文章编号：1003-0077(2018)10-0053-06

一种新的朝鲜语词性标注方法

金国哲，崔荣一

（延边大学 计算机科学与技术学院，吉林 延吉 133002）

摘　要：朝鲜语词性标注是朝鲜语信息处理的基础，其结果直接影响后续朝鲜语自然语言处理的效果。首先为了解决朝鲜语词性标注中遇到的形态素实际写法与原形不一致的问题，该文提出了一种在 seq2seq 模型的基础上融合朝鲜语字母信息的朝鲜语形态素原形恢复方法；其次，在恢复形态素原形的基础上，利用 LSTM-CRF 模型完成朝鲜语分写及词性标注。实验结果表明，该文提出的方法词性标注 F1 值为 94.75%，优于其他方法。

关键词：朝鲜语；词性标注；seq2seq；LSTM-CRF

中图分类号：TP391　　**文献标识码**：A

A Novel Korean POS Tagging Method

JIN Guozhe, CUI Rongyi

(Department of Computer Science and Technology, Yanbian University, Yanji, Jilin 133002, China)

Abstract: Korean POS tagging is the basis of the Korean information processing, and the result of POS tagging affects Korean Natural Language Processing directly. First of all, in order to solve the problem of inconsistency between the representation morpheme and original morpheme, this paper proposes a method of recovering the original form of Korean morpheme that integrates Korean Jamo information on the basis of seq2seq model. Then the LSTM-CRF model is used to achieve Korean spacing and POS tagging task. The experimental result shows that our method achieved 94.75% POS tagging F1-score, which is better than other methods.

Keywords: Korean; POS tagging; seq2seq; LSTM-CRF

0　引言

词性标注是指为句子中的每个单词标注一个正确词性的过程。词性标注是自然语言处理中的一项基本任务，是文本分类、机器翻译等其他自然语言处理任务的基础，同时在语音识别、信息检索等领域起着重要的作用。目前，英汉等语种词性标注研究比较成熟，而朝鲜语词性标注则较为落后，需要结合朝鲜语的语言特性，做深入研究。

朝鲜语句子由多个语节构成，而每个语节 eojeol 由多个形态素组成。其中语节是朝鲜语中的一个分写单位，而形态素则是具有意义的最小语言单位。例如，图 1 的句子中共有 3 个语节，其中每个语节由多个形态素构成，图中以“＋”作为形态素的分隔符。朝鲜语的词性标注任务就是以这些形态素作为单位的。例如，语节“나는”可以被标注为“나/代词＋는/助词”。

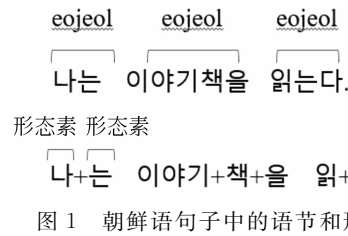

图 1　朝鲜语句子中的语节和形态素

中文词性标注任务中通常以分好词的单词作为标注单位，单词的词性变化较少。相比于中文，朝鲜语音节（字母）数较少（21 个元音＋19 个辅音＋27 个收音，如表 1 所示），而形态素组合较多，这也给朝鲜语词性标注增加了难度。另外，从图 1 中可以看到，朝鲜语词性标注任务伴随形态素分析的全过程，

收稿日期：2017-11-15　**定稿日期**：2017-12-21

基金项目：吉林省教育厅重点项目（吉教科合字[2016]第 250 号）

即把语节拆分成多个形态素的组合形式。然而,一部分形态素在句子中的写法与原形之间存在差异。例如,개발했다(实际写法)->개발+하+었+다(形态素原形)

表 1 朝鲜语字母表

辅音(共19个)	ㄱ, ㄲ, ㄴ, ㄷ, ㄸ, ㄹ, ㅁ, ㅂ, ㅃ, ㅅ, ㅆ, ㅇ, ㅈ, ㅉ, ㅎ, ㅊ, ㅋ, ㅌ, ㅍ
元音(共21个)	ㅏ, ㅐ, ㅑ, ㅒ, ㅓ, ㅔ, ㅕ, ㅖ, ㅗ, ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅝ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ
收音(共27个)	ㄱ, ㄲ, ㅆ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ

可以看到,在形态素分析过程中,形态素"했"转化为"하",这是因为"했"(表示已经做完)属于缩略语,其中包括了词干信息"做"和时态信息"已经"。解决这一问题的常用方法是利用语料库建立形态素变形词典,并利用词典完成形态素原形恢复。然而,基于词典的形态素原形恢复方法受限于语料库质量,存在处理不好未登录词等问题。针对这一问题,本文将形态素在句子中的实际写法和原形看作两种语言,用机器翻译中常用的 seq2seq 模型[1]解决这一问题。

