



(12) 发明专利申请

(10) 申请公布号 CN 103488627 A

(43) 申请公布日 2014.01.01

(21) 申请号 201310400123.X

(22) 申请日 2013.09.05

(71) 申请人 中国专利信息中心

地址 100088 北京市海淀区蓟门桥西土城路
6号

(72) 发明人 任智军 李进 蒋宏飞 杨婧

(74) 专利代理机构 北京瑞恒信达知识产权代理
事务所(普通合伙) 11382

代理人 苗青盛 王凤华

(51) Int. Cl.

G06F 17/28(2006.01)

G06F 17/27(2006.01)

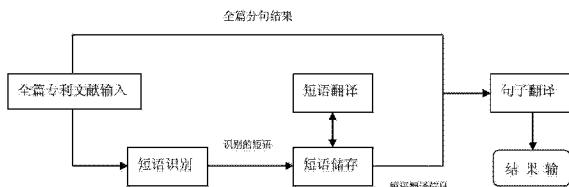
权利要求书4页 说明书15页 附图3页

(54) 发明名称

全篇专利文献翻译方法及翻译系统

(57) 摘要

本发明公开了一种全篇专利文献的机器翻译方法和系统，基于模板或规则方法或权重方法得到短语；然后通过短语频率或修正的短语频率或记忆借鉴等方法进行短语修正，最终得到识别名词短语 RNP；对全文中识别名词短语标注 RNP 信息，翻译识别名词短语 RNP 并在短语存储器中保存相关信息；之后对全文进行逐句翻译，在翻译时，对于标注 RNP 的短语不再展开，直接从短语存储器中取译文；翻译完毕后，根据原文的标题信息进行按顺序输出。本发明能够获取专利文献中常用复杂名词短语，减少含有常用复杂名词短语的句子的分析时间，提高了翻译速度，同时还保证了常用复杂名词短语翻译的一致性。



1. 一种全篇专利文献的机器翻译方法,包括 :

A 步骤 :针对文献全文,识别出各级标题信息并标注 ;

B 步骤 :对全文进行词法分析,得到分词和词性标注信息 ;

C 步骤 :根据 B 步骤的分词和词性标注信息进行短语识别,得到识别名词短语 RNP 并将所述识别名词短语 RNP 翻译成目标语言 ;和

D 步骤 :以句子为单位进行翻译,对于标注为 RNP 的短语直接使用 C 步骤所得的译文,翻译完毕后,按原文标题顺序输出。

2. 根据权利要求 1 所述的方法,其中,所述 C 步骤包括 :

C01 步骤 :采用模板提取法、规则提取法、权重计算法或所述三种方法任意结合对短语进行提取 ;

C02 步骤 :对提取的短语进行判定,得到候选短语 ;

C03 步骤 :对候选短语进行错误识别和修正,得到识别名词短语 RNP ;

C04 步骤 :为全文中出现的所有识别名词短语标注 RNP 标签 ;和

C05 步骤 :翻译最终识别名词短语并存放在短语存储器中。

3. 根据权利要求 2 所述的方法,其中,所述 C01 步骤中权重计算法的步骤包括 :

C0101 步骤 :对短语进行打分,方法可以为 TF-IDF 法、TFC 法或 ITC 法 ;

C0102 步骤 :根据标题信息设置位置权重系数,短语的权重等于短语打分乘以位置权重系数 ;

C0103 步骤 :判断短语是否存在于专利文档库的停用高频短语列表中,若存在,则排除该短语 ;停用高频短语列表的产生方法为 :在专利文档库中,短语频率为该短语在文档库中出现的次数与文档库中所有短语出现的总次数的比值,降序排列后前 N 个短语组成高频短语列表,N 为 20-1000 的整数 ;和

C0104 步骤 :当短语的权重高于设定值时,则判定其为候选短语,设定值为 $0.5 \times \omega^*$, ω^* 为当前专利文档中短语权重的最大值。

4. 根据权利要求 3 所述的方法,其中,所述的位置权重系数包括 :

β_1 ,表示说明书摘要、背景技术、具体实施方式部分的权重 ;

β_2 ,表示权利要求、技术领域部分的权重 ;

β_3 ,表示附图说明部分的权重 ;和

β_4 ,表示标题、权利要求主题名称部分的权重 ;

取值满足以下不等式 :

$\beta_1 < \beta_2 < \beta_3 < \beta_4$ 。

5. 根据权利要求 4 所述的方法,其中, β_1 、 β_2 、 β_3 和 β_4 的取值为 :

$0.1 < \beta_1 < 0.6$

$0.2 < \beta_2 < 0.8$

$0.3 < \beta_3 < 0.9$

$0.5 < \beta_4 < 1$ 。

6. 根据权利要求 4 所述的方法,其中, β_1 、 β_2 、 β_3 和 β_4 的取值为 :

$\beta_1 = 0.4$

$\beta_2 = 0.5$

$\beta_3=0.6$

$\beta_4=0.8$ 。

7. 根据权利要求 2-6 中任一项权利要求所述的方法,其中,所述 C02 步骤中判定方法为短语频率法,首先设定阈值,如果短语频率高于该阈值,并且短语不在专利文档库的停用高频短语列表中,则判定所述短语为候选短语,短语频率为该短语在全文中出现的次数与所有短语出现次数的比值;阈值 ϵ 范围为 [1 / 全篇专利文献中短语的总个数, 100 / 全篇专利文献中短语的总个数]。

8. 根据权利要求 2-6 中任一项权利要求所述的方法,其中,所述 C02 步骤中判定方法为修正的短语频率法,首先设定阈值,如果短语频率高于该阈值,并且短语不在专利文档库的停用高频短语列表中,则判定所述短语为候选短语,短语频率为该短语在全文中出现的次数与所有短语出现次数的比值与位置权重系数的乘积;阈值 ϵ 范围为 [1 / 全篇专利文献中短语的总个数, 100 / 全篇专利文献中短语的总个数]。

9. 根据权利要求 8 所述的方法,其中,所述 C02 步骤中的位置权重系数包括:

β_1 ,表示说明书摘要、背景技术、具体实施方式部分的权重;

β_2 ,表示权利要求、技术领域部分的权重;

β_3 ,表示附图说明部分的权重;和

β_4 ,表示标题、权利要求主题名称部分的权重;

并且取值满足以下不等式:

$\beta_1 < \beta_2 < \beta_3 < \beta_4$ 。

10. 根据权利要求 9 所述的方法,其中, β_1 、 β_2 、 β_3 和 β_4 的取值为:

$0.1 < \beta_1 < 0.6$

$0.2 < \beta_2 < 0.8$

$0.3 < \beta_3 < 0.9$

$0.5 < \beta_4 < 1$ 。

11. 根据权利要求 9 所述的方法,其中, β_1 、 β_2 、 β_3 和 β_4 的取值为:

$\beta_1=0.4$

$\beta_2=0.5$

$\beta_3=0.6$

$\beta_4=0.8$ 。

12. 根据权利要求 2-6 中任一项权利要求所述的方法,其中,所述 C02 步骤采用记忆鉴定法进行判定,对专利文档库中所有专利全文提取短语,经过人工判定得到正确的短语,并将其保存在记忆库,将记忆库中的短语和待判定短语通过编辑距离算法和最长公共字串法进行比较,生成候选短语。

