Modern Linguistics 现代语言学, 2021, 9(4), 1008-1017
Published Online August 2021 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2021.94136

Hans汉斯

基于词典和规则的土耳其语形态消歧系统实现

张贵林，易绵竹，李宏欣，李　建，易晓宇

信息工程大学洛阳校区，河南　洛阳
Email: lihongxin830@163.com

收稿日期：2021年6月29日；录用日期：2021年7月27日；发布日期：2021年8月4日

摘　要

本文提出了一种基于形态分析词典和上下文环境约束规则的土耳其语形态消歧方法，通过文本预处理、命名实体识别、固定搭配识别、未登录词处理、形态分析和形态消歧共6个模块，构建了一个实用的土耳其语形态消歧系统。实验中，系统对随机选取的15份新闻文本测试数据进行处理，结果显示，与未加入消歧规则的基线系统相比，文本中78.57%的形态歧义得到了解决，形态句法标注准确率达96.84%，提高了1.7个百分点。

关键词

形态消歧，土耳其语，黏着语，混合方法

Implementation of a Lexicon and Rule-Based Morphological System for Turkish Text

Guilin Zhang, Mianzhu Yi, Hongxin Li, Jian Li, Xiaoyu Yi

Luoyang Campus of Information Engineering University, Luoyang Henan
Email: lihongxin830@163.com

Received: Jun. 29th, 2021; accepted: Jul. 27th, 2021; published: Aug. 4th, 2021

Abstract

This paper proposed a hybrid approach that solves morphological disambiguation problem based on a Turkish Frequency List lexicon and contextual constraint rules. On this methodology, we have created a practical Turkish morphological disambiguation system consisting of text preprocessing, named entity recognition, fixed collocation recognition, unknown word recognition, morphologi-

文章引用: 张贵林, 易绵竹, 李宏欣, 李建, 易晓宇. 基于词典和规则的土耳其语形态消歧系统实现[J]. 现代语言学, 2021, 9(4): 1008-1017. DOI: 10.12677/ml.2021.94136

cal parsing and morphological disambiguation, a total of six modules. In the test, 15 online news texts were randomly selected, and by combining constraint rules the system gets 96.84% of all the morphosyntax features correctly parsed on the test data. Compared with the baseline system without disambiguation rules, 78.57% of the morphological ambiguities in the text were resolved and the accuracy increased by 1.7%.

Keywords

Morphological Disambiguation, Turkish, Agglutinative Language, Hybrid Approach

1. 引言

土耳其语形态分析问题的相关研究始于上世纪 90 年代初。最早，土耳其学者奥富拉茨等人选择使用有限状态转移网络(FSTN)来进行土耳其语形态分析，他们基于形态配列规则和音位规则，利用一个包含大约 2.3 万个土耳其语词根的词典，在 PC-KIMMO 环境下成功构建了双层描述形态分析器[1]。该分析器在输入土耳其语单词的表面形式之后，会从词根或词干开始，把每一次派生或屈折变化都视为由一种状态向另一种状态的转移，最终通过搜索－匹配的方式输出一个包含该单词词根和一组形态标记的可能形态分析结果。这项研究成果影响深远，时至今日，大多数土耳其语形态分析研究仍然以此为基础。

土耳其语是一种典型的黏着语，通过词根后接词缀的方式，一个简单的土耳其语词根可以衍生出成千上万个不同的单词形式。词根是土耳其语中表示本体、概念和行为的最小语义单位。词缀是构成单词的语素，可以改变词根的语义、词性及句法功能，但不能独立成词[2]。一个土耳其语单词可以同时带有单复数、时态、人称、格等多个不同的信息特征，在进行语言信息处理时，以恰当的形式对这些信息加以解释，可以很好地缓解数据稀疏问题，而形态分析就是一种有效可行的方法。其中，标记可视为对单词所含形态、句法和语义信息进行编码，以实现同类信息的统一形式化表示。

几乎在所有语言中，标记的过程都会伴有歧义现象的出现，对土耳其语形态分析来说也是如此，即一个土耳其语单词的形态分析结果可能产生不止一个。在具体上下文环境中，为给出单词正确的形态分析结果，往往需要对形态歧义进行歧义消解，而这一过程则通常被看作是一种形态句法标注，也称形态消歧[3]。土耳其语形态消歧对于更高层次的语言信息处理任务而言，是一个非常有用的预处理过程，它的准确性将直接影响形态句法标记的效果，对土耳其语进行准确的形态消歧，将有助于句子的语义消歧、语言模型构建、拼写校正、语义角色标注和机器翻译等一系列语言信息处理任务。