另外,从更小的字母尺度上观察,可以发现上述问题中音节的组成部分—字母中包含了丰富的形态素信息。例如,音节"했"中收音字母"ㅆ"含有强烈的过去式时态信息。因此本文在 seq2seq 模型中融入字母信息(元音、辅音、收音),以提高形态素原形恢复准确率。

本文的主要工作有:

(1) 提出基于字母嵌入和 seq2seq 模型的朝鲜语形态素原形恢复方法。

(2) 基于 LSTM-CRF[2] 的朝鲜语音节单位词性标注方法及形态素为单位的标注合并。

下文内容安排如下:第 1 节介绍朝鲜语词性标注相关的研究,第 2 节详细描述朝鲜语词性标注过程中的形态素原形恢复方法及基于 LSTM-CRF 的以音节为单位的词性标注方法,第 3 节是实验过程及实验结果分析,第 4 节是结束语。

1 相关研究

常用的词性标注方法有:隐马尔科夫模型、最大熵模型、条件随机场模型等,而随着近年来深度学习的流行,基于 LSTM-CRF 的模型在词性标注任务中取得了良好的效果。朝鲜语词性标注过程中同样可以借鉴和使用这些模型。

朝鲜语词性标注的研究分为两类:第一类直接应用形态素分析器再结合统计模型完成词性标注,第二类则在词性标注过程中间接完成形态素分析。

朝鲜语属于黏着语,形态素之间的组合变化较多,若直接以语节为单位进行词性标注,则容易引起数据不足的问题。采用第一类方法的学者认为,利用形态素分析器预先做好形态素分析有助于解决数据不足问题[3]。这些方法大都是基于概率的模型,例如,隐马尔科夫模型[4]或最大熵模型[5]等。另外也有学者对上述两种模型进行了改进,例如,Lee S 等人[6]提出融合朝鲜语的多种语言特征,同时放宽隐马尔科夫假设,以此解决训练数据不足的问题。

反观第二类方法,通常在词性标注过程中,间接完成形态素分析。Han 等人[7]提出将朝鲜语词性标注任务分成三个阶段:语节为单位的形态素原形恢复;语节为单位的复合型词性标注;利用规则从语节为单位的复合型词性标注中获取形态素词性标注。其中第一步和第三步均采用词典完成转换工作。Shim 等人[8-9]的方法同样采用复合型词性标注,但因为采用了更小的音节为单位的复合型标注,因此相比于 Han 的方法,复合型标注集较小,一定程度上解决了训练数据不足的问题。Na 等人[10]同样采用了分三步进行词性标注的方法:利用 CRF 模型将语节分解成形态素,同样利用 CRF 进行形态素单位的词性标注,通过形态素原形恢复词典获得最终结果。本文提出的方法属于第二类,同时借鉴了 Shim 等人提出的音节单位词性标注方法。

2 朝鲜语词性标注

2.1 朝鲜语词性标注流程

首先将朝鲜语句子以语节为单位输入到形态素原形转换模块(该模块是利用同一个语料库提前训练好的模型),通过该模块将句子中的每个语节转换成对应的形态素原形,如图 2 中的第 1 步所示。之后将第 1 步的结果句以音节为单位输入到训练好的 LSTM-CRF 模型中。由于 LSTM-CRF 模型训练过程中针对无分写的句子进行训练,因此 LSTM-CRF 的输出实际上包含两类标注,其中一类是分写标注,另一类是词性标注。图 2 中第 2 步的实际输出如表 2 所示。

本文采用了标注集合$\{B, I\}$作为分写标注,其中 B 表示一个语节的首音节,I 代表语节中除首音节外的其他音节。

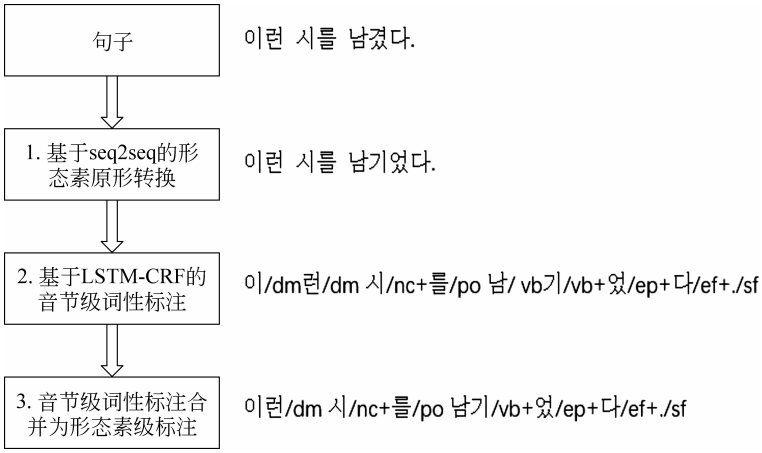

图 2　本文采用的朝鲜语词性标注流程

表 2　基于 LSTM-CRF 的词性标注、分写标注实例

音节	이	런	시	를	남	기	었	다	.
分写标注	*B*	*I*	*B*	*I*	*B*	*I*	*I*	*I*	*I*
词性标注	dm	dm	nc	po	vb	vb	ep	ef	sf

