面向汉阿句法短语库构建的阿语短语分析*

Alaa Mamdouh Akef [1,2]，杨尔弘 [1,2]，王莹莹 [1,2]
（1. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083
2. 北京语言大学 信息科学学院，北京 100083）

摘要：本文选择联合国官方资料库作为汉阿文本对齐语料，建立短语对齐。首先，利用斯坦福句法分析器对语料进行句法标注加工，进而抽取双语短语结构，构建汉阿句法短语库。在短语库的构建过程中，发现阿语语料的句法标注存在一些问题，因此本文着重对阿语标注中的这些问题进行分析。本文共抽取了 9 种阿语短语结构共 43796 条，统计短语结构分布，并着重对名词性短语结构和动词性短语结构中的句法标注错误实例进行归类总结，分析错误产生的原因。

关键词： 斯坦福句法分析器；汉阿平行语料库；短语结构

An analyzing of Arabic phrases for Chinese Arabic syntax phrase database study

Alaa Mamdouh Akef [1,2] , Yang Erhong[1,2] , Wang Yingying[1,2]
(1. National Print Media Language Resources Monitoring and Research Center, Beijing Language and Culture University, Beijing 100083, China;
2. School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: This paper collected Chinese Arabic aligned sentences which were drawn from UN official transcripts as this study parallel corpora, establish a Chinese Arabic syntax phrase database. First, corpus processing start with using the Stanford Parser to do syntactic tagging corpus processing, and then extracting bilingual phrase structure, build Chinese Arabic parallel phrases database. In the process of building a phrase database, we found that there are some problems in the syntactic tagging of Arab language corpus, so this paper analyzes these problems of the Arab language annotation. In this paper, we extracting 43796 phrases for 9 kinds of Arab language phrase structure, and after that, made Arabic phrase structure distribution, focusing on structure of noun phrases and verbal phrases structure of syntactic tagging error instances classified summarized, analyzed the causes of error.

Key words: Stanford Parser, Chinese-Arabic parallel corpora, phrase structure

---

* **基金项目：**北京语言大学来华留学博士研究生创新基金专项项目

作者简介：Alaa Mamdouh Akef（1987——），通信作者，博士研究生，主要研究领域为语言信息处理，Email: alaa_eldin_che@hotmail.com；杨尔弘（1965——），博士，教授，主要研究领域为语言信息处理、语言监测，Email:yerhong@126.com；王莹莹（1993——），硕士研究生，主要研究领域为语言信息处理，Email: ying_y_wang@126.com。

1. 引言

基于平行语料库的机器翻译研究是机器翻译的研究重点。平行语料库为机器学习翻译所需的句法规则提供了保障，也为统计机器翻译提供了资源。

对平行语料库建设而言，句法加工的质量决定机器翻译的质量。刘群仿照机器翻译中著名的金字塔结构，构建了一个统计机器翻译的金字塔，以确定各种统计机器翻译方法在金字塔中所处的相对位置[1]。基于短语的机器翻译方法处于核心位置，在两种语言分别进行词语对齐和句法分析的基础上，对两种短语结构进行对齐，其目的是为了获取短语翻译对。它更强调句法树与基于短语机器翻译的互动关系。

短语对齐是机器翻译领域中双语语料库加工中的一个难点问题[2]。双语语法体系的异构性导致很难把源语言句法树和目标语言句法树中的结点完全对齐起来[3]。阿拉伯语跟汉语的语法和句法体系相差很大，短语结构类型也有很大差别，这是汉阿短语对齐的一个很难解决的问题。

此外，句法分析作为短语对齐的基础，其准确率直接影响着短语对齐的结果[4]。本文在对汉语和阿拉伯语语料分别进行句法分析并构建汉阿平行短语库的过程中发现，阿拉伯语的句法分析结果正确率偏低，无法正确识别出短语结构，所以在汉阿短语对齐时出现了很多问题。究其原因主要是阿语特有的一套语法体系，机器无法正确处理其复杂多变的语法规则。