根据研究内容的总体框架，论文第一部分主要介绍了土耳其语形态消歧相关研究成果；第二部分简要阐述了形态分析词典的来源及构建方法；第三部分详细描述了形态消歧系统的整体框架及相关规则；第四部分对形态消歧系统的性能进行了人工评测和分析，验证了本文所提方法的有效性；第五部分对全文进行总结，并对今后的研究方向进行了展望。

2. 相关研究工作

在自然语言处理中，土耳其语形态消歧可视为与英语词性标注任务处在同一层次上。土耳其语形态消歧方法归纳起来可分为三类：基于规则、基于统计和混合方法。

在基于规则的方法中，大量人工制定的规则或约束条件被用来选择具体上下文环境中单词正确的形态句法标记。1994 年，土耳其学者奥富拉茨和库鲁奥兹采用局部相邻约束条件、启发式算法和有限数量统计数据构建形态消歧系统，在人工选取的文本中取得了 97%~99%的准确率[4]。1996 年，奥富拉茨和图尔提出以约束条件为基础结合使用人工制定规则与非监督学习规则进行形态消歧，取得了 93%~94%的准确率[5]。此后，两人在 1997 年再次提出采用一种规则投票的方法来解决规则排序问题，利用单个规则对匹配分析进行投票，最后选择最高得分的形态分析，据称消歧准确率可以达到 94%~95% [6]。基于统计的方法通常利用词根和标记序列的统计数据来选择最佳的形态句法标记。2002 年，奥富拉茨与图尔等人首次提出利用隐马尔可夫模型(HMM)模拟土耳其语形态分析分布，通过构建三元语法基准模型，再使用维特比算法(Viterbi)进行计算的方式，最终找出文本中最佳的形态标记，三种模型最好的消歧结果可以达到 93.95%的准确率[7]。2008 年，萨克等人使用基准三元语法统计模型枚举每个句子候选形态分析序列的最佳 N 元列表，然后通过感知器算法对最佳 N 元列表进行重排序，列表中使用一个含有 23 个特征的特征集，将基准模型的形态消歧准确率从 93.61%提高到了 96.28% [3]。2011 年，格尔金和耶尔德兹基于分类的方法，利用 13 种分类器分别对模型进行训练，实验结果显示其中使用的 J48 树分类器消歧效果最好，其消歧准确率达到了 95.61% [8]。2016 年，耶尔德兹等人提出利用深度学习框架表示单词及向量空间模型依存关系，通过各层顶部训练的 softmax 层预测单词序列的最大相似度，然后在 softmax 层输出结果上使用维特比算法找到最佳形态分析序列，该方法在包含 20M 单词的土耳其语测试文本中解决了 85%的形态歧义[9]。在规则与统计相结合的混合方法上，2006 年，尤莱克和图尔提出通过贪婪前置算法监督训练形态消歧规则，利用生成决策表分别对单词潜在形态分析进行加权，然后根据置信度选取最终结果，该系统的形态消歧准确率达到了 95.82% [10]。2013 年，库特鲁和齐切克力基于规则与统计相混合的方法，利用 342 条形态消歧规则、后缀标记概率和启发式算法进行土耳其语形态消歧，取得了 94.1%的形态消歧准确率[11]。

虽然现有土耳其语形态消歧相关成果众多，但由于土耳其语本身的复杂性，相关研究一直没有出现突破性进展。近几年，随着深度学习方法相关研究的不断深入，高质量大型标记语料库的构建逐渐成为亟待解决的问题，作为一项重要的基础性研究内容，基于语言学规则的土耳其语形态消歧方法仍然具有独特的研究价值。

3. 形态分析词典

本文主要选用土耳其 TS-Corpus 在线语料库相关研究成果来构建形态分析词典。TS-Coprus 在线语料库是一个通用的土耳其语形态标记语料库，收录单词规模约为 4.95 亿，该语料库在 2011 年首次发布，可免费在线访问[12]。语料库数据主要涵盖新闻报道、会议报告、社交媒体和学术论文等多个领域的内容。2013 年，系统开发人员在统计数据基础上构建了一份基础形态分析频度表，数据规模约为 100 万左右，表内将单词在语料库中出现频次最高的形态分析结果视为该单词的唯一形态分析结果，本文将这一默认结果视为这些单词的最常见形态分析，用于形态分析词典的构建。