第 3 步将每个语节中(根据分写标注 B、I 判定语节边界)词性标注相同的相邻音节进行合并,输出最终的词性标注结果。

2.2　形态素原形转换方法

通过“分析世宗 21 世纪语料库”(人工标注过词性)中的 1 000 万条语节,我们发现了以下几条规律。

(1) 将近 19%的朝鲜语形态素在词性标注过程中发生了变形(由句子中的写法转化成形态素原形)。

(2) 转为原形的形态素长度大部分比实际写法增加了 1～3 个音节单位。

(3) 形态素变化过程中,音节的字母信息起着关键的作用,例如图 3 中的“워”的元音“ㅝ”被拆分成“ㅜ”和“ㅓ”,并与辅音“ㅇ”结合、构成两个音节“우어”。

基于以上分析,本文提出朝鲜语音节嵌入融入 seq2seq 模型的形态素原形转换方法,如图 3 所示。

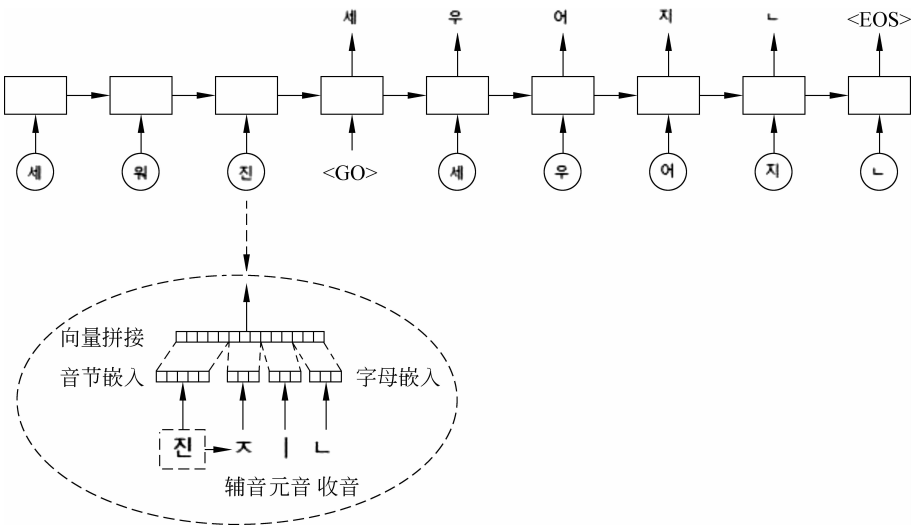

图 3　形态素原形转换模型

本文用 s 表示朝鲜语音节,其中 $s\in V_s$,V_s 为音节表。每个朝鲜语音节由辅音、元音、收音三个基本字母构成,分别用 J_f、J_y、J_s 表示这三个字母。例如,图 3 中的音节“진”可以拆分成 J_f=ㅈ、J_y=ㅣ、J_s=ㄴ。而有些音节,例如“세”,若不存在收音,则用特殊符号<NULL>替换之。具体的拆分可以利用朝鲜语在 unicode 编码的规则,按从上到下、从左到右的顺序完成。其中 J_f、J_y、$J_s\in V_J$,V_J 为字母表,$|V_J|$=

52(收音中去掉与辅音重复的字母共16个,再加上特殊符号<NULL>)。

下一步通过音节查询表 L_s 和字母查询表 L_J,将音节 s 和 s 对应的三个字母 J_f、J_y、J_s 转化成对应的音节嵌入向量和字母嵌入向量,计算过程如式(1)~式(4)所示。

$$\boldsymbol{e}_s = L_s(s), \boldsymbol{e}_s \in \mathbb{R}^d \tag{1}$$

$$\boldsymbol{e}_{J_f} = L_s(J_f), \boldsymbol{e}_{J_f} \in \mathbb{R}^k \tag{2}$$

$$\boldsymbol{e}_{J_y} = L_s(J_y), \boldsymbol{e}_{J_y} \in \mathbb{R}^k \tag{3}$$

$$\boldsymbol{e}_{J_s} = L_s(J_s), \boldsymbol{e}_{J_s} \in \mathbb{R}^k \tag{4}$$