属于闪含语系的阿拉伯语是一种屈折语，阿拉伯语词一般分为三类，名词、动词和虚词。名词和动词的词根由动词派生而来，动词由两个部分构成：辅音构成词根，元音构成内容，即元音有语法作用，可以用来表明这个派生词是动词还是名词。现代阿拉伯语文本中很少见甚至不包含元音标注，但这些元音有表明词性的语法作用。对于无元音标注的文本，人和机器都可能会出现无法正确判断这个词的词性的问题，所以这是阿拉伯语的词性标注和句法分析的一个很大的困难[5]。

另外，阿拉伯语里的词缀用法很丰富，基本形式是“前缀+词根+后缀”，同一个词根可以跟不同的词缀复合，每个词缀都可以代表不同的词性。除了前后缀以外，阿拉伯语加中缀的系统也非常发达，即在词根中插进各种成分，构成新义。所以词语必须被正确划分才能找到其正确的词性和词义，否则就会产生词内划分歧义，影响句法分析结果的正确性。

本文采用斯坦福句法分析器 3.6.0 版本中的阿文文法[1]对阿文语料进行句法分析。因本文的语料存储是 UTF-8 编码格式，所以选择 arabicFactored.ser.gz 文法对阿拉伯语文本进行句法分析。

2. 语料来源及语料预处理

2.1 语料来源

为使研究具有可操作性，本文首先建立了汉阿平行语料库，语料来自联合国官方资料库，共有人工对齐的 7125 个句子。其中阿语语料中共有 827500 词次，汉语语料共有 265427 词次。

2.2 语料预处理

为了保证语料本身的正确性，以及使阿拉伯语语料更适应斯坦福句法分析器的阿文文法 arabicFactored.ser.gz，本文首先对语料的内容进行以下几项简单修正：

---

1 http://nlp.stanford.edu/software/parser-arabic-faq.shtml 斯坦福分析器的阿文文法是基于阿文滨州树库（Arabic Penn Treebank）而编成的，包含了三种文法,默认编码格式包括 UTF-8 和一种针对阿拉伯语的 ACSII 格式，Buckwalter encoding of Arabic in ASCII。

1. 错别字；
2. 少数不合理的阿拉伯语句子或者搭配错误；

斯坦福句法分析器无法对不合理的阿文句子进行句法标注，错误搭配也会影响句法分析的效果和正确率。

3. 标点符号；

斯坦福句法分析器在分析带分号的句子时，不把分句当成一个单独的句子，而是把整个句子作为一个句子分析。这样的句子在处理时，会因为句子过长造成内存不足，进而导致句法分析器抛出错误而中止运行。所以本文在对语料预处理时把中文语料和阿拉伯语语料中对应位置的分号改成句号，以使句法分析器能正确运行。

4. 标题

斯坦福句法分析器在处理语料中的标题时，因为标题末尾不含有句号，所以分析器可能把一个标题作为标题后面一句话中的一部分去处理，这样肯定会造成句法分析错误。针对这种情况，本文给中文语料和阿语语料中的标题末尾标上句号，以使句法分析器能正确分隔标题和正文首句。

3. 阿语短语提取及修正

本文在汉阿平行语料库的基础上，利用斯坦福句法分析器对平行语料进行自动句法分析；接着使抽取阿语句法分析结果中的9种短语结构类型；经过检查每种短语结构，修正并统计短语结构的错误类型。整个实流程如图1所示：