13. 根据权利要求 2-6 中任一项权利要求所述的方法,其中,所述 C02 步骤采用短语频率法、修正的短语频率法、记忆鉴定法的任意组合进行判定,对不同判定方法的结果使用投票法进行选择,相同结果数量最多的短语为候选短语。

14. 根据权利要求 2 所述的方法,其中,所述 C03 步骤采用 CRF 方法、规则方法、错误模式方法或此三种方法任意结合进行辨识和修正,得到识别名词短语 RNP,同时修正短语标注信息。

15. 根据权利要求 2 所示的方法,其中,所述 C05 步骤包括:

判断短语是否已在短语存储器中,若不在,进行短语翻译;翻译后,按短语存储器格式保存该短语,该短语存储器格式包括短语、分词信息、词性标注信息、识别名词短语标签信息和译文信息。

16. 根据权利要求 15 所示的方法,其中,所述短语翻译包括以下步骤:

核心词修正,对短语进行句法分析,将短语的根节点修改为核心词 / 主题词;然后采用 CYK 算法进行翻译;

通过计算平均调序距离,保留得分高的至少一个候选译文;和

根据目标语言专利文档库信息进行译文候选重排序,将多个翻译候选结果通过利用目标语言专利文档库训练获得的语言模型进行语言模型评分,输出评分最高者。

17. 一种全篇专利文献的机器翻译系统,包括:

输入模块,用于接收并分析文献全文,首先识别各级标题,然后进行词法分析,标注分词、词性信息;

短语识别模块,所述短语识别模块用于得到识别名词短语 RNP;

短语翻译模块,所述短语翻译模块翻译识别名词短语,并保存在短语存储器中;

全文翻译模块,所述全文翻译模块对全文逐句翻译,对于识别名词短语 RNP 不再进行句法展开,直接从短语存储器中取译文;和

输出模块,所述输出模块将翻译结果按原标题顺序输出。

18. 根据权利要求 17 所述的系统,其中,所述短语识别模块还包括:

短语提取模块,所述短语提取模块根据模板法、规则法、计算权重法或其结合提取短语;

短语判定模块,所述短语判定模块根据短语频率方法、修正的短语频率法、记忆鉴定方法、投票法或其结合进行短语判定;和

错误修正模块,所述错误修正模块采用 CRF 方法、规则方法或错误模式方法或其结合对候选短语进行修正,最终得到识别名词短语 RNP。

19. 根据权利要求 17 所述的系统,其中,所述短语存储器包含短语、分词信息、词性标注信息、识别名词短语标签信息和译文信息。

20. 根据权利要求 17 所述的系统,其中,所述短语翻译模块包括:

判断单元,用于判断识别名词短语 RNP 是否存在于短语存储器中,如果存在,则不作处理转到下一条短语;如果不存在,进入修正单元;

修正单元,用于对识别名词短语 RNP 进行句法分析,并将所述识别名词短语结构修正为以核心词 / 主题词作为根节点的结构;

翻译及评分单元,对修正后的名词短语,采用 CYK 算法自底向上进行翻译,并结合平均调序距离进行评分;和

对比单元,用于根据目标语言专利文档集信息进行译文候选重排序,即将多个翻译候选结果通过利用目标语言专利文档库训练获得的语言模型进行语言模型评分,保存评分最高者。

21. 根据权利要求 17 所述的系统,其中,所述全文翻译模块包括:

句法分析单元,用于逐句分析句法,获取全文分析处理的分词、词性标注信息;和

翻译单元,对于识别名词短语 RNP 从短语存储器中取出译文,对于其他内容进行翻译。

全篇专利文献翻译方法及翻译系统

技术领域

[0001] 本发明涉及机器翻译技术,尤其涉及全篇专利文献的机器翻译方法及翻译系统。

背景技术

[0002] 机器翻译是使用计算机实现从一种自然语言文本到另一种自然语言文本的翻译。其研究方法分为规则和统计两种。由于规则系统开发周期长,资金和人力的需求大,所以规则系统进展缓慢。相对而言,统计方法开发周期短、便于处理大规模语料等优点而显出优势。在统计机器翻译方法中,基于短语的翻译方法得到充分的发展。但从目前看,对于专业的领域的翻译来说,比如在专利文档的翻译中,较长的短语常常被分词为几个短语进行翻译。例如,“所述超低温热封聚丙烯流延膜,...”,可能会被分词为“所述”、“超低温”、“热”、“封”、“聚丙烯”和“流延膜”。而在专利文献撰写中,“所述”后的词语通常是固定的,其本身就可以看为一个固定短语,所以能将“超低温热封聚丙烯流延膜”作为一个短语整体进行处理,则只需要一次分析和翻译,就可以在此专利文献中出现该短语时直接套用。另外,对于复杂短语,在句法分析的时候,会由于上下语境的不同而产生不同的短语分词结果,造成同一篇专利文档中译文前后不一致,但对于专利文献来说,很多复杂短语是固定的,在全文中会多次出现,因此只要在全文范围内识别出这样的短语,就可以在全文翻译中直接套用其译文,而不必再对同样的内容进行分析。

[0003] 公开号为 CN103116578A 的中国专利申请,公开一种融合句法树和统计机器翻译技术的机器翻译方法与装置,该方法首先建立不同语种语言之间的词典库、语法规则库、短语翻译概率表以及目标语语言模型,然后对原文输入句子进行切分、词性消兼和语法分析,生成句法树,然后采用自顶向下的策略遍历该句法树,对单个节点和部分跨句法的连续节点,取其叶节点的原文与统计机器翻译所训练出的短语翻译概率表进行智能匹配,利用短语翻译表的译文和目标语言的语言模型来达到提高输出译文流利度和准确度的目的。此方法对短语的提取不是基于全文的,因此会存在同样的短语翻译不一致以及多次分析、翻译的情况。

[0004] 因此,在现有技术的翻译过程中,复杂名词短语不能保持一致性,同时,同一短语被多次地分析、翻译,耗时费力。

发明内容

[0005] 为了克服现有的缺陷,本发明提出一种全篇专利文献的机器翻译方法和系统。

[0006] 根据本发明的一个方面,提出了一种全篇专利文献的机器翻译方法,该方法包括以下步骤:A 步骤:针对文献全文,识别出各级标题信息并标注;B 步骤:对全文进行词法分析,得到分词和词性标注信息;C 步骤:根据 B 步骤的分词和词性标注信息进行短语识别,得到识别名词短语 RNP 并将该识别名词短语 RNP 翻译成目标语言;和 D 步骤:以句子为单位进行翻译,对于标注为 RNP 的短语直接使用步骤 C 所得的译文,翻译完毕后,按原文标题顺序输出。

- [0007] 根据本发明的另一个方面,提供了一种机器翻译系统,包括:
- [0008] 输入模块,用于接收并分析文献全文,首先识别各级标题,然后进行词法分析,标注分词、词性信息;
- [0009] 短语识别模块,所述短语识别模块用于得到识别名词短语 RNP 短语翻译模块,所述短语翻译模块翻译识别名词短语,并保存在短语存储器中;
- [0010] 全文翻译模块,所述全文翻译模块对全文逐句翻译,对于识别名词短语 RNP 不再进行句法展开,直接从短语存储器中取译文;和
- [0011] 输出模块,所述输出模块将翻译结果按原标题顺序输出。
- [0012] 本发明提供一种全篇专利全文机器翻译方法和翻译系统,解决了现有技术中常用复杂名词短语翻译不一致及翻译效率低的问题。

