

# 俄汉机器翻译系统中基于多知识交叉分析的俄语词法分析

张心红\* 李向东\*\* 张蕾\* 闪烁\*

(\*中国科学院计算机语言信息工程研究中心 100083)

(\*\*北京外国语大学俄语学院 100081)

**摘要:** 词法分析是机器翻译的第一阶段,是后续的句法与语义分析的基础。以往俄汉机译系统中的词法分析主要依赖于词尾形态,用削、加词尾的方法查原形词,很难保证词法分析的准确性。本文介绍了一种俄汉机器翻译系统中基于语音、语法、语义多知识交叉分析的俄语词法分析系统,它将俄语词形变化特点与语音、语法、语义等信息联系起来,充分利用它们之间的相关规律,大大提高了俄汉机译系统 MAP-RCMT 词法分析的准确率。

## RUSSIAN MORPHOLOGICAL ANALYSIS BASED ON MULTIKNOWLEDGE ALTERNATE PROCESSING (MAP) IN RUSSIAN-CHINESE MACHINE TRANSLATION SYSTEM

Zhang Xinhong\* Li Xiangdong\*\* Zhang Lei\* Shang Shuo\*

(\*Research Center of Computer & Language Engineering, Chinese Academy of Sciences, 100083)

(\*\*Beijing Foreign Studies University, Beijing, 100081)

**Abstract:** Morphological analysis, the first step of machine translation, is basic for further syntactic and semantic analyses. In some RC MT systems, morphological analysis is mainly based on suffix forms, and the method of cutting or adding suffixes to the word is used in order to find its original form. This method can hardly ensure the correctness of analysis. This paper proposes a method of morphological analysis in Russian-Chinese MT system. It connects word forms with their phonetic, grammatical and semantic characteristics, and utilizes the coherence among them, therefore greatly increases the correctness of morphological analysis.

### 一、引言

词法分析是机器翻译句子分析处理的开始,其任务是根据词形变化规律鉴别具体输入的单词的词形特征,形成相应的词形信息,以供后续的句法与语义分析使用。从机器翻译角度看,俄语的词法特点可以归结为以下四点:1)形态范畴丰富,包括性、数、格、体、式、时、态等;2)词法变化形式多样;3)同形多义、同形歧义现象突出;4)各种词类的形态范畴有不同的组合关系(一致关系、支配关系、附加关系等)要求。如何总结利用俄语词形变化中的各种规律,准确地分析出原形词,并为后续的句法与语义分析提供尽可能多的语言信息,

这是俄汉机器翻译系统俄语词法分析的任务。

我国俄汉机器翻译研究工作开始于五十年代末。曾有过以中国科学院语言研究所和计算技术研究所为主研制的 59 型方案, 哈尔滨工业大学研制的 60 型方案, 中国科学院语言研究所、计算技术研究所、北京外国语学院和中国科学技术情报研究所研制的 61 型方案, 哈尔滨工业大学的 HLT-80 型方案等。就词法分析部分而言, 这些方案对单词的还原主要依赖于词尾形态, 用削、加词尾的方法查原形词。有些词法分析利用了俄语中词干同形的词往往属于不同词类这一规律, 将词典分成独立的组成部分(名词词典, 形容词词典和动词词典), 以减少各种同形现象[1]。但俄语的词形变化多种多样, 仅仅依靠词类划分, 很难保证词法分析的准确性。例如: жил(生活, 动词 жить 的过去式)一词会错误还原到 жилой(住人的, 形容词), 分析为 жилой 的形容词短尾形式。这是因为这种削、加词尾的查词方法, 不能判断 жилой 属于关系形容词, 没有短尾形式, 所以得出错误分析。

由此可以看出, 如果想使词法分析结果准确, 必须充分利用与俄语词形变化有关的各种规律性信息。究竟有哪些因素能影响一个词的形态变化呢?