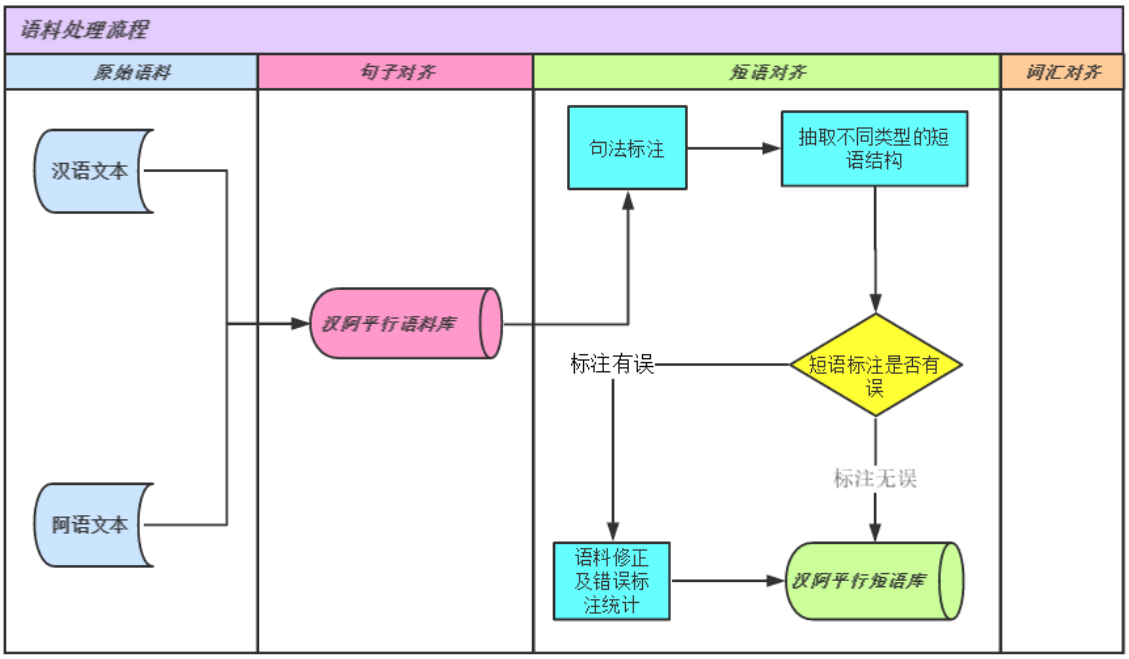

图1 语料处理流程

其中对短语语料的修正和错误类型统计是人机交互和自动修正相结合的。在开始修正语料之前，必须要先观察语料，为了减少人工的工作量，首先要尽可能多地找出那些可以用程序处理的标注错误类型。机器处理语料后需要人工地逐条检查剩余的短语结构，如果标注无误，则放入汉阿句法短语库中，如果标注错误，则判断其错误类型，放入标注错误的短语库中；最后对标注错误的短语库中的短语标注错误类型进行统计，形成错误标注类型统计表。