附图说明

[0013] 本发明的上述及其它方面和特征将从以下结合附图对实施例的说明清楚呈现,在附图中:

- [0014] 图 1 是全篇专利文献机器翻译方法流程图;
- [0015] 图 2 是短语处理模块工作流程图;
- [0016] 图 3 是短语翻译器句法分析的一个例子;
- [0017] 图 4 是全篇专利文献机器翻译系统的结构图;
- [0018] 图 5 是短语识别模块的工作流程图;和
- [0019] 图 6 是短语翻译模块的工作流程图。

具体实施方式

[0020] 下面结合附图和具体实施例对本发明提供的一种全篇专利文献机器翻译方法和系统进行详细描述。

[0021] 如图 1 所示,图 1 提供了专利文献机器翻译方法总体技术方案实现流程图。该方法包括以下步骤:A 步骤:接收全文,识别各级标题信息、XML 标签信息、特征内容并标注;B 步骤:对全文进行词法分析,得到分词和词性标注信息;其中,根据需要还可以进行浅层句法分析或完整的句法分析;C 步骤:根据 B 步骤的分词结果对短语进行提取、判定、识别和修正,得到识别名词短语 RNP;翻译识别名词短语 RNP 并存放在短语存储器中;D 步骤:以句子为单位进行翻译,翻译时遇到标注为 RNP 的短语,直接从短语存储器中取译文,不再对短语进行分析,翻译完后按原文标题顺序输出译文。

[0022] 在步骤 A 中,专利内容部分包括名称、摘要、权利要求书、说明书(技术领域、背景技术、发明内容、附图说明、具体实施方式);标注的方法举例如下:权利要求 1 可以标注为 <claim1>。

[0023] 在步骤 C 中,包括以下步骤:C01 步骤:短语提取;C02 步骤:短语判定;C03 步骤:短语识别和修正;C04 步骤:为全文中出现的所有该短语标注 RNP 标签;和 C05 步骤:短语翻译。

[0024] 在步骤 C01 中,短语提取可以使用模板提取方法,即通过一些设定的边界信息,利用模板进行短语提取。

[0025] 【例 1】一种用于控制飞机飞行的系统，其特征在于，...

[0026] 可以将“一种”、“其特征在于”作为起始边界信息，利用模板：{一种}+{短语 A}+{，其特征在于}，提取短语“用于控制飞机飞行的系统”。

[0027] 短语提取方法还可以为规则提取方法，即利用词性标注特征 POS(part-of-speech) 加前后缀组合方法进行短语提取，撰写的规则例子如下：(-1) CAT(V)+(0)CAT[N]+(1)Suffix → NP[0,1]。

[0028] 【例 2】... 提供词性标注方法

[0029] 其中，后缀为“方法”，词性标注特征为：提供 / v 词性 / n / 标注 / nv 方法 / n。

[0030] 将后缀“方法”与“词性 / n / 标注 / nv”结合，得到短语“词性标注方法”。

[0031] 短语提取方法可以为计算权重法，对其权重进行打分，如果其权重高于设定值，比如 $0.5 \times \omega^*$ ，则判定为候选短语， ω^* 为当前专利文档中短语权重的最大值。此外，在计算 ω^* 时，要排除在停用高频短语列表中的短语。

[0032] 权重打分方法可以为 TF-IDF 法：

$$[0033] \omega_{NP} = f_{NP} \times \log \frac{N}{n_{NP}}$$

[0034] 其中 ω_{NP} 为短语的权重， f_{NP} 为短语在全文中的频率（其计算公式根据上文中公式）， n_{NP} 为在专利文档库中出现的该短语的文档数，N 为专利文档库中文档数。

[0035] 打分方法还可以为 TFC 法：

$$[0036] \omega_{NP} = \frac{f_{NP} \times \log \left(\frac{N}{n_{NP}} \right)}{\sqrt{\sum_{NP} \left[f_{NP} \times \log \left(\frac{N}{n_{NP}} \right) \right]^2}}$$

[0037] 其中， ω_{NP} 为短语的权重， f_{NP} 为短语在全文中的频率（其计算公式根据上文中公式）， n_{NP} 为在专利文档库中出现该短语的文献数，N 为专利文档库中文档数。 \sum_{NP} 表示对全文中所有短语求和。

[0038] 打分方法还可以为 ITC 法：

$$[0039] \omega_{NP} = \frac{\log(f_{NP} + 1.0) \times \log \left(\frac{N}{n_{NP}} \right)}{\sqrt{\sum_{PN} \left[\log(f_{NP} + 1.0) \times \log \left(\frac{N}{n_{NP}} \right) \right]^2}}$$

[0040] 其中， ω_{NP} 为短语的权重， f_{NP} 为短语在全文中的频率（其计算公式根据上文中公式）， n_{NP} 为在专利文档库中出现该短语的文档数，N 为专利文档库中文档数， \sum_{NP} 表示对全文中所有短语求和。

[0041] 权重打分方法还可以为 TF-IWF 法：

$$[0042] \omega_{NP} = f_{NP} \times \log \left(\frac{\sum_{NP} C_{NP}}{C_{NP}} \right)$$

[0043] ω_{NP} 为短语的权重， f_{NP} 为短语在全文中的频率（其计算公式根据上文中公式）， C_{NP} 为短语在全文中出现的次数， \sum_{NP} 表示对全文中所有短语求和。

[0044] 在计算出权重之后，根据短语出现的位置设置位置权重系数 β_i ，对权重进行调

整,公式如下:

[0045] 【公式 1】 $\omega^* = \omega * \beta_i$

[0046] 其中 β_i 为位置权重系数。 β_i 根据其在分析处理阶段 (A 步骤) 中识别出的各标题部分的位置信息,取不同的值,具体如下:

[0047] β_1 表示说明书摘要、背景技术、具体实施方式部分的权重;

[0048] β_2 表示权利要求、技术领域部分的权重;

[0049] β_3 表示附图说明部分的权重;

[0050] β_4 表示标题、权利要求主题名称部分的权重。

[0051] β_i 取值范围的关系满足不等式 1:

[0052] $\beta_1 < \beta_2 < \beta_3 < \beta_4$

[0053] β_i 优选为:

[0054] $0.1 < \beta_1 < 0.6$

[0055] $0.2 < \beta_2 < 0.8$

[0056] $0.3 < \beta_3 < 0.9$

[0057] $0.5 < \beta_4 < 1$

[0058] 且满足不等式 1 所限定的取值范围。

[0059] β_i 更加优选为:

[0060] $\beta_1 = 0.4$

[0061] $\beta_2 = 0.5$

[0062] $\beta_3 = 0.6$

[0063] $\beta_4 = 0.8$

[0064] 停用高频短语列表是通过计算短语频 f_j^L ,降序排列后取排名 1 至排名 n 的短语而构成,计算短语频率的公式为:

[0065] 【公式 2】
$$f_{NPL} = \frac{C_{NPL}}{C_L}$$

[0066] 其中 f_{NPL} 表示该短语在专利文档库 L 中的频率, C_{NPL} 为该短语在专利文档库中出现的次数, C_L 表示专利文档库中所有短语出现的总次数,计算公式为:

[0067] 【公式 3】

[0068]
$$C_L = \sum_i N_i^L$$

[0069] N_i^L 表示专利文档库中短语 i 出现的次数。排名 n 为 20–1000,优选为 50–500,更优选为 100。

[0070] 该专利文档库可以是大于或等于一万篇的专利文档库,优选与所述被翻译的专利文档技术领域相同或相似的专利文档库。

[0071] 进一步地,在步骤 C01 中可以使用上述三种方式的任意组合来进行短语提取。

[0072] 在步骤 C02 中,短语判定方法可以为短语频率方法,即计算专利全文中该短语出现的频率,按照设定的选择阈值 ϵ ,如果出现频率小于该阈值,则该短语不属于候选短语。

[0073] 短语频率的计算公式为:

[0074] 【公式 4】 $f_{NP} = \frac{C_{NP}}{C}$

[0075] 其中, f_{NP} 为该短语的频率, C_{NP} 为该短语在专利全文中出现的次数, C 为专利全文中所有短语出现的总次数。 C 的计算公式为 :

[0076] 【公式 5】

[0077] $C = \sum_i N_i$

[0078] 其中, N_i 为短语 i 在专利全文中出现的次数。

[0079] 阈值 ε 的计算公式为 :

[0080] 【公式 6】 $\frac{1}{N_{ALL}} \leq \varepsilon \leq \frac{100}{N_{ALL}}$

[0081] 更优选为 :

[0082] 【公式 7】 $\frac{1}{N_{ALL}} \leq \varepsilon \leq \frac{20}{N_{ALL}}$

[0083] 最优选为 :

[0084] 【公式 8】 $\varepsilon = \frac{5}{N_{ALL}}$

[0085] 其中, N_{ALL} 为全篇专利文献中短语的总个数。

[0086] 同时, 查询该短语是否存在于停用高频短语列表中, 若存在, 则该短语不属于候选短语。

[0087] 短语判定方法还可以是修正的短语频率法, 计算方法为 :

[0088] 【公式 9】 $f_{NP}' = f_{NP} * \beta_i$

[0089] 其中 β_i 为位置权重系数, 具体的取值在前面已有描述。

[0090] 短语判定方法还可以为记忆鉴定方法, 首先从一个专利文档库的所有专利全文中提取短语, 经过人工判定等方式得到正确的短语, 存入记忆库。判定时, 使用边际编辑距离算法和最长公共字串法对提取的短语与记忆库中的短语进行比较, 生成候选短语。

[0091] 进一步地, 短语判定方法还可以是上述 3 种方法的任意组合。对于多种判定方法, 可以通过投票法对结果进行选择。所述投票法表示用多种方法获得的短语中, 取相同结果数量最多的一种。例如, 有两种方法得出结果为 A, 有一种方法得出结果为 B, 则取 A 为最终结果, 即候选短语。

[0092] 经过短语判定得到的短语为候选短语。

[0093] 在步骤 C03 中, 对候选短语进行识别和修正以得到识别名词短语 RNP。所述错误修正方法, 可以用 CRF 方法对短语标注结果进行概率打分, 根据打分结果对于错误进行修正。打分公式为 :

[0094] $p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left(\sum_j \lambda_j F_j(y, x) \right)$

$$[0095] F(y, x) = \sum_i^n f(y_{i-1}, y_i, x, i)$$

[0096] 其中, $f(y_{i-1}, y_i, x, i)$ 为转移概率或发射概率, y_{i-1} , y_i 是第 $i-1$ 和第 i 个标记, x 为观察序列。 i 为短语在观察序列中的位置。 $Z(x)$ 是归一化因子。 λ_j 是训练获取的参数。

[0097] 所述错误修正方法可以为规则方法, 根据上下文和相应的语法规则, 对错误进行修正。

[0098] 所述错误修正方法可以为错误模式方法, 对预先获得的所有错误模式进行记录, 放入存储器, 当判定后的短语符合错误模式时, 根据错误模式进行修正。下面举例说明:

[0099] 【例 3】[其中气体发生器] 由两个部分构成 => 其中 [气体发生器] 由两个部分构成。

[0100] 上例中, 左边为原短语边界, 右边为修正后的短语边界, 左边原短语边界标注时, 错误地将“其中”合并到名词短语中, 发现这种错误模式后, 根据错误模式进行修正, 将“其中”排除在名词短语之外。

[0101] 所述错误的修正方法, 还可以是结合上述 2 种或 2 种以上方法, 综合进行错误修正。其中, 错误修正包括修改短语标注信息。

[0102] 经过错误修正步骤后获得的短语为识别名词短语 RNP。

[0103] 在步骤 C05 中, 判断识别名词短语 RNP 是否存在于短语存储器中。如果存在, 则不作处理, 直接对下一条短语进行判断, 否则, 执行下面步骤。

[0104] 首先, 对输入短语进行句法分析并进行核心词修正。目的是将句法分析默认的以动词为根节点的结构修正为以核心词 / 主题词作为根节点的结构。

[0105] 【例 4】词性 / n / 标注 / nv 方法 / n

[0106] 其修正后句法分析结果如图 3 所示。

[0107] 其次, 基于修正后的句法结构, 采用 CYK(Cocke–Younger–Kasami) 算法, 自底向上进行翻译。在此过程中, 结合平均调序距离进行翻译评分。

[0108] 再次, 对 CYK 翻译过程获得的翻译结果, 保留翻译评分最高的 N 个为候选译文, N 优选为 100, 然后再根据目标语言专利文档集训练获得的语言模型评分进行重排序, 确定最优译文。

[0109] 所述平均调序距离公式为:

[0110] 【公式 10】

$$[0111] \sum D = \sum_i L_i / Z$$

[0112] 其中 ω_i 表示第 i 个词调序前后所处位置的距离 $L_i = L_i^2 - L_i^1$ 。Z 为词总数。

[0113] 【例 5】执行 [0] 命令 [1] 超时 [2] => Command[0] execution[1] timeout[2]

[0114] 执行 [0] => execution[1] D1=1

[0115] 命令 [1] => Command[0] D2=1

[0116] 超时 [2] => timeout[2] D3=0

$$[0117] \text{因此 } D = \frac{(1+1+0)}{3} \approx 0.667$$

[0118] 作为调序结果选择的一项评分, D 与预先设定的调序距离阈值 D_f 进行比较, 排除评

分大于 D_f 的译文。所述 D_f 为经验值, 优选 $0.5 \leq D_f \leq 3$, 更加优选为 $1 \leq D_f \leq 2$, 最优选为 $D_f=1.5$ 。

[0119] 所述根据目标语言专利文档集信息进行候选译文重排序, 是将多个翻译候选结果通过利用目标语言专利文档库训练获得的语言模型进行语言模型评分, 输出评分最高者。所述专利文档库是一个专利全文数据库, 其所含专利文档数量优选为一万篇以上。优选为根据待翻译的所述专利文档相同或相似的技术领域的专利文档库。

[0120] 最后, 将识别名词短语 RNP 按短语存储器格式保存在短语存储器中, 供后续翻译使用。信息存放的数据格式为 : 短语、分词信息、词性标注信息、识别名词短语标签信息、译文信息。