同样是未完成体动词以-ать 结尾, 动词 делать(做)和 плакать(哭)的变位不同:

делать: делаю; делаешь; делают  
плакать: плачу; плачешь; плачут

同样以软音符号-ь 结尾, 名词 вождь(领袖)和 вещь(东西)的变格不同:

вождь: вождя; вождю; вождя; вождём; о вожде  
вещь: вещи; вещи; вещь; вещью; о вещи

同样是以-к 结尾的阳性名词, мальчик(男孩)和 учебник(课本)的第四格变化形式不同:

мальчик: мальчика  
учебник: учебник

делать 与 плакать 的变位不同, 是因为词尾-ать 之前的辅音 л 和 к 具有不同的语音性质; вождь 和 вещь 变格不同, 是因为它们分属阳性名词和阴性名词; 而 мальчик 是表人名词, учебник 是表物名词, 所以变格也不同。总结这些现象, 可以得出这样一个结论: 单词的形态变化不仅与词类、词干构成和词尾形式有关, 而且与语音、语法、语义有着非常密切的关系。因此, 要使词法分析准确、有效, 必须将词形变化特点与语音、语法、语义等信息联系起来, 充分利用它们之间的相关规律。

## 二、基于多知识交叉分析的俄语词法分析规则库

在中国科学院计算机语言信息工程研究中心与北京外国语大学俄语学院合作开发的俄汉机器翻译系统 MAP-RCMT 中, 我们根据上述原则建立了一套基于多知识交叉分析的俄语词法分析规则体系, 它总结了单词形态变化特点与其语音、语法、语义等特征的关系, 大大提高了俄汉机译系统 MAP-RCMT 词法分析的准确率。

俄语的词形变化如何与语音、语法、语义相关,俄汉机译系统 MAP-RCMT 的词法分析体系系统又如何利用它们之间的关系正确地分析出原形词,并为后续句法与语义分析提供准确、丰富的词法信息呢?下面我们就从语音、语法、语义三个方面分别讲述。

## 1、俄语词形变化与语音的关系

俄语的词形变化规律非常复杂.如名词的变格,根据其语法属性和形态特征,分别有三种变格法.其中,第一变格法和第二变格法根据词干末尾的辅音性质又有硬变化和软变化之分;形容词的变格根据其词干末尾辅音的不同有硬变化、软变化和混合变化三种变格法;动词的变位,根据其现在时和将来时单、复数不同人称词尾,分出第一变位法和第二变位法.这些纷繁复杂的变化在语音上有没有一定的规律可循呢?答案是肯定的.比如词干末尾是元音还是辅音,词干是否以唏音(ж, ш, ч, щ)或舌前塞音ц或舌根塞音(г, к, х)结尾,都与词形变化密切相关.以后缀-е ц为例:

当-е ц前面是元音字母时,е变成й.如китаец(中国人)---китайца;

当-е ц前面是辅音(л除外)字母时,е脱落.如отец(父亲)---отца;

当-е ц前面是-л时,е变成ь.如палец(手指)---пальца;

当-е ц前面是两个并列辅音时,е通常保留.如хитрец(狡猾的人)---хитреца.

由此可见,进行俄汉机译系统的词法分析,必须对单词内部的语音环境进行描述,才能得出正确的词形变化结果.所以,我们在制定词法分析规则时,设置了“适用条件”这一分析过程,即对词缀所在的语音环境进行限定,规定必须在何种语音条件下,才有何种词形变化.例如:

形容词词尾-ие还原为-ой的条件是必须能在词尾左边找到г, к, х, ж, ш这几个音,否则这一还原就不能成立.

同样,适用条件还可以限定词尾左边必须是元音或必须是辅音,元音或辅音的个数是几个,甚至可以限定词尾左边是一个将某些特殊辅音除去的辅音集合.适用条件这一分析过程的设置,将与单词形态变化相关的语音因素运用到词法分析中,大大提高了分析的准确性.

不仅如此,俄汉机译系统 MAP-RCMT 的词法分析中语音条件的引入,还解决了一个一直令俄汉机器翻译词法分析研究者比较头疼的问题---俄语词形变化中的不连续形态变化,语音交替.在俄语单词形变(如变格、变位等)时,词素中的某个音或音组可能被另一个音所代替,或者其中的某个元音消失,这种有规律的变化叫做语音交替[2].如动词писать(写)变位(пишу, пишешь,...)时,词根中的с变成ш.语音交替是俄语词形变化中一种常见的现象.这种交替有时发生在词尾,有时发生在词干.因此这种变化无法完全用词尾替换的方法来解决,而且这种词干的变化只出现在特定的语音环境中.对这一类单词进行词法分析,必须考虑语音环境条件.