该语料库共包含四种标记信息：单词原形、词根、词性标记和形态分析标注序列。在语料库中，根据语义和语法特征，土耳其语单词被分为名词、动词、形容词、副词、代词、后置词、连词、限定词、重叠词、感叹词和疑问词等 11 个大的类别。之后，随着推特语料的引入，又逐渐增加了网络用语缩略词、网络用语强调词、表情符号、网络俚语、错词、HTML 实体、缩写、土耳其语式英语等 8 种词类，并分别赋予了相应标记。据统计，在语料库中除词性和词类标记之外，其它形态句法标记数量共计 86 个，部分标记如表 1 所示。本文形态分析词典所使用的标记符与 TS-Coprus 标记符基本保持一致，但在词类部分会对缩略语、错词、强调词等词类标记进行形式转换，并对部分单词的形态句法标记进行增删和修改。

Table 1. Part-of-speech tags and partial morphological feature tags of TS-Corpus
表 1. TS-Corpus 语料库词性标记及部分形态特征标记

词类	词性标记	词缀	形态标记
名词	Noun	主格	Nom
动词	Verb	宾格	Acc
形容词	Adj	从格	Abl
副词	Adv	向格	Dat
代词	Pron	位格	Loc
后置词	Postp	工具格	Ins
连词	Conj	等同格	Equ
限定词	Det	第一人称单数词缀	A1sg
重叠词	Dup	第二人称复数词缀	A2pl
感叹词	Intej	第三人称复数词缀	A3pl
疑问词	Ques	使动词缀	Caus
数词	Num	...	

4. 形态消歧策略及方法框架

本文通过选择最常见形态分析与规则相结合的形态消歧策略，来实现土耳其语的形态句法标注。选择最常见形态分析指的是，在经过准确形态句法标注的大规模语料库中，统计找出单词出现频次最高的形态分析形式，并将该形态分析选为该词在所有文本中的形态分析结果。通过最常见形态分析词典，可以快速实现土耳其语单词的形态分析，同时能够解决部分形态歧义问题。对于难以通过词典获得准确形态分析结果的歧义问题，则采用语法和上下文约束规则来完成歧义单词的形态句法标注。整个形态消歧系统主要包括文本预处理、命名实体识别、固定搭配识别、未登录词处理、形态句法标注和形态消歧等6个基本模块，整体框架如图 1 所示。

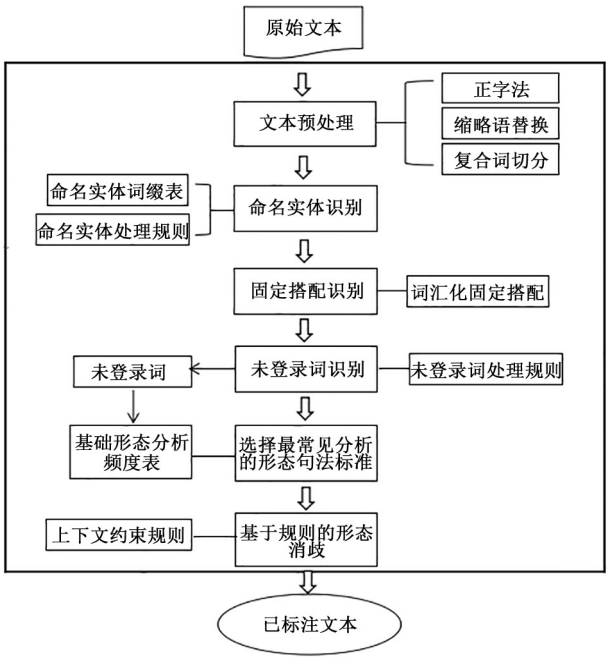

Figure 1. Flow chart of morphological disambiguation
图 1. 形态消歧流程图

4.1. 文本预处理

拼写错误、复合词和缩略语作为集外词(OOV)的重要组成部分，是影响土耳其语形态句法标注的三个重要因素，三者的处理与否将会直接影响系统的整体性能。我们将可做预处理的土耳其语拼写错误大致分成三类，第一种是单词拼写错误，第二种是多个首字母大写单词连写引起的拼写错误，第三种是字母“â”和“î”导致的同一单词的两种不同写法。对于包含字母“â”和“î”的单词，只需使用字母“a”和“i”进行替换即可。对于常见拼写错误和第二种拼写错误，则基于形态分析词典构建常见拼写错误替换词表，通过遍历、匹配和替换实现拼写校正。

