文章编号:1671-4598(2020)04-0161-04 DOI:10.16526/j.cnki.11—4762/tp.2020.04.033 中图分类号:TP391.1 文献标识码:A

基于改进 GLR 算法的智能识别英语翻译模型设计

党莎莎，龚小涛
（西安航空职业技术学院 通识教育学院，西安 710089）

摘要：广义最大似然比检测（GLR）算法模型翻译识别结果存在数据点重合的情况，精确度无法得到有效保障；为了准确地识别短语，设计了基于改进 GLR 算法的短语智能识别算法，该算法构建标记规模约 74 万个英汉单词的短语语料库，使短语具备可搜索功能，通过短语中心点构建短语结构，可获得词性识别结果，依据解析线性表的句法功能校正词性识别结果中的英汉结构歧义，最终获得识别的内容；综合的测评结果上看，基于改进 GLR 算法识别精度 95%以上，综合得分 92.3 分，该算法克服了 GLR 的弊端，相对统计算法和动态记忆算法提高了运算速度和处理性能，更加适合机器翻译任务，为在智能机器翻译领域提供了新的思路。

关键词：智能识别；改进 GLR 算法；机器翻译

Design of Intelligent Recognition English Translation Model Based on Improved GLR Algorithm

Dang Shasha, Gong Xiaotao
(School of General Education, Xi'an Aviation Vocational and Technical College, Xi'an 710089, China)

Abstract: Generalized maximum likelihood ratio detection (GLR) algorithm model translation recognition results have data points that overlap, and accuracy cannot be effectively guaranteed. In order to accurately identify phrases, an optimized GLR algorithm based on intelligent recognition is designed. This algorithm constructs a corpus of phrases with a scale of about 740, 000 English—Chinese words, makes the phrases searchable, and constructs the phrase structure through the center of the phrase to obtain part—of—speech recognition As a result, the ambiguity between English and Chinese structures in the part—of—speech recognition results was corrected according to the syntactic function of the analytical linear table, and finally the content of recognition was obtained. In terms of comprehensive evaluation results, based on the improved GLR algorithm, the recognition accuracy is more than 95%, and the comprehensive score is 92.3, the algorithm overcomes the disadvantages of GLR, improves the operation speed and processing performance relative to the statistical algorithm and dynamic memory algorithm, and is more suitable for machine translation tasks, providing a new way of thinking in the field of intelligent machine translation.

Keywords: intelligent recognition; optimized GLR algorithm; machine translation

0 引言

近些年，伴随教育、科技的不断发展，机器翻译应用产品的数量也越来越多[1]，这些应用主要集中在学术文献、搜索引擎等外文翻译方面。因此，机器翻译技术有着庞大的市场应用需求，发展前景较好。以往的机器翻译技术或多或少有些弊端，翻译的精准性太低，是阻碍机器翻译技术进一步发展的巨大瓶颈。在实际的机器翻译产品测试环节，比如百度、GOOGLE 翻译软件，翻译结果和实际专业人工翻译的质量相差较大[2]，暴露出现有的机器翻译水平已经无法适应当前翻译需求的状况，市场亟需一款高性能、翻译准确率高的机器翻译技术。得益大数据的发展，许多研究者寻求通过计算机辅助翻译（computer aided translation，CAT）来帮助完成翻译工作。计算机辅助翻译的核心思想是：翻译的结果通常被当成辅助性的参考，最后是由用户来判断翻译的优劣，进行人工选择；另外一方面对语料库的运用，能够把各个行业领域的词汇进行归类整理，让翻译的质量得到改进，更加贴近用户的实际需求[3]。合理使用翻译频数较高的专业词汇语料库能够很大程度地、较少重复、翻译工作量，而且还能极大提高翻译的准确性。

周亚婷[4]分析了英语篇章机器翻译符合单位属于句号局的特性，其单位为 NT 小句，对其翻译单位体系中的 PTA 模型实现了英汉翻译的过程，实现面向篇章翻译英汉小句语料库的建设，对其中的 PTA 模型进行了详细的讲解，彰显了语料库的重要性。卢蓉[5]改进了传统基于规则的机器翻译模型，使用基于语义网络的英语机器翻译模型，在具体的实现过程中，使用基于向量混合的短语合成语义统计英语机器翻译方法，在翻译相似度模型的度量过程中，使用余弦相似度计算方法获取两个向量的语义相似度，加入加权向量法计算规则辨别两个相似向量的不同之处，获