针对语音交替这一现象,我们在词法分析中首先利用“适用条件”,对语音交替发生的语音环境进行描述,再运用“条件替换”的形式解决语音交替后的音变问题.“条件替换”指的是只有当单词的语音环境符合“适用条件”所限定的条件时,才进行适当的词尾替换.而替换之

前首先要将发生音变的字母还原为其所对应的单词原形中的字母。比如动词词尾-у还原到-ать的适用条件是必须在词尾左边找到ч, ж, ш, щ这几个音, 而还原时ч要首先还原为к或т(如плачу到плакать), ж要首先还原为з或г(如важу到вазать), ш要首先还原为с或х(如пишу到писать), щ要首先还原为ск或т(如ищу到искать), 最后再将-у替换成-ать。这样, 语音交替的问题就得到了解决。

## 2、俄语词形变化与语法的关系

语法与词形变化的关系更加明显。首先是不同的词类变化规律不同, 在同一词类中不同语法范畴变化规律亦不相同。同样是名词, 阳性、阴性、中性的变格方式不同; 同样是动词, 完成体和未完成体的变位方式不同。还有, 名词中可数名词和非可数名词的划分, 决定名词是否有复数形式; 一部分词不发生形态变化; 动词中有些词, 没有命令式的形式; 及物动词与非及物动词的划分, 决定该词是否有被动形动词的形式。上述这些影响词形变化的语法特征都应该成为词法分析的依据。

为此, 俄汉机译系统 MAP-RCMT 的词法分析系统引入了与语法相关的限定形式, 在词法规则中设置了〈属性检查〉这一项。其中的属性是指用该条规则还原后的原形单词所应具有的属性。

输入单词以-нный结尾, 将-нный换成-ть之后, 得到的原形单词必须是动词, 且必须是完成体。如果该原形单词在字典中的定义满足上述要求, 则规则运用成功。如果得到的原形单词在字典中不满足上述要求, 则不执行此规则。

这样, 单词деланный(形容词, 假装的)就不会还原成делать(未完成体动词, 做), 只有сделанный(完成体动词被动态, 做完), 才能还原到正确原形сделать。

由此可见, 语法条件的设置, 有效限制了词法规则的适用范围, 提高了词法分析准确率。

## 3、俄语词形变化与语义的关系

语义对词形变化的影响表现在许多方面。一个形容词是关系形容词还是性质形容词, 影响到这个形容词是否有短尾形式, 是否有比较级; 动物名词和非动物名词的变格不同, 动物名词第四格同于第二格, 非动物名词第四格同于第一格; 有些动词, 多表示客观状态或感觉, 只有第三人称形式, 如ныть(疼), хотеться(想要)等。这些较细微的语义内容, 都与词形变化有关。如果不加以描述并运用在词法分析中, 同样会出现许多错误分析结果。

因此, 俄汉机译系统 MAP-RCMT 的词法分析系统进一步引入了与语义相关的限定形式。这一限定形式同样运用了〈属性检查〉这一手段。语义属性则包括了形容词是关系形容词还是性质形容词, 名词是动物名词还是非动物名词, 动词是否表示客观状态感觉等特征。

仍以жил一词为例, 对其进行还原:

按照某一条词法规则, 空词尾被替换为-ой, 得到жилой。但是当程序检查字典中

жилой的语义属性时,发现这个形容词是关系形容词.而这条规则在属性检查部分限定,该规则只适用于性质形容词,不适用于关系形容词,所以不执行此规则.这样,错误的分析结果就被排除了.

以上介绍了俄汉机译系统 MAP-RCMT 的词法分析系统如何在词形变化与语音、语法、语义知识之间建立起相互约束关系,从而更加准确地对单词进行分析.而在实际规则应用中,这些语音、语法、语义知识并不是孤立地存在和发挥作用,而是相互联系,相互制约,从不同角度为规则词形变化的分析提供依据,有力地保证了词汇分析规则的高效、准确.