包含字母“â”和“î”的单词实例：

pekâlâ　pekala

askerî　askeri

单词拼写错误实例：

alıcam　alacağım

yukaridakiler　yukarıdakiler

首字母大写单词连写实例：

AarhusBelediyesi　Aarhus Belediyesi

AbdullahAbdülkadiroğlu Abdullah Abdülk-adiroğlu

土耳其语复合词可以连写，也可以分写，一个新的复合词的意思一般可以通过拆分来进行理解。在不同的文本中，土耳其语复合词的拼写形式通常也会有所不同。例如，在规模为 500 万句的新闻语料中，土耳其语复合词 işadamları(企业家)的连写和分写形式出现的次数分别 1881 次和 1374 次。这种情况的存在会严重增加数据稀疏性，为此，需要对两种拼写形式进行统一。除了上述类型的复合词之外，系统还将对带有连词符“-”的复合词进行切分，切分后形态句法标注处理方式与复合名词相同。

复合名词切分实例：

Işadamları　iş adamları

sıradışı sıra dışı

Asya-Pasifik　Asya Pasifik

在缩略语的处理上，本文通过构建常用缩略语词表来实现缩略语与相应词语的转换，以进一步减少未登录词的个数。常用缩略语替换词表中的缩略语主要来自 TS-Corpus 形态分析词典，为了扩展词表的规模，本文还对训练语料中出现的未登录缩略语进行统计，通过频率选取常用缩略语，之后经过人工处理添加至词表。词表中主要包括三种类型的缩略语，第一类是不带句点“.”的土耳其文缩略语，第二类是翻译成土耳其语的英文缩略语，第三类是带句点“.”的缩略语。其中，对于带句点“.”缩略语的处理，可以减轻数据稀疏问题，也有助于提高分句的准确率。

常用缩略语词表转换实例：

带句点“.”的缩略语：

Dr.　Doktor

T.B.M.M.　Türkiye Büyük Millet Meclisi

翻译成土耳其语的英文缩略语：

Gov　Hükümet

Asp　Etkin Sunucu Sayfaları

不带句点“.”的缩略语:

TBMM Türkiye Büyük Millet Meclisi

TOBB Türkiye Odalar Ve Borsalar Birliği

4.2. 命名实体识别

根据土耳其语单词的书写规范，人名、地名、组织机构名等命名实体中的单词首字母均需要大写，且后接词缀时要求以单引号(’)隔开，其形态句法标记主要取决于单引号之后词缀的具体形态，想要得到该类单词的正确形态分析结果，只需要确定分隔符单引号后接词缀的形态句法信息即可。根据这一特点，本文在处理带单引号的单词时，首先通过分隔符单引号将单词分为实体单词与带单引号的词缀字符串两个部分，然后再对两者分别进行形态句法标记。其中，实体单词直接通过形态分析词典转换为单词词根形式，带单引号的词缀部分则使用命名实体词缀形态分析表转换为形态句法标记。

命名实体词缀实例：

'lerinizle Noun+Prop_Apos_A3pl_P2pl_Ins

'larınca Noun+Prop_Apos_A3pl_P3sg_Equ

对于不带单引号(’)的命名实体，实体类名词可以通过形态分析词典直接输出形态分析结果。对于时间类和数字类实体采用上下文约束规则的方法进行形态分析，在规则设置上，结合实体本身独有的特征，如“日－月－年”、“日－月”等固定型式，利用词类、单词和空格等约束成分判定实体类别。

4.3. 固定搭配识别

土耳其语固定搭配可分为词汇化与非词汇化固定搭配两类。词汇化固定搭配与短语(或词组)之间并无严格的界限，但词汇化固定搭配更加具有整体性，其作为一种固定的单一结构意义单位，并不是意义单位的简单相加，且单词之间不可再插入其它成分。例如：表示特定概念的固定搭配“çamaşır makinesi”(洗衣机)与类别短语“ahşap ev”(木屋)，“çamaşır makinesi”插入数词“bir”之后，固定搭配会失去原有意义，且不再具有固定搭配性质；表示材质的名词短语“ahşap ev”则可插入数词“bir”，插入之后意思基本不变，且仍是一个符合句法规范的名词短语“ahşap bir ev”(一座木屋)。鉴于词汇化固定搭配成份间的紧密关系，本文对词汇化固定搭配进行形态分析时，将参照复合词的处理标准，在考虑词汇化固定搭配整体形态句法属性前提下，不再对构成词汇化固定搭配的词根词进行标记。