收稿日期:2020-01-16； **修回日期**:2020-02-21。

作者简介:党莎莎(1991-)，女，山西临汾人，硕士，助教，主要从事英语教学与翻译等方向的研究。

取精准翻译的结果，保证翻译的质量。黄登娴[6]克服了采用管道式逐层分析技术对机器翻译进行解析，将切分的短语单词与短语语料库对比分析词性和句法，进一步获得待翻译的英文的句法结构的方法错误具有逐步传递和累积，最终导致翻译准确率较低的弊端，设计基于知网的词汇语义相似度以及对数线性模型，采用汉英依存树到串的方式保存对应的双语语料，提供对语言依存结构化的处理，确保汉英双语的对应关系，计算知网运算输入需要翻译句子同实例库内源语言中词汇的语义相似度，进一步提高了翻译的准确率，翻译结果具备较高的准确性。

经过对以上文献的总结，发现在翻译过程中某个句子的短语包含的语义通常是这个句子中的核心内容，对短语的智能识别是语言识别中重要的环节，其原理就是通过对句子中的短语进行识别汇总，然后分析短语的词性和句法，对照短语语料库进行翻译和自动组合，最终得到原文句子的翻译结果[7]。在机器翻译领域，短语的智能识别是关键技术，可以满足翻译样本的选调、平行语料的精确对齐，采用短语智能识别的技术能够有效减少语法上的歧义。结构歧义是当前英语翻译领域中的难点，需要运用词性识别算法来解决，本文使用基于改进的 GLR[8]（generalized maximum likelihood ratio，广义最大似然比检测，简称 GLR）算法的机器翻译算法，该算法构建标记规模约 74 万个英汉单词的短语语料库，使短语具备可搜索功能，通过短语中心点构建短语结构，可获得词性识别结果，依据解析线性表的句法功能校正词性识别结果中的英汉结构歧义，最终获得识别的内容，确定翻译中短语的实际位置范围，以期一定程度上缓解结构歧义在当前英语翻译领域中的弊端，提高短语识别的效率。

1 短语智能识别算法

1.1 短语语料库构建

语料库在智能英语翻译模型中扮演了重要的角色，将双语短语资料存储在语料库中，能够对汉语、英语中的短词语的词性进行精准的标注，规范每个短语的功能，能够大幅度地提高英汉机器翻译过程中的短语自动识别算法的精确性和时效性[9]，协助英汉机器翻译地更加准确。众所周知，通常的英汉机器翻译都是将长句转换成多对短词语形式，然后匹配语料库中的语料，采用打分算法评估翻译后的上下文环境和相应的翻译短语的优劣，加大标记范围等方法能有效提升得分，这也是一些新兴的算法创新的思路，最终形成机器翻译的结果。所以，构建的短语语料库的整体功效对机器翻译算法有着至关重要的作用。图 1 对短语语料库信息的流程进行了展示。

本文基于智能识别的英语翻译模型构造的短语语料库包含了 74 万个单词，能够满足构造 2.2 万个句子、1.2 万个短语的需求，从图 1 中的短语语料库信息可以看出，短语语料库是具有针对性的，本文选用的是英汉机器翻译的短语语料库，分别对英汉的短语语料进行了标注，区分了不同短语语料的时态；语料的标记方式由数据、层次和加工方式三个部分组成，数据的类型是文本格式，层次选用了词性和对齐的方式，加工方式采用人机主动沟通方式直接互动，进行英文翻译的一系列常规流程操作，促使短语语料翻译的准确性。