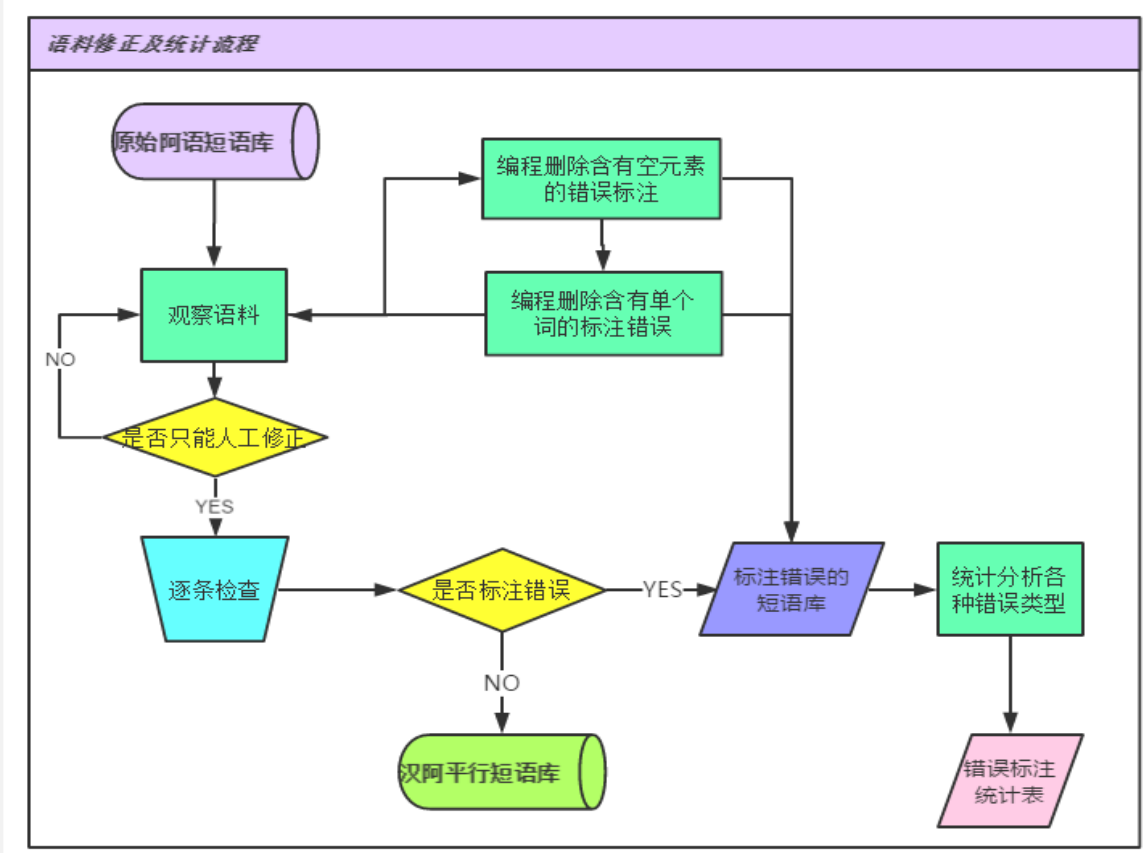

图 2 语料修正及统计流程

4. 阿语中的短语结构类型分布

斯坦福句法分析器的阿文文法使用的是阿文滨州树库（Arabic Penn Treebank）的词性标注集[2]和短语结构集，主要有 9 个短语类型，它们是 NP(名词性短语)、S（单句型短语）、VP(动词性短语)、PP（介词短语）、ADJP（形容词性短语）、SBAR（疑问副词开头短语）、WHNP（疑问副词短语）、FRAG（不完整的句子）和 ADVP（副词性短语）[6]。

本文从斯坦福句法分析器的分析结果中提取了以上 9 种短语结构类型，共 43796 条短语结构，各种类型的短语结构分布如表 1 所示：

2 斯坦福句法分析器使用形态分析器 Buckwalter analyzer 改善了阿文滨州树库的词性标注集，称为 augmented Bies 标注集，其标注与原标注的映射表参见 http://catalog.ldc.upenn.edu/docs/LDC2010T13/atb1-v4.1-taglist-conversion-to-PennPOS-forrelease.lisp

表 1 阿拉伯语中 9 种短语结构的分布

短语结构	功能类型	实例数	占比	频次	频率
NP	名词性短语	17479	39.91%	22435	37.36%
PP	介词短语	7437	16.98%	9068	15.10%
VP	动词性短语	6592	15.05%	6857	11.42%
S	单句型短语	5789	13.22%	11818	19.68%
SBAR	疑问副词性开头短语	3923	8.96%	4064	6.77%
ADJP	形容词性短语	1817	4.15%	2559	4.26%
WHNP	疑问名词性短语	23	0.05%	1980	3.30%
FRAG	不完整句子	634	1.45%	683	1.14%
ADVP	副词性短语	102	0.23%	590	0.98%
总计		43796	100.00%	60054	100.00%

根据表 1，我们可以得到名词性短语 NP 无论是短语数量还是在语料中出现的频次都排在第一位，排在第二位的是介词性短语 PP，第三位是动词性短语 VP。这三种短语类型跟汉语短语一样都是基本结构。阿语中的特殊短语是 SBAR（从属连词引导的从句短语）、WHNP（连接名词性短语，即先行词之后的短语）、FRAG（不完整句子）。

得到每种类型的短语结构后，经观察发现，每种类型的短语结构都包含有以下两个明显的错误：

1. 短语结构中有空元素

即句法分析器对一个空白的元素标注了词性或句法成分的情况，例如名词性短语实例*(NP (NN))*，NN 标注的是空元素，而且该实例在语料中的频次为 71，在名词性短语结构的频次统计中排第 5 位；再如动词性短语结构中按频次排序的前三位的短语实例是：(VP (VBD))、(VP (VBP))、(VP (VBD ( ) (NP (NN ج) (JJ)))，这三个短语里全都含有空元素，频次分别为 39、19、16。由此可以想象这种标注错误对句法分析的正确率的影响之大。