[0121] 在步骤 C 中, 可以组合使用各分步骤中的方法。

[0122] 在步骤 D 中, 逐句翻译, 对于标注为 RNP 的短语, 作为名词 NN 处理, 不再对其进行句法树展开。

[0123] 【例 6】本发明提供一种全篇专利文献机器翻译方法及系统, 其句法分析结果如图 2 所示。在译词选择阶段, 对于标注为 RNP 的短语, 从短语存储器中取出其译文作为短语译文。当句子中不含 RNP 标签时, 根据句法分析结果进行翻译。将翻译后的目标语言翻译结果按原文标题顺序输出。

[0124] 根据本发明的另一个方面, 提出一种全篇专利文献翻译系统, 图 4 是全篇专利文献翻译系统的结构图。所述全篇专利文献翻译系统包括 : 输入模块, 接收输入的专利全文, 并对专利全文进行标题标识和标注, 进行词法分析 ; 短语识别模块, 根据词法分析结果对短语进行识别, 得到识别名词短语 RNP, 具体包括短语提取模块、短语判定模块、错误修正模块 ; 短语翻译模块, 包括判断单元、修正单元、翻译及评分单元、对比单元, 对识别名词短语 RNP 进行翻译并在短语存储器中保存相关信息 ; 专利全文翻译模块, 是以句子为翻译单位的机器翻译模块或翻译器, 对专利全文逐句进行翻译, 在翻译过程中, 如果遇到 RNP 短语, 则不对其展开, 直接取短语存储器中的译文 ; 和输出模块, 从专利全文语句翻译模块获取所有句子翻译结果, 按照原文标题顺序输出译文。

[0125] 输入模块首先识别各个专利内容部分, 包括名称、摘要、权利要求书、说明书 (技术领域、背景技术、发明或实用新型内容、附图说明、具体实施方式)。识别方法主要是以专利各部分的标题信息、XML 标签信息、特征内容信息进行识别, 并在识别后进行相应标注。例如权利要求 1 可以标注为 <claim1>。

[0126] 然后, 在进一步确定段落单元及语句单元后, 利用现有开源词法分析工具和句法分析工具对每条语句进行词法分析, 也可以根据需要进行适度的句法分析, 并给出语句的分词结果、词性标注结果以及句法分析结果。

[0127] 短语识别模块, 包括短语提取模块、短语判定模块、错误修正模块, 图 5 是短语识别模块的工作流程图。

[0128] 短语提取模块用于提取短语, 方法可以为模板提取方法, 根据设定的边界信息, 利用模板进行短语提取。例如, 一种用于控制飞机飞行的系统, 其特征在于, ...。可以将“一种”、“其特征在于”作为起始边界信息, 利用模板 : { 一种 }+{ 短语 A }+{, 其特征在于 }, 提取短语“用于控制飞机飞行的系统”。

[0129] 提取方法还可以为规则提取方法, 利用词性标注特征 POS(part-of-speech) 加前

后缀组合方法,规则的一个例子为:

[0130] (-1) CAT(V)+(0) CAT[N]+(1) Suffix → NP[0,1]。

[0131] 【例 7】... 提供词性标注方法,其中,后缀为“方法”,词性标注特征为:提供 / v 词性 / n / 标注 / nv 方法 / n。将后缀“方法”与“词性 / n / 标注 / nv”结合,得到短语“词性标注方法”。

[0132] 提取方法还可以为计算权重法,对其进行打分计算权重。如果高于设定值,比如 $0.5 \times \omega^*$,则判定其为候选短语。 ω^* 为去掉停用高频列表中的短语后全文剩余短语的权重的最大值。

[0133] 所述停用高频短语列表是通过计算短语频率 f_i^L ,降序排列后取排名 1 至排名 n 的短语而构成,计算短语频率的公式为:

[0134] 【公式 11】

$$[0135] f_{NPL} = \frac{C_{NPL}}{C_L}$$

[0136] 其中 f_{NPL} 表示该短语在专利文档库 L 中的频率, C_{NPL} 为该短语在专利文档库中出现的次数, C_L 表示专利文档库中所有短语出现的总次数,计算公式为:

[0137] 【公式 12】

$$[0138] C_L = \sum_i N_i^L$$

[0139] N_i^L 表示专利文档库中短语 i 出现的次数。排名 n 为 20~1000,优选为 50~500,更优选为 100。

[0140] 该专利文档库中专利文献的数量大于或等于一万篇,优选与所述被翻译的专利文档技术领域相同或相似的专利文档库。

[0141] 权重打分方法可以为 TF-IDF 法,

$$[0142] \omega_{NP} = f_{NP} \times \log \frac{N}{n_{NP}}$$

[0143] 其中 ω_{NP} 为短语的权重, f_{NP} 为短语在全篇专利文献中的频率(其计算公式根据上文中公式), n_{NP} 为在专利文档库中出现的该短语的专利文档数, N 为专利文档库中文档数。

[0144] 打分方法还可以为 TFC 法:

$$[0145] \omega_{NP} = \frac{f_{NP} \times \log \left(\frac{N}{n_{NP}} \right)}{\sqrt{\sum_{NP} \left[f_{NP} \times \log \left(\frac{N}{n_{NP}} \right) \right]^2}}$$

[0146] 其中, ω_{NP} 为短语的权重, f_{NP} 为短语在全篇专利文献中的频率(其计算公式根据上文中公式), n_{NP} 为在专利文档库中出现的该短语的专利文献数, N 为专利文档库中文档数, \sum_{NP} 表示对全篇专利文献中所有短语求和。

[0147] 打分方法还可以为 ITC 法:

$$[0148] \quad \omega_{NP} = \frac{\log(f_{NP} + 1.0) \times \log\left(\frac{N}{n_{NP}}\right)}{\sqrt{\sum_{PN} \left[\log(f_{NP} + 1.0) \times \log\left(\frac{N}{n_{NP}}\right)\right]^2}}$$

[0149] 其中, ω_{NP} 为短语的权重, f_{NP} 为短语在全篇专利文献中的频率 (其计算公式根据上文中公式), n_{NP} 为在专利文档库中出现的该短语的专利文献数, N 为专利文档库中文档数, \sum_{NP} 表示对全篇专利文献中所有短语求和。

[0150] 打分方法还可以为 TF-IDF 法 :

$$[0151] \quad \omega_{NP} = f_{NP} \times \log\left(\frac{\sum_{NP} C_{NP}}{C_{NP}}\right)$$

[0152] ω_{NP} 为短语的权重, f_{NP} 为短语在全篇专利文献中的频率 (其计算公式根据上文中公式), C_{NP} 为短语在全篇专利文献中出现的次数, \sum_{NP} 表示对全篇专利文献中所有短语求和。

[0153] 在计算出权重之后, 根据短语出现的位置, 对权重进行调整, 利用下面等式进行计算,

[0154] 【公式 13】 $\omega^* = \omega * \beta_i$

[0155] 其中 β_i 为位置权重系数。 β_i 根据其在分析处理阶段 (A 步骤) 中识别出的各标题部分的位置信息, 取不同的值, 具体如下 :

[0156] β_1 表示说明书摘要、背景技术、具体实施方式部分的权重 ;

[0157] β_2 表示权利要求、技术领域部分的权重 ;

[0158] β_3 表示附图说明部分的权重 ;