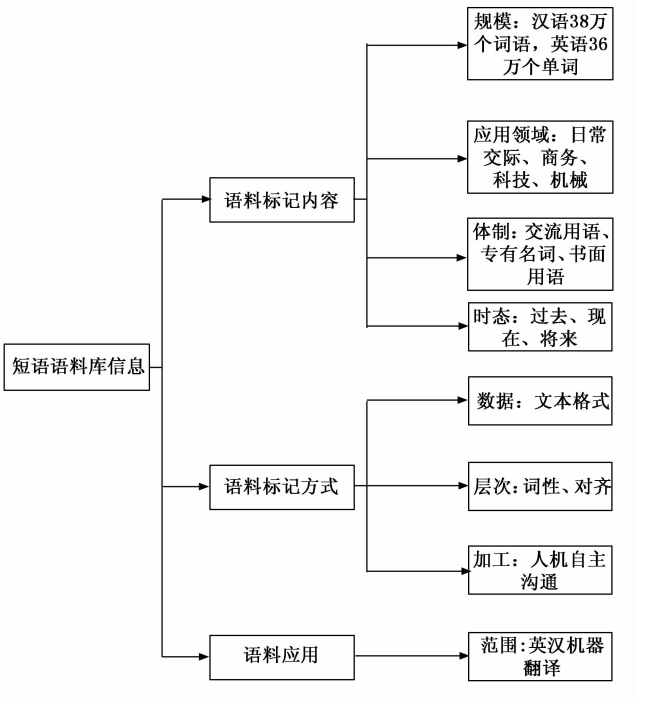

图 1 短语语料库信息流程

1.2 短语语料库词性识别

短语的词性识别是机器翻译智能识别算法中关键的核心步骤，能够对大量的句子、短语、单词的语法歧义进行处理[10]。通过对短语语料库中内容进行词性标注，每个语句都会划分为数个单词，对于英文句子，每个单词都是独立的存在，中文语句需要进行“分词”处理，处理后的单词经过对齐处理后形成了短语，其间通过对翻译句子上下文的判断会标记单词的词性，最后通过句法分析短语的依存关系，形成句子的句法树。通过这种方法使得机器翻译的时效性和准确性提升，另外还使得短语语料库的处理能力得到显著增加。GLR 算法是词性识别当中常用的一种算法，主要用于判断短语前后文关系，其核心理论是基于动态识别表单和无条件转移语句[11]。

经典的 GLR 算法每个步骤的运转都是使用多种移位指令和精简的操作，期间的每个操作的开端和终端都是使用特使的标准来展示。在进行短语翻译的过程中，当 GLR 算法没有检测到语法歧义的状况，就会重新开始进行去重和校准操作；如果检测到语法歧义，就需要使用句法分析的几何结构线性表来对解析线性表进行调取，对短语的内容展开识别，根据局部最优原则提供最优的内容，输送至不同的识别通道中进行符号的识别，根据识别的结果选择最优的结果。

通常情况下，由于 GLR 算法在词性识别的结果中存在较大的偶然性，识别的数据点重合概率较高，仍然无法满足现有的词性识别精确度[12]。本文对经典的 GLR 算法进行了改进，提出使用短语中心来分析短语的结构，有效降低了数据点重合的概率，提升了词性识别的精确度。改进的 GLR 算法对短语前后文的似然性计算借助四元集群来实现，算法如式（1）所示：

$$G_E = (V_N, V_T, S, \alpha) \tag{1}$$

在式（1）中，V_N 代表循环符号集群，$V_N \neq \phi$；V_T 代表终止符号集群，$V_T \neq \phi$ 且 V_T 与 V_N 中的元素不重合；S 代表开始符号集群，是 V_N 中的元素；α 代表短语动作集群。

假设 P 是 α 中的任意动作且 P 又存在于 V_N 中，经过推导可以得到式（2）：

$$P \rightarrow \{\theta, c, x, \delta\} \tag{2}$$

在式（2）中，θ, c, x, δ 分别代表动作右侧符号、中心点符号、约束值和标记方式，θ 和 c 同时位于 V_T 与 V_N 中，δ 可位于 V_T 中，也可位于 V_N 中。

改进的 GLR 算法规定识别结果线性表最上面的符号与 θ 一致，约束值 x 需为真，中心点符号 c 需数值，不能为空值。只有达到了以上 3 个标准的识别结果，才是短语词性识别的结果。

1.3　短语智能识别算法的校正流程

目前现行的英汉机器翻译算法中，对切分的短语与短语语料库匹配得到的结果往往作为最终的机器翻译结果，缺乏对短语所处的上下文环境的分析，过分依赖短语语料库的词性分析，导致最终的翻译结果不够准确[13]。因此本文进一步考虑对词性分析的结果进行校正处理。在对改进的 GLR 算法进行词性分析校正的过程中，针对 GLR 算法使用解析线性表对短语进行词性识别的结果中出现错误点的状况，校正过程通过核对短语语料库中的标记内容进行，详细的短语校正算法流程如图 2 所示。

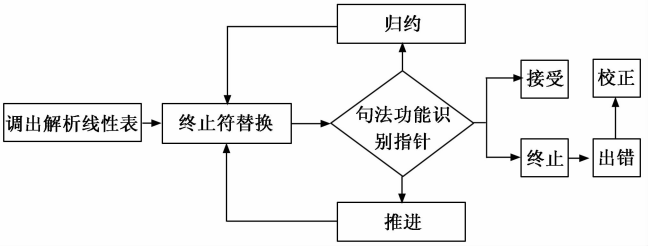

图 2　智能识别算法校正流程图

从图 2 中可以看到归约与推进指标的关系，具体的关系如表 1 所示。

在改进的 GLR 算法运行的过程中，对终止符展开更换前，要先识别指针的类型，如果是规约指针，需要检测指针的约束条件是否存在于短语语料库中；如果不存在，就直接进入终止指针。终止指针一般会出现在有结构歧义的后备点的位置上，当查询到终止指针后，就会形成短语结构树，然后标记符号栈，研究后备点的中心点符号是不是有，是不是放置在准确的语句结构上，如果没有或者放置不正确，那算法就会调用出错指针，进行校正词性的识别结果[14]。