2. 短语结构中只包含一个词

即句法分析器把单个词标注为短语的情况，例如名词性短语结构中按频次排序的前四位的短语实例是：(NP (DT ذلك))、(NP (DTNN المادة))、(NP (NNP وقد))、(NP (PRP هي))这三个短语里都只包含一个词语，且频次分别为146、145、121、81。

删除上述两种错误实例后的短语结构分布表如表 2 所示：

表 2 删除空元素和单个词的短语实例后的短语结构类型分布

短语结构	功能类型	删除的实例数	剩余实例数	删除后的占比
NP	名词性短语	2440	15039	39.04%
PP	介词短语	329	7108	18.45%
VP	动词性短语	671	5921	15.37%
S	单句型短语	692	5097	13.23%

SBAR	疑问副词性开头短语	339	3584	9.30%
ADJP	形容词性短语	640	1177	3.06%
WHNP	疑问名词性短语	15	8	0.02%
FRAG	不完整句子	129	505	1.31%
ADVP	副词性短语	17	85	0.22%
总计		5272	38524	

5. NP、VP、PP 中的句法分析错误统计分析

现代阿拉伯语语法中一般把句子分为名词性句子[7]，即句首是名词，和动词性句子，即句首是动词。所以名词性短语和动词性短语在阿拉伯语中占有十分重要的地位。而且根据表 1 和表 2 的结果，我们可以得到，名词性短语 NP、介词性短语 PP 和动词性短语 VP 一直都是占比和频次最高的三种短语结构类型。观察抽取出的介词性短语，发现其标注正确率为 95%以上，错误率十分低。究其原因是阿拉伯语中只有 19 个虚词，常用的只有 11 个，句法分析器处理时很容易识别并标注正确这些虚词，其介词相关的错误主要体现在动词性短语和名词性短语中会出现一些介词标为动词或者名词的情况。因此，接下来我们选取名词性短语 NP 和动词性短语 VP 统计分析其中的错误类型。

5.1 NP 中的错误类型

阿拉伯语中的名词性短语 NP 是以名词为中心词组成的短语，需要注意的是阿语的名词与汉语中的名词有所不同，阿语名词有阴性、阳性之分，亦有单、双、复之分，其中双数和复数名词都是单数名词的变化形式，阿语名词也泛指名词和带冠词。

名词性短语 NP 中的错误类型主要包括以下几种：

1. 把介词标错为名词

斯坦福句法分析器的阿文文法中虽然有形态分析器 buckwalter，但是它有可能无法识别连写的介词，比如图 3 中“بالقوة用力”是介词和名词，其中的介词是连写的，句法分析器无法识别，所以造成了标注错误。

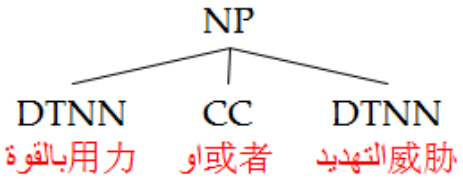

图 3 名词性短语错误实例 1

(NP (DTNN بالقوة用力) (CC او或者) (DTNN التهديد威胁))

2. 把动词标为名词

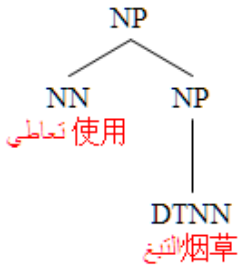

图 4 名词性短语错误实例 2

(NP (NN تعاطي使用) (NP (DTNN التبغ烟草)))

图 4 中的“تعاطي使用”是个动词，但却被标注为名词。在没有语境也没有元音标注表明其词性的情况下，句法分析器无法正确判断词性，故出现此种类型的错误。

3. 其他类型。

名词性短语中的各种错误类型的数量和占比如表 3 所示：

表 3 名词性短语中的错误类型统计

错误标注为名词的词性	实例数	占比
介词	939	6.24%
动词	786	5.23%
其他	203	1.35%
错误总计	1928	12.82%
标注无误	13111	87.18%

5. 2 VP 中的错误类型统计分析

阿拉伯语中的动词性短语 VP，即以动词为主体的短语，阿语的动词按照动作发生时的先后可以分成过去式动词和现在式动词，另外还有命令式动词。动词在句子中除了充当开头以外，就充当谓语。斯坦福句法分析器将阿语动词分为 VBD（动词性过去式）、VBP（动词性，非第三人称单数现在式）、VBD（动词性过去分词）、VBG（动名词性或现在分词）。