[0159] β_4 表示标题、权利要求主题名称部分的权重。

[0160] 取值范围的关系满足不等式 1 :

[0161] $\beta_1 < \beta_2 < \beta_3 < \beta_4$

[0162] β_i 优选为 :

[0163] $0.1 < \beta_1 < 0.6$

[0164] $0.2 < \beta_2 < 0.8$

[0165] $0.3 < \beta_3 < 0.9$

[0166] $0.5 < \beta_4 < 1$

[0167] 且满足不等式 1 所限定的取值范围。

[0168] β_i 更加优选为 :

[0169] $\beta_1 = 0.4$

[0170] $\beta_2 = 0.5$

[0171] $\beta_3 = 0.6$

[0172] $\beta_4 = 0.8$

[0173] 进一步地, 提取方法可以使用上述方法的任意组合。

[0174] 短语提取模块将其提取的短语发送给短语判定模块。短语判定模块对提取的短语进行判定, 短语判定方法可以为短语频率方法, 即计算专利全文中该短语出现的频率, 按照

设定的选择阈值 ϵ ，如果出现频率小于该阈值，则排除该短语。短语频率的计算公式为

[0175] 【公式 14】

$$[0176] f_{NP} = \frac{C_{NP}}{C}$$

[0177] 其中， f_{NP} 为该短语的频率， C_{NP} 为该短语在专利全文中出现的次数， C 为专利全文中所有短语出现的总次数。 C 的计算公式为：

[0178] 【公式 15】

$$[0179] C = \sum_i N_i$$

[0180] 其中， N_i 为短语 i 在专利全文中出现的次数。

[0181] 阈值 ϵ 的计算公式为，【公式 16】

$$[0182] \frac{1}{N_{ALL}} \leq \epsilon \leq \frac{100}{N_{ALL}}$$

[0183] 更优选为：

[0184] 【公式 17】

$$[0185] \frac{1}{N_{ALL}} \leq \epsilon \leq \frac{20}{N_{ALL}}$$

[0186] 最优选为：

[0187] 【公式 18】

$$[0188] \epsilon = \frac{5}{N_{ALL}}$$

[0189] 其中， N_{ALL} 为全篇专利文献中短语的总个数。

[0190] 查询该短语是否存在于停用高频短语列表中，若存在，则排除该短语。

[0191] 短语判定方法还可以根据短语出现位置修正的短语频率法，

[0192] 【公式 19】 $f_{NP}' = f_{NP} * \beta_i$

[0193] 其中 β_i 为位置权重系数。在上面已有描述。

[0194] 短语判定方法还可以为记忆鉴定方法，所述专利文档库是一个专利全文数据库，其所含专利文档数量优选为一万篇以上。优选为根据待翻译的所述专利文档相同或相似的技术领域的专利文档库。短语判定方法还可以是上述 3 种方法的任意组合。如果应用了多种判定方法，可以通过投票法对结果进行选择。所述投票法表示用多种方法获得的短语中，取相同结果数量最多的一种。例如，有两种方法得出结果为“概率打分方法”，有一种方法得出结果为“打分方法”，则取“概率打分方法”为最终结果。

[0195] 经过短语判定的短语为候选短语。错误修正模块，对候选短语中可能的识别错误进行修正，同时修改句子中的标注信息。

[0196] 错误修正方法可以用 CRF 方法对候选短语进行概率打分，根据打分结果对于错误进行修正。打分公式为：

$$[0197] \quad p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left(\sum_j \lambda_j F_j(y, x) \right)$$

$$[0198] \quad F(y, x) = \sum_i^n f(y_{i-1}, y_i, x, i)$$

[0199] 其中, $f(y_{i-1}, y_i, x, i)$ 为转移概率或发射概率, y_{i-1} , y_i 是第 $i-1$ 和第 i 个标记, x 为观察序列。 i 为短语在观察序列中的位置。 $Z(x)$ 是归一化因子。 λ_j 是训练获取的参数。

[0200] 错误修正方法可以为规则方法, 根据上下文和相应的语法规则, 对错误进行修正。

[0201] 错误修正方法可以为错误模式方法, 对预先获得的所有错误模式进行记录, 放入存储器, 当判定后的短语符合错误模式时, 根据错误模式进行修正。

[0202] 【例 8】[其中气体发生器] 由两个部分构成 => 其中 [气体发生器] 由两个部分构成。上例中, 错误是将“其中”合并到名词短语中, 发现这种错误模式后, 根据错误模式进行修正, 将“其中”排除在名词短语之外。

[0203] 错误的修正方法, 还可以是结合上述 2 种或 2 种以上方法, 综合进行错误修正。在错误修正模块中, 还修改上述短语标注信息。经过错误修正步骤后获得的短语为识别名词短语 RNP。

[0204] 短语翻译模块, 用于翻译 RNP 短语并将结果保存到短语存储器中。短语翻译模块包含判断单元、修正单元、翻译及评分单元、对比单元, 图 6 是短语翻译模块的工作流程图。

[0205] 首先, 识别名词短语 RNP 进入判断单元, 判断其是否存在于短语存储器中, 如果存在, 则不作处理, 对下一条短语进行判断; 如果不存在, 进入修正单元。

[0206] 在修正单元中, 对识别名词短语 RNP 进行句法分析, 并将所述识别名词短语结构修正为以核心词 / 主题词作为根节点的结构;

[0207] 【例 9】词性 /n/ 标注 /nv 方法 /n, 其修正后句法分析结果如图 3 所示。在翻译及评分单元中, 对修正后的名词短语采用 CYK (Cocke–Younger–Kasami) 算法自底向上进行翻译, 在此过程中结合平均调序距离进行评分。所述平均调序距离 D , 作为调序结果选择的一项评分, 与预先设定的调序距离阈值 D_f 进行比较, 排除评分大于 D_f 的译文。

[0208] 平均调序距离公式为:

[0209] 【公式 20】

$$[0210] \quad \sum_i D = \sum_i L_i / Z$$

[0211] 其中 ω_i 表示第 i 个词调序前后所处位置的距离 $L_i = L_i^2 - L_i^1$ 。 Z 为词总数。

[0212] 【例 10】执行 [0] 命令 [1] 超时 [2] => Command[0]execution[1]timeout[2]

[0213] 执行 [0] => execution[1] $D_1 = 1$

[0214] 命令 [1] => Command[0] $D_2 = 1$

[0215] 超时 [2] => timeout[2] $D_3 = 0$

$$[0216] \quad \text{因此} D = \frac{(1+1+0)}{3} \approx 0.667$$

[0217] 所述 D_f 为经验值, 优选 $0.5 \leq D_f \leq 3$, 更加优选为 $1 \leq D_f \leq 2$, 最优选为 $D_f = 1.5$ 。

[0218] 接着,对 CYK 翻译过程获得的候选译文,保留得分最高的 N 个候选,N 优选为 100。

[0219] 在对比单元中,根据目标语言专利文档集信息进行重排序,就是将多个候选译文通过利用目标语言专利文档库训练获得的语言模型进行语言模型评分,评分最高者为最优译文,将其存储在短语存储器中,保存的信息包括名词短语、分词信息、词性标注信息、识别名词短语标签信息、译文信息。所述专利文档库是一个专利全文数据库,其所含专利文档数量优选为一万篇以上。优选为根据待翻译的所述专利文档相同或相似的技术领域的专利文档库。