表 1　归约与推进指令的比较表

名称	相同点	不同点
归约	两者作用接近，都是要更换解析线性表中的终止符位置。	归约表达的是之前的约束条件已经没有效果或者是循环过程存在问题，需要再次明确句法功能识别约束条件。
推进		推进表示的是正在进行的句法功能识别中没有结构存在歧义的位置，短语词性识别结果是准确无误的，这时候就要选择接受指针进行留用。接受指针和推进指针通常都会一起出现，如果在流程中没有满足这个条件，仅仅只是出现了某个指针，这就说明循环出现差错或者算法有误。然后就需要再次调出解析线性表，撤回之前已经默认同意的词性识别结果。

2　模型验证

2.1　验证方案

为了验证改进后的 GLR 算法的实际英汉翻译效果，需要进行相关的测评，展示改进的 GLR 算法的性能，测评的英汉翻译任务主要性能指标包括：翻译精度、翻译速度、更新能力。实验的测评小组由专业的英汉翻译人员、3 台英汉翻译机器和专业的评分人员组成，其中三台英汉翻译机器的词性分析阶段算法分别选择的是统计算法、动态记忆算法、GLR 算法、改进的 GLR 算法。

测评的过程：三台英汉机器翻译对指定的 50 条短语和 50 条网络随机语句进行翻译，英汉翻译的专业人员同样对对指定的 50 条短语和 50 条网络随机语句进行翻译，评分人员通过对比机器翻译和人工翻译，然后对三台英汉机器的算法进行进行评分，评分的规则如表 2 所示。

表 2　评分规则表

项目	评分规则
识别精度	根据表达的内容是否清晰、语法结构是否正确，按满分 100 分进行打分。
识别速度	算法的总识别时间乘以权值后进行求和，再除以短语识别数量。
更新能力	算法的总更新时间乘以权值后进行求和，再除以短语识别数量。

注：各项分值权重为识别精度 0.8，识别速度 0.1，更新能力 0.1。

2.2　实验结果

本次测评实验对 50 条短语和 50 条网络随机语句进行短语识别，详细描述见 3.1 小节，详细的实验结果如表 3 所示。

从图 3 的测试结果来看，无论是从识别精度、识别速度、更新能力上，基于改进的 GLR 算法词性识别的机器翻译都是同类最优的。从图 4 综合的测评结果上看，最高得

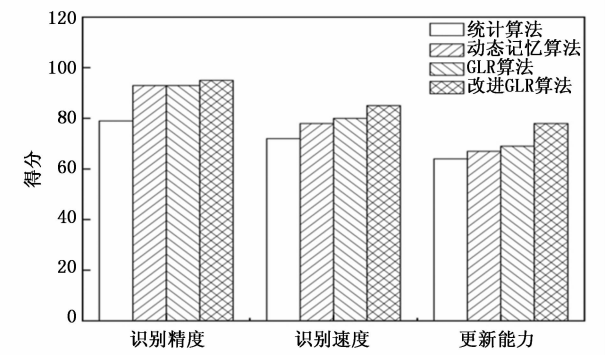

图 3 4 种英汉翻译算法评价结果

分是基于改进 GLR 算法 92.3 分，最低得分是统计算法 76.8 分，动态记忆算法在最后的测试得分上与改进的 GLR 算法得分差异不大，两者的主要差距集中在更新能力方面的得分。结合图 3、图 4，显然，改进 GLR 算法较其他算法的性能优势明显。

本文的比对实验还采用了对实际翻译案例的实验，选择“西安市物价局就牛肉面限价”语句进行翻译，最终得到的基于统计算法、动态记忆算法、改进的 GLR 算法的机器翻译和人工翻译译文的实验比对结果如表 4 所示。

