}{\sum_{e'}\sum_{n=1}^{N}\delta(e_{n,i},e')\cdot p(f_1^J,e_{n,1}^{n,I_n})},\qquad(1)$$

$p_i(e/f_1^J)$ 表示给定源语言输入 f_1^J, 最优翻译假设 e_1^I 中在位置 i 的目标单词 e 的后验概率; $\delta(x,y)$ 代表克罗奈克函数; $p(f_1^J,e_{n,1}^{n,I_n})$ 表示 *N*-best 列表中翻译假设的后验概率, 由机器翻译系统给出。

2.2　基于滑动窗的单词后验概率

2.1 节中固定位置后验概率方法的缺点是严格确定目标词的位置, 这与 *N*-best 列表的实际情况不符, 即翻译假设的句长在某一范围内是动态变化的, 使得由于不同的翻译假设因句长不同等因素的影响, 导致在同一位置 i 上的目标词 e_i 有可能是不同的, 即对应不同的源语言词, 因此, 难以保证单词 e 出现的位置 i 是固定不变的, 但有可能出现在位置 i 附近, 即上下文中。因此, 若将初始的固定位置 i 变为一个动态值, 使其能在初始值的某一范围内滑动, 当目标单词出现在限定范围内时, 参与后验概率计算。这就是在固定位置基础上改进的基于位置窗的方法思想。

记位置窗为 $i\pm t$, t 为窗口大小, 为自然数。如果单词在位置窗范围内出现, 同样认为单词出现在当前翻译假设。因此, 单词的后验概率可以由此窗

范围内的位置上单词出现的后验概率之和来决定，计算公式如下:

$$p_{i,t}(e/f_1^J)=\sum_{k=i-t}^{i+t}p_k(e/f_1^J)。\quad(2)$$

2.3 基于词对齐的单词后验概率

基于词对齐即采用 Levenshtein 对齐计算词后验概率的方法最早是由 Ueffing 等[5]提出的，并通过实验验证了其有效性。其基本思想是将生成的最优目标译文 e_1^I，即 1-best 与 *N*-best 列表中其他句子做 Levenshtein 对齐，也就是最短编辑距离对齐，然后根据对齐信息，最优译文 e_1^I 中每个单词 e 的后验概率 $p(e/f_1^J,e_1^I)$ 可以通过计算具有 Levenshtein 对齐位置 i 上包含 e 的所有翻译假设的概率之和得到。

具体来讲，记 $L(e_1^I,e_{n,1}^{n,I_n})$ 为最优译文 e_1^I 和 *N*-best 列表中其他翻译假设 $e_{n,1}^{n,I_n}$ 之间的 Levenshtein 对齐关系，则对于位置 i 的单词 e，具有 Levenshtein 对齐关系的后验概率可由下式给出:

$$p_{lev}(e/f_1^J,e_1^I)=\frac{p_{lev}(e,f_1^J,e_1^I)}{\sum_{e'}p_{lev}(e',f_1^Je_1^I)}。\quad(3)$$

其中，

$$p_{lev}(e,f_1^J,e_1^I)=\sum_{n=1}^{N}\delta(e,L_i(e_1^I,e_{n,1}^{n,I_n}))\cdot p(f_1^J,e_{n,1}^{n,I_n}),\quad(4)$$

$\delta(x,y)$ 代表克罗奈克函数，即

$$\delta(x,y)=\begin{cases}1 & x=y,\\ 0 & 其他。\end{cases}$$

$p(f_1^J,e_{n,1}^{n,I_n})$ 表示 *N*-best 列表中翻译假设的后验概率，本文使用统计机器翻译系统生成此翻译假设的概率值。

3 语言学特征

我们已经知道在统计机器翻译结果中常见的错误主要有语法、句法及词序等错误，而第 2 节中所描述的词后验概率本质上是由系统内部特性来决定的，无法提供足够的语法或句法知识用于错误检测。因此，与统计机器翻译系统或其他自然语言处理任务类似，在对译文置信估计或错误检测中，也加入多种外部知识源如句法、语义等语言学特征以提高检测准确率。本节主要介绍两类常用的语言学特征，即句法特征和词汇化特征。