词汇化固定搭配形态分析实例：

çamaşır Noun + A3sg + Pnon + Nom makine Noun + A3sg + P3sg + Nom → çamaşır makine Noun + A3sg + P3sg + Nom

doğru + Adj dürüst + Adj → doğru dürüst Adj

非词汇化固定搭配主要体现在语义和句法功能的变化上，利用这些词汇化和非词汇化固定搭配设置相应的规则，可以解决土耳其语部分形态歧义问题。例如，非词汇化结构“gelir gelmez”中，当单词“gelir”作为动词单独出现时，其形态分析结果为 gel + Verb + Pos_Aor + A3sg (词根 + 动词 + 肯定_一般现在时 + 第三人称单数)；当单词“gelmez”作为动词单独出现时，其形态分析结果为 gel + Verb_Neg_Aor + A3sg (词根 + 动词_否定_一般现在时 + 第三人称单数)；当两者均作为动词同时出现时，其句法功能相当于副动词。类似的非词汇化结构在土耳其语中还有许多，它们通常可以形式化表示为 R + X R + Y，其中 R 为词根，X 和 Y 为相同或不同词缀的形态句法标记。此类非词汇化固定搭配可以通过已标记文本固定搭配规则来进行处理。

4.4. 未登录词的处理

虽然系统中使用的标记语料库单词数量高达 100 多万，但是由于土耳其语严重的数据稀疏问题，导

致在处理过程中仍会遇到许多数据库中没有存储的单词形式。考虑到未登录词大多为名词和形容词，本文从句法功能角度出发将所有未登录词全都按照名词原形处理。此外，系统还设置了未登录词记忆模块，用户可以通过手动或者形态分析器对未登录词的形态分析结果进行调整，无歧义的形态句法分析可直接添加到基础形态分析频度表中。

4.5. 上下文约束规则

原始文本经过初步的形态消歧之后，生成的已标记文本存储的形式如下：

$[A_n, <B_n> <C_n>]$

$[A_{n+1}, <B_{n+1}> <C_{n+1}>]$

$[A_{n+2}, <B_{n+2}> <C_{n+2}>]$

……

其中，A_n 代表的是已标记文本中的第 n 个词符，n 为大于等于 1 的整数。B_n 代表的是已标记文本中的第 n 个词符的词根形式。C_n 代表的是已标记文本中的第 n 个词符的句法分析。

根据以上特点，我们共建立了以下几种类型的规则：

1) 动词型式构成的非词汇化固定搭配处理规则

如果 C_n = a，C_{n+1} = b，则$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_n + A_{n+1}, <A_n + A_{n+1}> <D_n>]$。其中，a 和 b 代表的是具体的词形、词根或者形态句法分析，D_n 是符合条件的情况下规定搭配具体的句法分析。

2) 形容词型式构成的非词汇化固定搭配处理规则

如果 $B_n = B_{n+1}$，$C_n = C_{n+1}$ = a，

则$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_n + A_{n+1}, <A_n + A_{n+1}> <D_n>]$；

如果 $B_n = B_{n+2}$，C_{n+1} = a，

则$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$ $[A_{n+2}, <B_{n+2}> <C_{n+2}>]$转换为$[A_n + A_{n+1} + A_{n+2}, <A_n + A_{n+1} + A_{n+2}> <D_n>]$。

3) 名词型式构成的非词汇化固定搭配处理规则

如果 B_n = a，$C_n = C_{n+1}$ = b，

则$[A_n,<B_n><C_n>][A_{n+1}, <B_{n+1}><C_{n+1}>]$ $[A_{n+2},<B_{n+2}><C_{n+2}>]$转换为$[A_n + A_{n+1} + A_{n+2}, <A_n + A_{n+1} + A_{n+2}><D_n>]$。