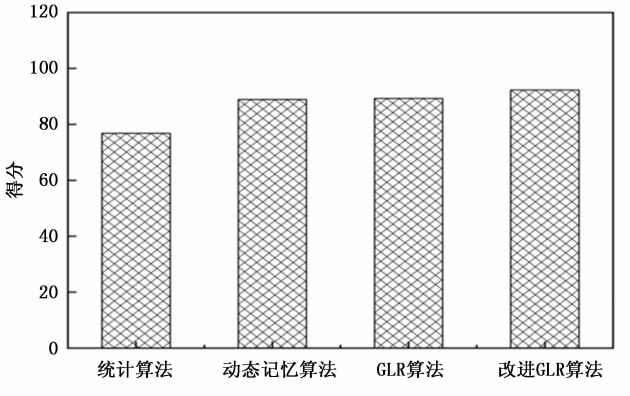

图 4 4 种英汉翻译算法综合测试得分比较

表 4 翻译实例结果对比

翻译方法	翻译内容
统计算法	Xi’an explained beef noodles reduce:only because of the excessive price.
动态记忆算法	Xi’an explained that beef noodles reduce:only because ofthe excessive price increase.
GLR 算法	Xi’an price bureau explained that beef noodles reduce:only because of the excessive price raises.
改进 GLR 算法	Xi’an price bureau gives the explanations of beef noodles reduce:only because of the excessive price raises.
人工翻译译文	Xi’an price bureau gives explanations of price controlon beef noodles: it is only because the raises have been too large.

从表 4 中可以发现，基于统计算法和动态记忆算法的机器翻译对“物价局”这个词没有进行翻译，而基于改进 GLR 算法的机器翻译正确的翻译出来了。在对“做出解释”，进行翻译的时候，只有基于改进 GLR 算法的机器翻译和人工翻译译文最接近，可以明显地看到本文设计的基于改进 GLR 算法的机器翻译对比统计算法和动态记忆算法翻译得更加准确，识别精度可达了 95％以上，达到了与人工翻译同等级别的水平，表明了基于改进 GLR 算法在机器翻译中的高效可行性。

3 结论

针对英语翻译领域中结构歧义的难点，同时克服了传统 GLR 算法在翻译模型中词性识别存在数据点重合的弊端，提出了改进的 GLR 算法。改进 GLR 算法运用短语中心点来设计短语的结构，依据解析线性表的句法功能校正词性识别结果中的英汉结构歧义，从而有效缓解了传统统计算法和动态记忆算法中识别结果精度不高的现状，为识别的短语指定了最合理的位置。实验的结果表明，基于改进 GLR 算法的机器翻译同其他算法相比，具有计算简单快捷、难度不高、实用性更强的特性，适合英语机器翻译工作。

参考文献：

[1] 白瑞芳. 基于 RNN 编码器的交互式机器翻译平台控制技术 [J]. 计算机测量与控制，2019，27 (7)：89－92.
[2] 钟 梅. 基于 CAT 技术的大学英语翻译教学实践 [J]. 英语教师，2019，19 (9)：20－25.
[3] 周亚婷. 英语篇章机器翻译单位及模型设计及应用 [J]. 电子测试，2018 (10)：118－119.
[4] 卢 蓉. 基于语义网络的英语机器翻译模型设计与改进 [J]. 现代电子技术，2018，41 (14)：126－129.
[5] 黄登娴. 英语翻译软件翻译准确性矫正算法设计 [J]. 现代电子技术，2018，41 (14)：170－172，177.
[6] 宋 柔，葛诗利. 面向篇章机器翻译的英汉翻译单位和翻译模型研究 [J]. 中文信息学报，2015，29 (5)：125－135.
[7] 梁国龙，陶 凯，范 展. 声矢量阵自适应波束域广义似然比检测算法 [J]. 电子学报，2015，43 (1)：135－139.
[8] 王柳莎. 基于 K－均值聚类算法的英语教学岗位胜任能力评估系统设计 [J]. 微型电脑应用，2019，35 (7)：128－130.
[9] 蒋亚芳，严 馨，李思远，等. 多重 CCA 算法的柬汉双语词向量构建方法 [J/OL]. 计算机工程与应用：1－10[2020－01－16]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20190820. 1023. 004. html.
[10] 刘永芳，郝晓燕，刘 荣. 中国英语新词语料库构建技术研究 [J/OL]. 计算机工程与应用：1－7 [2020－01－16]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20190823. 1106. 010. html.
[11] 郭 蕾. 基于自然语言处理的英语翻译计算机智能评分系统设计 [J]. 现代电子技术，2019，42 (4)：158－160，165.
[12] 高成吉. 一种英语口语识别算法 [J]. 信息技术，2018，42 (8)：148－151，158.
[13] 尹 陈. 中国大学生英语翻译自动评分方法的研究与设计 [D]. 合肥：中国科学技术大学，2018.
[14] 付宇博. 基于决策树的英语文本难度评估研究 [D]. 武汉：华中师范大学，2018.