3.1 语法特征

在统计机器翻译系统中，尤其是基于句法的统计机器翻译系统，如树到串、串到树、树到树等所引入的句法知识，一般是通过句法分析器(Syntactic Parser)对源语言或目标语言进行句法分析后得到的，而且通常是符合语法规范的正确知识。因此，我们可以直观地认为，如果对翻译假设进行句法分析，在句法分析失败或分析结果中不合语法的地方，则所在位置单词或该单词的上下文是翻译错误的可能性更大。基于以上考虑，可以引入句法分析，从中抽取句法特征作为外部资源用于翻译错误检测。

Xiong 等[10]采用 LG 句法分析器抽取句法特征，并根据所分析单词与上下文的语法连接关系来定义句法特征。该 LG 句法分析器的特点是当分析器无法对整个句子进行分析时，则忽略问题单词以找到其他剩余单词之间的联系，从而完成句法分析过程。被忽略的单词就成为与句子中其他单词没有连接的单词，称为 null-link 单词。这些与其他单词没有联系的 null-link 单词存在语法错误的可能性更大，而存在连接关系的单词则符合语法规范的可能性更大。因此，利用此信息可以来定义二值句法特征，如下式所示:

$$\text{link}(e)=\begin{cases}\text{yes}, & e\text{与其他词存在连接关系},\\ \text{no}, & \text{其他},\end{cases}\quad(5)$$

e 表示翻译假设中的单词。关于此特征详细过程可参阅文献[10]。

3.2 词汇化特征

词汇化特征是自然语言处理领域分类任务中常用的特征，主要有实体单词和词性等。在统计机器翻译错误检测任务中，词汇化特征也可以作为外部资源在一定程度上降低分类错误率。对于译文中的多数目标单词，可以认为其出现频率与分类结果存在一定关系，即频率高的单词序列和词性标注序列为正确的概率要大于出现频率较少的单词序列和词性标注序。本文采用文献[10]中的方法，考虑单词及词性标注的上下文信息，即滑动窗的方法，对于每个单词或词性标注，取其前 2 个和后 2 个及其本身构成特征向量。其具体形式如下所示:

单词实体特征 Word: $(w_{-2},w_{-1},w,w_1,w_2)$,

词性标注特征 Pos: $(\text{pos}_{-2},\text{pos}_{-1},\text{pos},\text{pos}_1,\text{pos}_2)$。

4 实验与分析

4.1 中文–英文翻译系统

SMT 系统 本文所使用的统计机器翻译系统为基于短语的 Moses 系统[12]，翻译语言对和方向为

中到英翻译。

训练语料　翻译模型训练语料为 LDC 提供的数据集, 主要包括香港新闻、FBIS、ISI 汉英网络数据及新华新闻等, 共计 3397538 句对。翻译模型中短语长度限定为 10 个词。语言模型为五元, 训练语料由以上双语语料的英文部分加英文 Gigaword 的新华部分组成, 共计约 1000 万句。

开发与测试数据　机器翻译系统的开发集使用 NIST 机器翻译评测任务 2006 年 current 集, 共计 1664 句, 每个源语言输入对应 4 个参考答案。测试集为 NIST MT 2005 数据集和 NIST MT 2008 数据集, 分别包含 1082 句和 1357 句, 每句对应 4 个参考答案。在测试集解码过程中, 设定每个源语言句子输出 10000 个最好翻译假设, 即 N=10000。统计机器翻译系统在两个集合上的性能见表 1。

表 1　统计机器翻译性能和标注为"正确"词比例

Table 1　SMT performance and the ratio of correct words (RCW)

数据集	BLEU4/%	WER/%	TER/%	RCW/%
NIST MT 2005	33.17	69.50	61.40	41.59
NIST MT 2008	25.97	69.79	63.56	37.99

4.2　翻译错误检测任务

文献[10]采用最大熵分类器对翻译错误进行检测和分类, 并将翻译假设中的单词分为 2 类: 正确(correct)和错误(incorrect)。为进行对比验证, 本文也采用此两类类别和最大熵分类器。