4) 基于词符所处位置的形态消歧规则

如果 A_n = a，C_{n+1} = <Punct>，则$[A_n, <B_n> <C_n>]$替换为$[A_n, <B_{n2}> <C_{n2}>]$。其中，a 是存在歧义的词符，B_{n2} 是该词符第二种形态分析的词根，C_{n2} 是该词符第二种形态分析的句法分析。

5) 基于上下文语法约束条件的形态消歧规则

如果 A_{n+1} = a，C_n 格标记 = c，则$[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_{n+1}, <B_{(n+1)2}> <C_{(n+1)2}>]$。

即 A_{n+1} = 单词 Bakan，C_n 格标记 = Dat，则<Noun + A3sg + Pnon + Nom>转换为<Verb + Pos + Adj + PresPart>。

6) 基于单词同现的消歧规则

在某些情况下中，两个单词同时出现通常具有较为固定的句法功能，此时，两者中的一个或者两个单词存在歧义，则可以建立下列规则形式进行处理：

如果 A_n = a，A_{n+1} = b，则$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_n, <B_n> <C_{n2}>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$；

或$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_n, <B_n> <C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{(n+1)2}>]$，或$[A_n, <B_n>$

$<C_n>]$ $[A_{n+1}, <B_{n+1}> <C_{n+1}>]$转换为$[A_n, <B_n> <C_{n2}>]$ $[A_{n+1}, <B_{n+1}> <C_{(n+1)2}>]$。

5. 实验结果及分析

5.1. 评价指标

为了能够较好地评价本文土耳其语形态消歧系统的性能，我们定义了以下测试评价标准：

形态句法标注的准确率：

$$准确率=\frac{正确标注的词符数量}{所有词符的数量}\times 100\%$$

该公式是为了对形态消歧系统整体进行评价，公式上方为测试时正确形态句法标注的词符的数量，公式下方为测试语料库中总的词符数量。

基于规则的形态消歧准确率：

$$消歧准确率=\frac{基于规则消歧的词符数量}{初步消歧后剩余歧义词符数量}\times 100\%$$

该公式是对消歧规则有效性进行评价，公式下方为初步态消歧后仍存在形态句法歧义的词符数量，公式上方为基于上下文环境约束规则能够正确形态消歧的词符数量。

5.2. 结果分析

为了评估消歧系统的性能，本文从土耳其早报、英国广播公司、土耳其外交部网站随机选取了 15 篇土耳其语新闻文本作为测试语料。每篇文章的词符个数约为 300 字左右，测试语料库总的词符数量为 4967 个，其中包含 690 个标点符号。对选取的文本，我们以选择最常见形态分析消歧作为形态句法标注基线系统，对利用上下文环境约束规则进行消歧的结果进行对比，根据人工统计，实验结果如表 2 和表 3 所示：

Table 2. Morphological disambiguation based on frequency table of basic morphological analysis
表 2. 基础形态分析频度表的形态消歧实验结果

未识别的非词汇化固定搭配	3
数词结构歧义	48
专有名词歧义	14
其他歧义	18
标注准确率	95.81%
标点符号除外的词符标注准确率	95.14%

Table 3. Morphological disambiguation based on context constraint rules
表 3. 基于上下文约束规则的形态消歧实验结果

未识别的非词汇化固定搭配	1
数词结构歧义	6
专有名词歧义	5
其他歧义	6
执行消歧规则后的标注准确率	97.14%
标点符号除外的词符标注准确率	96.68%

将表 2 与表 3 对比后可以看到，应用上下文环境约束规则进行形态消歧之后，测试语料库整体的形态句法标注准确率从 95.81%提高到了 97.14%，基于形态分析频度表未能解决的 84 个歧义词符中的 66 个词符得到了正确处理，消歧规则的应用处理了 78.57%的形态句法标注歧义。

在实验中，系统通过固定搭配规则正确地识别出了测试文本中出现的两个非词汇化固定搭配 olursa olsun (不管怎么样)和 devam edip etmemesinin (他是否继续的)。其中值得一提的是，在 devam edip etmemesinin 的处理中，devam edip 是一个词汇化固定搭配形式，由于两种固定搭配识别规则在处理过程和机制上相对独立，使得在非词汇化固定搭配规则的识别处理上可以有效避免与词汇化固定搭配规则的潜在冲突。