[0220] 专利全文翻译模块是以句子为翻译单位的机器翻译模块或翻译器,对专利全文语句逐句进行翻译。

[0221] 根据本发明的机器翻译方法相对于现有的机器翻译方法的改进在于进行句法分析,对于标注为 RNP 的短语,作为名词 NN 处理,不再对其进行句法树展开,保留 RNP 为附加信息。进行翻译,对于标注为 RNP 的短语,从短语存储器中取出其译文作为短语译文;其他部分按现有的统计方法与规则方法、模板方法的一种或它们的结合翻译。

[0222] 输出模块从专利全文翻译模块获取所有句子翻译结果,按照原文的标题顺序输出译文。

[0223] <实施例 1>

[0224] 用根据本发明的机器翻译方法翻译如下专利全文,以下内容仅作为实施例给出本发明的工作方法的示例,省略了要旨之外的内容,本发明不限于本实施例。

[0225] 权利要求书

[0226] 1. 一种超低温热封聚丙烯流延膜,由热封层、聚丙烯芯层和聚丙烯电晕层三层流延共挤复合而成,其特征是所述热封层主要由以下组分按重量比制成:聚丙烯无规共聚物 10 ~ 80 份,聚烯烃弹性体 20 ~ 90 份,爽滑剂 0.1 ~ 0.5 份,防粘连剂 0.1 ~ 0.5 份。

[0227] 2. 根据权利要求 1 所述的超低温热封聚丙烯流延膜,其特征是所述热封层各组分的重量比为:聚丙烯无规共聚物 10 ~ 20 份,聚烯烃弹性体 80 ~ 90 份,爽滑剂 0.1 ~ 0.5 份,防粘连剂 0.1 ~ 0.5 份。

[0228] 3. 根据权利要求 1 所述的超低温热封聚丙烯流延膜,其特征是所述聚丙烯电晕层主要由以下组分按重量比制成:聚丙烯 100 份,防粘连剂 0.1 ~ 0.5 份。

[0229] 4. 根据权利要求 1 所述的超低温热封聚丙烯流延膜,其特征是所述聚丙烯芯层主要由以下组分按重量比制成:聚丙烯均聚物 100 份,苯乙烯 - 乙烯 - 丁稀 - 苯乙烯嵌段共聚物 3 ~ 5 份,爽滑剂 0.1 ~ 0.5 份。

[0230] 5.....

[0231]

[0232] 首先在用户界面中输入该文本,短语提取模块提取在全文中多次出现的短语:

[0233]

1	所述超低温热封聚丙烯流延膜
2	热封层
3	聚丙烯无规共聚物

4
---	-------

[0234] 经过短语判定模块进行判定,得出候选短语为:

[0235]

1	所述超低温热封聚丙烯流延膜
2	热封层
3	聚丙烯无规共聚物
4

[0236] 错误修正模块进行错误修正,例如,识别出1“所述超低温热封聚丙烯流延膜”有误,修正后结果如下。

[0237]

1	超低温热封聚丙烯流延膜
2	热封层
3	聚丙烯无规共聚物
4

[0238]

[0239] 经过错误修正模块进行错误修正后的短语,作为识别出的短语,对识别出的短语标注名词短语标签 RNP,识别模块将上述短语的短语原文、分词信息、词性标注信息、标签信息放入存储器。如下表所示,

[0240]

编号	短语原文	短语译文	词性标注信息	分词信息
1	超低温热封聚丙烯流延膜	...	RNP	...
2	热封层	...	RNP	...
3	聚丙烯无规共聚物	...	RNP	...
...

[0241] 短语翻译模块从存储器中取得短语原文进行翻译,翻译译文分别为:

[0242]

1	ultra-low temperature seal polypropylene cast film
---	--

2	sealant layer
3	random polypropylene copolymer
4

[0243] 短语翻译模块将译文存入存储器供其他模块使用。

[0244]

编号	短语原文	短语译文	词性标注信息	分词信息
1	超低温热封聚丙烯流延膜	ultra-low temperature seal polypropylene cast film	RNP	...
2	热封层	sealant layer	RNP	...
3	聚丙烯无规共聚物	random polypropylene copolymer	RNP	...

[0245]

...
-----	-----	-----	----	----

[0246] 句子翻译器根据分句结果,取得句子的分词、词性标注结果,在句法分析阶段,对标注为 RNP 的短语,作为名词 NN 处理,不再进行句法树展开,并保留 RNP 标签。在生成阶段,句子翻译器从词典中查找译文时,优先从存储器中获取译文,获得上述短语的译文,如下所示。

[0247] Claims

[0248] 1. An ultra-low temperature seal polypropylene cast film, by cast co-extruding a heat sealing layer, a polypropylene core layer and a polypropylene corona layer, Wherein said heat seal layer is mainly composed of the following components by weight ratio, random polypropylene copolymer of 10 to 80 parts, polyolefin elastomers of 20 to 90 parts, slippery agent of 0.1 to 0.5 parts, anti-blocking agent of 0.1 to 0.5 parts.

[0249] 2. The ultra-low temperature seal polypropylene cast film as claimed in claim 1, characterized in that each component of said heat-sealing layer weight ratio is : random polypropylene copolymer of 10 to 20 parts, polyolefin elastomer of 80 to 90 parts, slip agent of 0.1 to 0.5 parts, anti-blocking agent of 0.1 to 0.5 parts.

[0250] 3. The ultra-low temperature seal polypropylene cast film as claimed in claim 1, wherein said polypropylene alkenyl corona layer mainly consists of the following components by a weight ratio : 100 parts of polypropylene, 0.1 to 0.5 parts

of anti-blocking agent.

[0251] Copies.

[0252] 4. The ultra-low temperature seal polypropylene cast film as claimed in claim1, wherein said polypropylene alkenyl corona layer mainly consists of the following components by a weight ratio :100parts of polypropylene homopolymer, 3-5parts of Styrene-ethylene-Ding dilute-styrene block copolymer,0.1to0.5parts of slip agent.

[0253] 5.....

[0254]

[0255] 根据本发明的全篇专利文献机器翻译方法可以提高复杂名词短语的翻译准确性，降低了含有高频复杂名词短语的句法分析的难度，提高了句法分析的准确性，从而提高了翻译准确性，并减少了对高频短语进行句法分析的时间，从而提高了翻译速度。