### 三、俄汉机译系统词法分析规则库的构成

俄汉机译系统 MAP-RCMT 的词法分析规则库由以下三部分组成:

1) 规则形态变化规则库,即俄语中有规律可循,有语音、语法、语义等知识作为分析依据的词形变化之形式描述,其一般形式为:

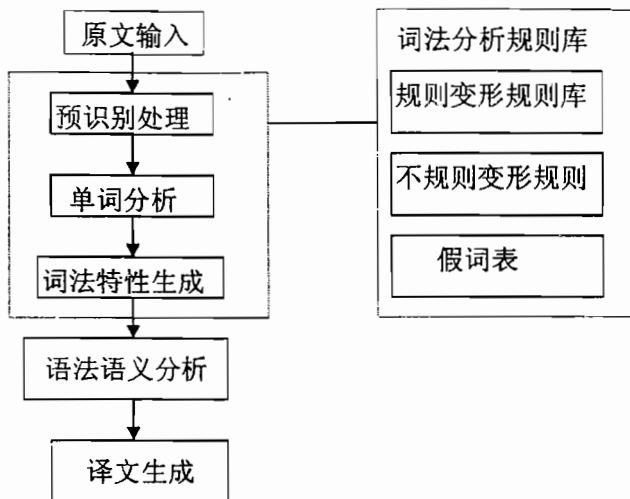
〈词尾模式〉 → 〈适用条件〉 〈还原操作〉 〈属性检查〉 〈词法特征〉

其中,适用条件和属性检查已经在前面作了介绍.“还原操作”则是指对输入单词进行还原处理的操作,主要是词尾的替换工作.为了适应不同情况的处理需要,设置了两种替换操作:“无条件替换”和“条件替换”.词法特征是对应于每一种词形变化的特征信息,如名词的性、数、格的变化,动词的变位等.词法分析的特征信息是词法分析的结果,它为后续的语法-语义分析提供了依据.

2) 假词表.假词表是为了有效处理俄语词法变化规律的局部例外情况而设立的.比如 беру (拿,动词第一人称)还原到 брать.一些词的词尾变化伴随词干尾部元音的脱落,比如 любви (爱,名词第二、三、六格)还原到 любовь.针对这类词,我们建立了假词表,当程序无法从字典中检索出该还原形式的定义时,就去检索假词规则库,取得正确的原形.假词表的设立可以解决一些通用规则难于覆盖的语言现象.

3) 不规则形态变化表.象大多数语言一样,俄语中有一些单词,词形变化特殊,无法用规则概括,比如 лет (年代,名词复数二格)还原到 год, выше (更高些,形容词和副词比较级)还原到 высокий和 высоко.对这类词汇我们作了特殊处理,建立了不规则形态变化表.

下面是俄汉机译系统 MAP-RCMT 词法分析器的结构图.其中,虚线框内的部分是词法分析器的组成.



#### 四、结束语

实际例句的翻译试验以及用户单位的试用结果表明,基于语音、语法、语义多知识交叉分析的俄语词法分析系统具有很强的分析能力和判断能力,能够对源语言绝大部分词汇做出正确的词法分析,准确率达 99%(千词测试).同时,它为后续的句法与语义分析提供了丰富而准确的信息.

当然,目前的俄汉机器翻译系统MAP-RCMT的词法分析系统也存在一些有待解决的问题.例如:词法规则中无法测知单词的音节特征;分析机制只能对输入单词进行逆向还原,没有进行正向变化检查;没有按照单词不同词法特征出现概率的大小,对输出候选词进行排序等.所有这些都对系统的翻译准确率和翻译速度产生影响,同时也提供了进一步研究的方向.

#### 参考文献

- [1] 刘涌泉、高祖舜、刘倬, 俄汉机器词典,《《机器翻译浅说》》,科学普及出版社,1964.
- [2] 钟国华、阎家业、龙翔 编著,《《实用俄语语法》》,辽宁人民出版社,1985.
- [3] Harald Trost: The applion of two-level mophology to non-concatentive German morphology.
- [4] 陈志忠,陈肇雄,高庆狮: 通用的自然语言词法分析机制,计算机学报,1991年2月,P93-99
- [5] Chang & Li, Symbolic logic and machanical theorem proving, Academic Press, New York, 1984.
- [6] J. HAJIC, Formal Morphology, COLING' 88, Budapest, 1988.
- [7] K. Koskenniemi and K. W. Church, Complexity, two-level morphology and Finnish, COLING' 88, Budapest, 1988.
- [9] L. Kataja and K. Koskenniemi, Finite-state description of semantic morphology: a case study of ancient Accadian, COLING' 88, Budapest, 1988.