实验结果表明，系统利用消歧规则能够正确地处理测试语料中符合具体上下文约束条件的词符，特别是在多次出现的日期和动词形式歧义处理上效果明显。仍未能处理的歧义词符中，大多难以根据上下文环境建立有效的消歧规则，比如单词 Şayet (如果)，在频度表中的形态分析为<Conj>连词，文中作为名词与单词 Ömer 构成一个名词词组，两种情况通常都可以出现在句首的位置，对于这样的词通常很难构架一个有效的上下文约束规则。再比如单词 ne 可以有<Conj>和<Pron + Ques + A3sg + Pnon + Nom>两种不同的句法分析，当作为连词时通常会见到 ne...ne (即不也不)的形式，虽然此时具有一定的上下文关系，但是按照本文消歧规则的生成模式，并不能针对这种情况进行有效处理。由此可以看出，基于规则的形态消歧方法具有一定的局限性。

6. 结语

本文提出了一种选择最常见形态分析与上下文环境约束规则相结合的形态消歧策略，并在此基础上设计建立了一个土耳其语形态消歧系统。系统实验主要包括两大部分：一个是利用基础形态分析频度表对土耳其语原始文本进行形态句法标注，另一个是针对土耳其语上下文环境约束条件来制定相应的形态消歧规则。在系统设计过程中，我们通过添加手工标注的专有名词和未登录词，对原有基础频度表进行了一定程度的扩展，并针对常见的词汇化固定搭配形式建立了形态句法分析标记库。在实验中，我们对随机选取的 15 份新闻文本构成的测试语料依次经过两个消歧过程的处理，最终获得了 78.57%的形态消歧效果。实验结果表明，选择最常见形态分析与基于规则的方法相结合的形态消歧策略，可以有效处理形态句法标注问题。在后续的研究工作中，消歧规则库的扩展和完善，基础形态分析频度表的修正与扩充，以及未登录词问题处理方法的优化与探索，都是需要我们进一步研究的重点方向。

参考文献

[1] Oflazer, K. (1994) Two-Level Description of Turkish Morphology. *Literary and Linguistic Computing*, **9**, 137-148. https://doi.org/10.1093/llc/9.2.137

[2] Kaasandık, A. (2013) Türk Dili. https://turkdili.gen.tr/

[3] Sak, H., Güngör, T. and Saraçlar, M. (2007) Morphological Disambiguation of Turkish Text with Perceptron Algorithm. *Computational Linguistics and İntelligent Text Processing*, Mexico City, 18-24 February 2007, 107-118. https://doi.org/10.1007/978-3-540-70939-8_10

[4] Oflazer, K. and Kuruöz, İ. (1994) Tagging and Morphological Disambiguation of Turkish Text. *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart, October 1994, 144-149. https://doi.org/10.3115/974358.974391

[5] Oflazer, K. and Tür, G. (1996) Combining Hand-Crafted Rules and Unsupervised Learning in Constraint-Based Morphological Disambiguation. *Proceedings of the ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, May 1996, 1-13.

[6] Oflazer, K. and Tür, G. (1997) Morphological Disambiguation by Voting Constraints. *Proceedings of ACL'97*, Madrid, 11 July 1997, 222-229. https://doi.org/10.3115/979617.979646

[7] Hakkani-Tür, D.Z., Oflazer, K. and Tür, G. (2002) Statistical Morphological Disambiguation for Agglutinative Lan-

guages. *Computers and the Humanities*, **36**, 381-410. https://doi.org/10.1023/A:1020271707826

[8] Görgün, O. and Yıldız, O.T. (2011) A Novel Approach to Morphological Disambiguation for Turkish. In: Gelenbe, E., Lent, R. and Sakellari, G., Eds., *Computer and Information Sciences II*, Springer, London, 77-83.

[9] Yildiz, E., *et al.* (2016) A Morphology-Aware Network for Morphological Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**. https://ojs.aaai.org/index.php/AAAI/article/view/10355

[10] Yuret, D. and Türe, F. (2006) Learning Morphological Disambiguation Rules for Turkish. *HLT-NAACL'*06: *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, June 2006, 328-334. https://doi.org/10.3115/1220835.1220877

[11] Kutlu, M. and Çiçekli, İ. (2013) A Hybrid Morphological Disambiguation System for Turkish. *Turkish Natural Language Processing*, 53-67.

[12] TS-Corpus 土耳其语语料库[Z/OL]. https://tscorpus.com, 2012-2019.