数据标注　本文利用 TER 工具包[13]中的 WER 准则来确定翻译假设中每个单词的真实分类结果。对于分类任务中的开发集和测试集的分类标注, 首先将源语言句子的翻译假设分别与 4 个参考译文进行词对齐, 并得到 WER 得分, 选择得分最小的参考译文为标注基准, 即编辑距离最小的译文为标准, 观察对齐结果, 在对齐的同一位置上如果翻译假设中的单词与参考译文中的单词一致标为"正确", 反之为"不正确", 分别记为 c (correct)或 i (incorrect)。

开发集与测试集　4.1 节翻译任务中使用了两个测试集, 即 NIST MT 2005 和 NIST MT 2008 数据集。在翻译错误检测任务中, 将 NIST MT 2008 数据集作为最大熵模型参数训练的开发集, 将 NIST MT 2005 数据集作为分类任务的测试集。SMT 系统产生的 NIST MT 2008 翻译结果中, 1-best 翻译假设包含 38587 个单词, 其中标为 c 的为 14658 个, 标为 i 的 23929 个。NIST MT 2005 翻译结果的 1-best 翻译假设包含 36497 个单词, 其中标为 c 的 15179 个, 标为 i 的 21318 个。两个数据集中样本为"正确"类别(ratio of correct words, RCW)的比例见表 1。

最大熵分类器　采用开源的最大熵工具包[14], 高斯先验设为 1。

4.3　评价准则

采用的错误检测方法有效性评价标准有: 分类错误率(CER)、准确率(P)、召回率(R)和 F 准则。

分类错误率如下式所示:

$$\mathrm{CER}=\frac{\text{分类类别为错的词数}}{\text{总词数}}。\tag{6}$$

在中到英的翻译错误检测分类任务中, 因翻译假设中真实类别为"不正确"的个数要大于真实类别为"正确"的个数, 所以在确定分类错误率的基线水平时, 通常的做法是, 将"正确"的单词全部标为"不正确"时所得到的评价准则得分, 即: 分类错误率的基线水平="正确"样本个数/总的样本个数。

准确率为分类器将真实为类别 i 的单词准确分类的个数 n_m 与分类器标记为 i 的单词总个数 n_t 之比, 即

$$P=\frac{n_m}{n_t}。\tag{7}$$

召回率为分类器将真实为类别 i 的单词准确分类的个数 n_m 与真实类别为 i 的单词总个数 n_g 之比:

$$R=\frac{n_m}{n_g}。\tag{8}$$

F 准则为准确率和召回率的权衡, 即

$$F=\frac{2\times P\times R}{P+R}。\tag{9}$$

4.4　错误分类实验

最大熵分类器的特征函数为考虑上下文的特征向量, 即除每个当前特征变量外, 同时考虑其前后两个。

实验设计: 1) 对 3 种典型的单词后验概率特征进行分类实验, 比较其性能并分析; 2) 对单独语言学特征进行最大熵模型分类实验, 并进行分析; 3) 将 3 种典型的单词后验概率特征分别与语言学特征组合, 进行分类实验, 并进行比较和分析。

4.4.1　基于单词后验概率特征的分类实验

基于 3 种典型的单词后验概率的分类实验结果见表 2。

表 2 3 种典型的 WPP 特征的错误检测结果

Table 2 Results of three typical WPP features for translation error detection

特征	CER/%	*P*/%	*R*/%	*F*/%
基线	41.59			
WPP_Dir	40.48	63.44	72.46	67.65
WPP_Win	39.70	63.82	73.95	68.51
WPP_Lev	40.12	60.24	92.07	72.83

表 2 中 WPP_Dir 表示基于固定位置的单词后验概率特征; WPP_Win 表示基于滑动窗的单词后验概率特征, 滑动窗 $t=2$; WPP_Lev 表示基于 Levenshtein 对齐的单词后验概率特征。在将 *N*-best 列表中的 1-best 翻译假设与其他翻译假设进行对齐时, 使用的是开源工具包 TER[13], 并将其"shift"功能关闭, 即为 WER 对齐。以上 3 个后验概率在使用前都进行了离散化处理[10]。