动词性短语 VP 中的错误类型主要包括以下几种：

1. 把介词标错为动词

阿拉伯语介词总共有 19 个，其中有几个介词也可以当虚词，这些虚词一般是连写的形式，所以也会造成前面提到的词内划分歧义。这些虚词使用时往往连接两个动词，从语法层面来看，这些虚词的作用是说明这两个动词的关系，如表示强调、解释、说明动作发生的原因是另一个动作等。所以它跟动词连接时会被句法分析器错误标注为动词，而不是介词。

2. 把名词标错为动词

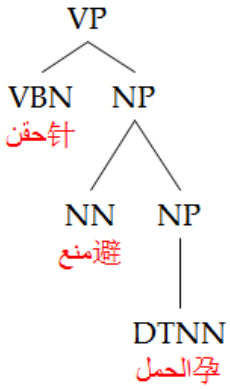

图 5 错误实例“避孕针”

(VP (VBN حقن针) (NP (NN منع避) (NP (DTNN الحمل孕))))

上述的例子是汉阿短语里边的一个子标题，是属于“حَقَن”的音标引起的歧义，医疗用语种的打针的针“حَقَن”和针的“حُقَن”元音标注（Vocalization）是不一样的。现代阿拉伯语文本中很少见甚至不包含元音标注，但这些元音有表明词性的语法作用。对于无元音标注的文本，斯坦福句法分析器会无法正确判断这个词是否动词，所以会有把名词错误标注为动词的情况。

3. 把“ان”标错为动词

阿文中的“إن”有两种标注，第一个标注是 VBP（非第三人称单数现在时动词）和 IN（介词），因为阿语传统语法理论不能够

解释“ان”的不同的用法，斯坦福句法分析器经常出现句法标注混乱的情况。

4. 其他类型

包括把虚词、指示词、代词、专有名词、数字和数词、连接名词性、疑问词标错为动词等 7 种类型，数量较少。

动词性短语中的各种错误类型的数量和占比如表 4 所示：

表 4 动词性短语中的错误类型统计

错误标注为动词的词性	实例数	占比
介词	425	7.18%
名词	393	6.64%
ان	56	0.95%
虚词	53	0.90%
指示词	52	0.88%
代词	48	0.81%
专名	37	0.62%
数字和数词	20	0.34%
连接名词性	2	0.03%
疑问词	2	0.03%
错误总计	1088	18.38%
标注无误	4833	81.62%

6. 结语

本文在句子层面对齐的汉阿平行语料库的基础上，利用斯坦福句法分析器对语料进行句法标注，抽取 9 种类型的短语结构；然后对短语结构中的错误进行修正，并以名词性短语和动词性短语为例对其中的标注错误类型进行统计分析；最后把标注无误的汉语短语和阿语短语人工对齐，构建汉阿平行句法短语库。在对名词性短语和动词性短语中的标注错误进行统计分析时发现阿语句法错误主要集中在名词、动词和介词两两之间的错误标注上。这些错误出现的原因在于现代阿语缺乏基于大规模的全面详细的语法信息词典以及作为阿文数据训练的滨州树库本身存在某些阿文句法标注问题。

参考文献

[1] 刘群. 统计机器翻译综述[J]. 中文信息学报, 2003, 17(4):1-12.

[2] 冯敏萱. 汉英平行语料库的平行处理[M]. 北京：世界图书出版公司, 2011

[3] 张春祥，高雪瑶著. 基于短语评价的翻译知识获取[M]. 哈尔滨：哈尔滨工业大学出版社, 2012.

[4] 刘群. 汉英机器翻译若干关键技术研究[M]. 北京：清华大学出版社, 2008.

[5] Ali Farghaly. Arabic Computational linguistics [M]. California: CSLI Publication , 2010

[6] pence Green and Christopher D. Manning, Better Arabic Parsing: Baselines, Evaluations? [C]. California: Stanford University, 2010.

[7] 周文巨，陆培勇. 阿拉伯语语法教程[M]. 上海：上海外语教育出版社, 2011.