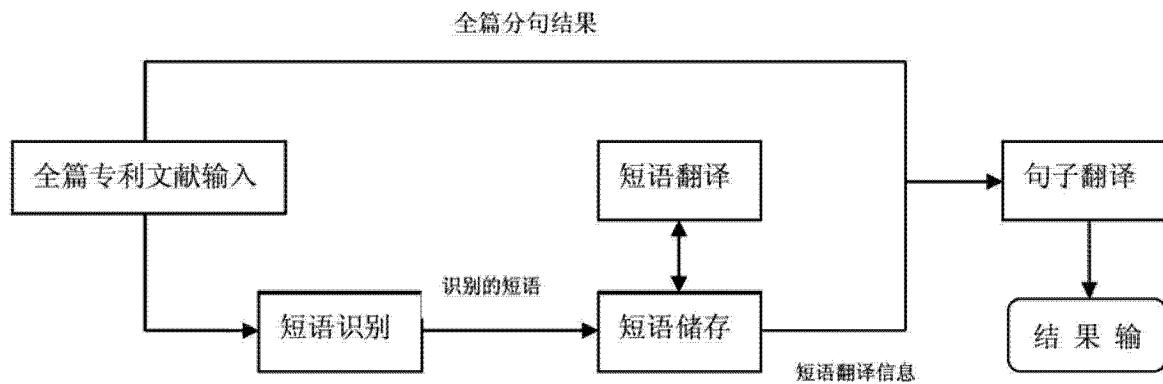


图 1

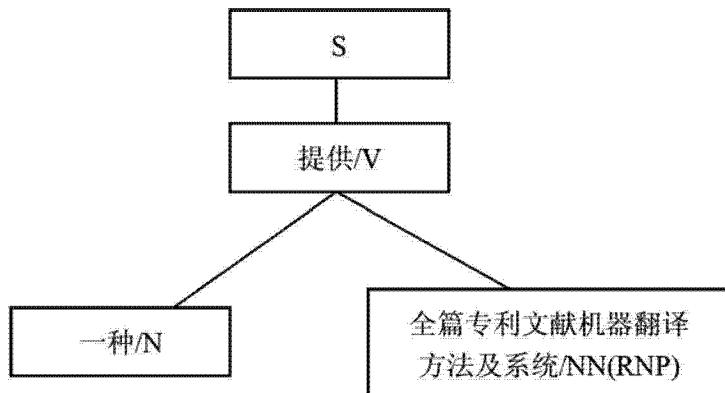


图 2

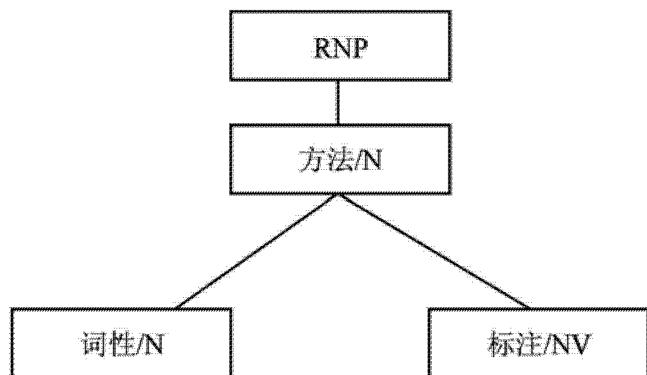


图 3

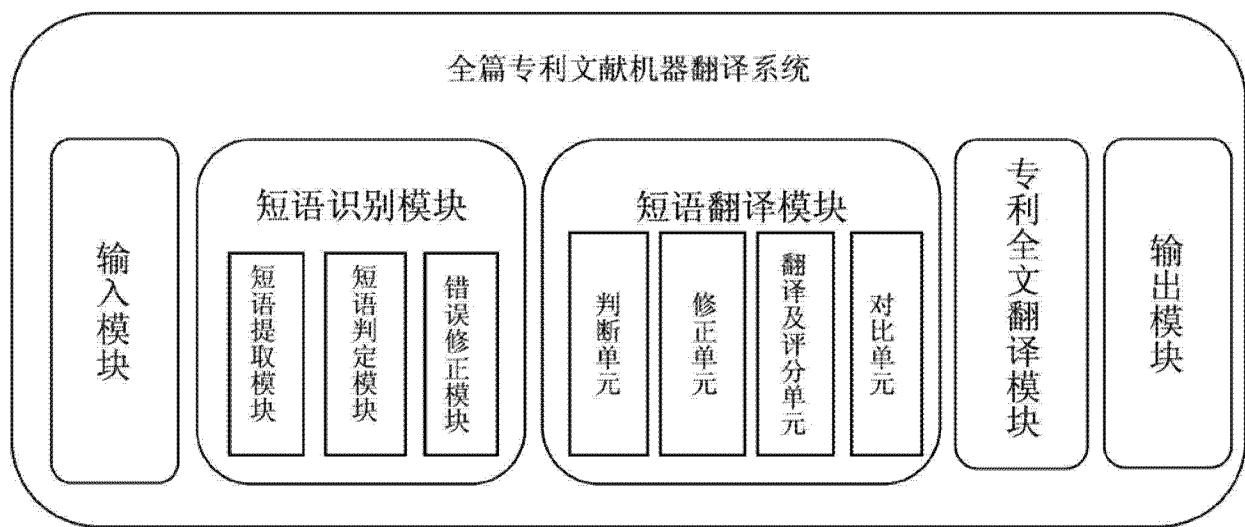


图 4

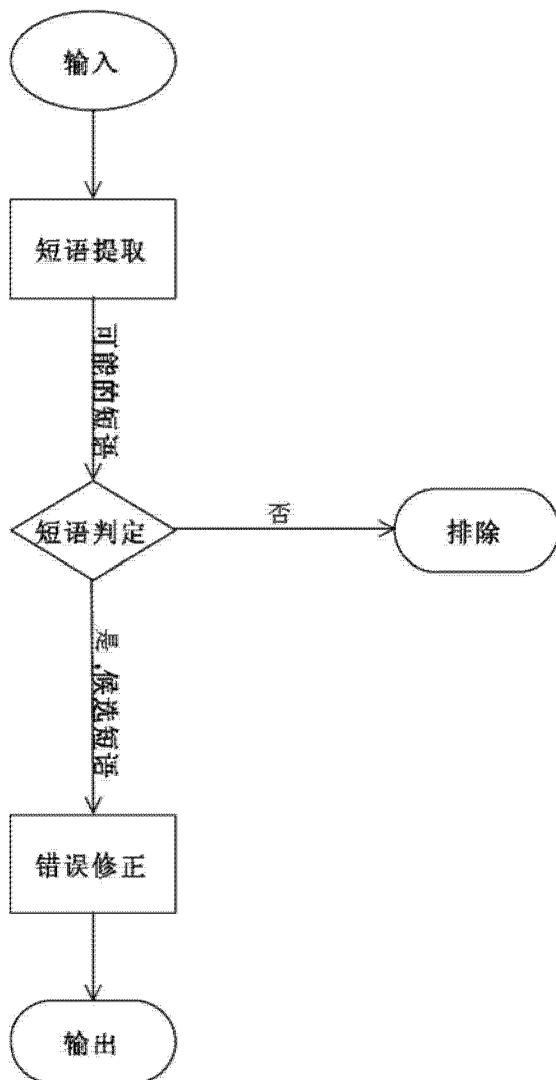


图 5

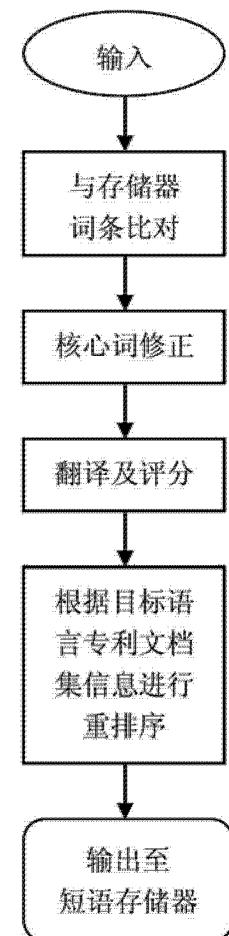


图 6