由表 2 可以看出, 就 CER 而言, 相比于基线系统, 特征 WPP_Dir, WPP_Win 和 WPP_Lev 分别降低了 2.67%, 4.54%和 3.53% (相对值), 其中 WPP_Win 表现最好。对以上结果进行分析, 可得: 1) WPP_Win 特征因其将固定位置改变为滑动窗, 对齐灵活度更高, 因此更符合源语言和目标语言因词序不同而导致的调序现象, 但滑动窗仅限于有限的局部调序; 2) WPP_Lev 特征是基于 Levenshtein 对齐的, 因此对齐性更好, 但由此也引入过多的编辑操作, 即插入、删除、替换等, 而且也因不存在词序调整, 虽对齐好于 WPP_Dir, 但灵活性低于 WPP_Win。

由以上分析和数据可以知道, 结合 CER 和 *F* 值, 特征 WPP_Win 的综合性能最好。

4.4.2 基于语言学特征的分类实验

基于语言学特征即单词实体(Word)、词性标注(POS)和句法关系(Link)的错误检测结果见表 3。

相比于基线系统, Word, Pos 和 Link 分别在 CER 上降低了 5.96%, 5.03%和 1.68%(相对值), 其中 Word 表现最好。就 *F* 值而言, Link 表现优于其他两个特征, 而 Pos 又优于 Word。对以上结果进行分析, 并与表 2 进行比较, 可得: 1) 除 Link 特征外, 语言学特征中的 Word 和 Pos 的分类错误率都低于 3 个词后验概率特征的分类错误率; 2) Link 特征的召回率最高, 而准确率最低, 这主要是由于 Link 特征的特征个数相对较少, 因此在分类时, 分类结果更倾向于将目标单词标为类别 i, 导致类别 c 的个数相对较少, 从而使召回率高而准确率较低; 3) Word 特征的分类结果好于 Pos 特征的原因可能在于开发集和测试集的相关性较强(皆为新闻领域), 而且特征个数要远远多于 Pos 特征个数, 因此在分类能力上, 将目标单词预测为类别 i 的倾向(或概率)要低于特征个数相对较少的 Pos, 导致其召回率降低, 但准确率更好。

表 3 语言学特征错误检测实验结果

Table 3 Performance of the error detection task based on linguistic features

语言学特征	CER/%	*P*/%	*R*/%	*F*/%
基线	41.59	—	—	—
Word	39.11	64.20	76.67	69.04
Pos	39.50	61.52	86.46	71.89
Link	40.89	59.55	93.55	72.77

4.4.3 组合特征分类实验

在自然语言处理研究的分类任务中, 特征组合往往可以更有效地降低分类错误率。表 4 列出了本文所描述的 3 种典型的词后验概率特征与 3 种语言学特征组合后基于最大熵模型的分类实验结果。

由表 4 可以看出, 就 CER 而言, 与基线系统相比, 3 种组合特征的 CER 分别降低了 13.61%, 14.52%和 14.35%(相对值), 而且 *F* 值也得到显著提高。虽然 3 种特征组合的分类性能差异不显著, 但其分类特性表现出与单个 WPP 特征时的一致性, 即组合"WPP_Win + Word + Pos + Link"的分类错误率最低, 而组合"WPP_Dir + Word + Pos + Link"的 *F* 值最高, 说明基于滑动窗位置的词后验概率特征可以捕获更多上下文信息, 从而使其区分翻译错误的能力强于基于固定位置的词后验概率特征。这种能力不仅表现在单个特征对比时, 同时在组合特征中也得到展现。表 4 在比较 3 种不同 WPP 特征组合效果的同时, 也揭示了语言学特征对于错误检测的贡献, 表明了语言学特征可以有效降低分类错误率, 提高错误预测能力。

表 4 3 种不同的 WPP 特征与语言学特征组合的错误检测实验结果

Table 4 Performance of the error detection task based on combination of three typical WPP and linguistic features

特征组合	CER/%	*P*/%	*R*/%	*F*/%
WPP_Dir + Word + Pos + Link	35.93	63.93	88.30	74.17
WPP_Win + Word + Pos + Link	35.55	64.77	85.83	73.83
WPP_Lev + Word + Pos + Link	35.62	65.31	83.22	73.15

5　结论与展望

本文分析和比较了 3 种典型的用于机器翻译译文错误检测的词后验概率特征的计算方法, 并结合 3 种语言学特征, 分别对其组合特性进行了实验验证和分析。基于汉英 NIST 机器翻译数据集的实验结果表明: 1) 不同方法的单词后验概率特征可以有效降低分类错误率, 但其错误检测能力是不同的, 差异是显著的; 2) 语言学特征可以有效降低分类错误; 3) 就 CER 准则而言, 组合特征的性能要远远优于单个特征的性能, 而且其特性与单个特征的特性具有一致性。

以后的工作中主要有以下问题需要解决: 1) 本文中的分类类别 c 与 i 的标注仅仅使用 WER 准则, 其要求对齐的词必须完全一致才为“正确”, 而实际情况存在同义不同形的情形, 因此, 下一步将采用复述等技术, 改进标注方法, 使得能够识别同义不同形的情况, 以更符合人工判断词语对错的情形, 提高与人工判断相一致的相关性; 2) 相比于其他特征, Link 特征的表现相对较弱, 这主要是因为其无法提供更丰富的上下文信息和依存关系, 下一步将抽取更丰富的句法特征以加强对非语法现象的约束, 从而提高分类性能。

参考文献

[1] Yamada K, Knight K. A syntax-based statistical translation model // Proceedings of ACL-EACL. Toulouse: Morgan Kaufmann, 2001: 523–530

[2] Koehn P, Och F J, Marcu D. Statistical phrase-based translation // Proceedings of HLT-NAACL. Edmonton: Association for Computational Linguistics, 2003: 127–133

[3] Chiang D. A hierarchical phrase-based model for statistical machine translation // Proceedings of ACL. Ann Arbor: Association of Computational Linguistics, 2005: 263–270

[4] Gandrabur S, Foster G. Confidence estimation for translation prediction // Proceedings of HLT-NAACL. Sapporo: Association for Computational Linguistics, 2003: 95–102

[5] Ueffing N, Macherey K, Ney H. Confidence measures for statistical machine translation // Proceedings of MT Summit IX. New Orleans: Springer-Verlag, 2003: 394–401

[6] Blatz J, Fitzgerald E, Foster G, et al. Confidence estimation for machine translation // Proceedings of COLING. Geneva: Yale University Press, 2004: 315–321

[7] Ueffing N, Ney H. Word-Level confidence estimation for machine translation. Computational Linguistics, 2007, 33(1): 9–40

[8] Specia L, Cancedda N, Dymetman M, et al. Estimating the sentence-level quality of machine translation systems // Proceedings of the 13th EAMT. Barcelona: European Association for Machine Translation, 2009: 28–35

[9] Specia L, Saunders C, Turchi M, et al. Improving the confidence of machine translation quality estimates // Proceedings of the 12th MT Summit. Ottawa: International Association for Machine Translation, 2009: 136–143

[10] Xiong Deyi, Zhang Min, Li Haizhou. Error detection for statistical machine translation using linguistic features // Proceedings of the 48th ACL. Uppsala: Association for Computational Linguistics, 2010: 604–611

[11] Bach N, Huang F, AI-Onaizan Y. Goodness: a method for measuring machine translation confidence // Proceedings of the 49th ACL. Portland: Association for Computational Linguistics, 2011: 211–219

[12] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation // Proceedings of ACL. Prague: Association for Computational Linguistics, 2007: 177–180

[13] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation // Proceedings of AMTA. Cambridge: Association for MT in the Americas, 2006: 223–231

[14] Zhang Le. Maximum entropy modeling toolkit for Python and C++ [CP/OL]. (2004) [2012–02–10]. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html