其中音节嵌入 $\boldsymbol{e}_s$ 为 d 维实数向量,字母嵌入 $\boldsymbol{e}_{J_f}$,$\boldsymbol{e}_{J_y}$,$\boldsymbol{e}_{J_s}$ 均为 k 维实数向量。下一步通过向量拼接操作将向量 $\boldsymbol{e}_s$,$\boldsymbol{e}_{J_f}$,$\boldsymbol{e}_{J_y}$,$\boldsymbol{e}_{J_s}$ 拼接成(公式中用;表示)长向量 $\boldsymbol{e}$,作为 seq2seq 中编码器和解码器 Cell 的输入向量,计算如式(5)所示。

$$\boldsymbol{e} = [\boldsymbol{e}_s; \boldsymbol{e}_{J_f}; \boldsymbol{e}_{J_y}; \boldsymbol{e}_{J_s}], \quad \boldsymbol{e} \in \mathbb{R}^{d+3k} \tag{5}$$

为了保持字母嵌入向量的位置信息,未采用向量加或取向量平均,取而代之的是拼接操作。用 $S=\{s_1, s_2, \cdots, s_m\}$ 表示输入序列(一个朝鲜语语节的音节序列),用 $Y=\{y_1, y_2, \cdots, y_n\}$ 表示该语节的形态素原形序列。模型通过上述输入向量的生成方法把 S 中的每个音节转化成向量表示 $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_m\}$,作为编码器 LSTM 的输入。每一个 LSTM Cell 的计算如式(6)所示。

$$\boldsymbol{h}_t = f(\boldsymbol{e}_t, \boldsymbol{h}_{t-1}), \quad \forall t = 1, \cdots, m \tag{6}$$

其中 f 表示一个 LSTM Cell 的映射函数,通过当前输入 $\boldsymbol{e}_t$ 和 Cell 的前一个状态 $\boldsymbol{h}_{t-1}$,输出 Cell 的当前状态 $\boldsymbol{h}_t$。模型将最后一个状态 $\boldsymbol{h}_m$ 作为编码器对输入音节序列的编码向量,传递给解码器。

解码过程在训练和预测阶段有所不同。训练阶段将正确的形态素原形序列右移一个单位,左侧填充特殊符号<GO>(表示解码过程的开始),以此序列作为解码器的输入。反观预测阶段,由于形态素原形序列需要从模型中通过预测获取,因此将解码器上一个 Cell 的输出结果作为当前 Cell 的输入。

训练阶段解码过程如式(7)、式(8)所示。

$$\boldsymbol{h}'_t = f(\boldsymbol{e}'_t, \boldsymbol{h}'_{t-1}), \quad \forall t = 1, \cdots, n, \boldsymbol{h}'_0 = \boldsymbol{h}_m \tag{7}$$

$$\boldsymbol{o}'_t = g(\boldsymbol{e}'_t, \boldsymbol{h}'_{t-1}), \quad \forall t = 1, \cdots, n, \boldsymbol{h}'_0 = \boldsymbol{h}_m \tag{8}$$

其中解码器的第一个 Cell 的输入为代表起始音节的<GO>对应的音节向量,初始状态为编码器的终止状态 $\boldsymbol{h}_m$。g 为 Cell 的输出映射函数,最后通过 softmax 函数,将每一步 Cell 的输出向量 $\boldsymbol{o}'_t$,映射到 $y'_t \in \mathbb{R}^{|V_s|}$,计算过程如式(9)所示。

$$y'_t = \text{softmax}(\boldsymbol{W} \cdot \boldsymbol{o}'_t) \tag{9}$$

2.3 基于 LSTM-CRF 的朝鲜语词性标注方法

通过训练好的形态素原形转换模型,把原始朝鲜语句子转化成形态素原形表示的句子,我们用 $X=\langle x_1, x_2, \cdots, x_n\rangle$ 表示这样的一个朝鲜语句子,其中 x_i 为代表第 i 个音节的索引值,$Y=\langle y_1, y_2, \cdots, y_n\rangle$ 为一个句子的分写—词性标注序列。模型首先把 X 输入到音节查询表,通过查询将每个音节 x_i 转化成固定长度的低维实数向量。训练过程中将音节查询表当作可训练参数,进行动态更新。我们用 LT(X) 表示经过向量化的输入句。

下一步,为了更好地捕获音节前后上下文信息,将 LX(X)输入到双向 LSTM 网络中。假设输入音节 x_i 经过前向 LSTM 的 Cell 后的输出结果为 $\overrightarrow{h_i} \in \mathbb{R}^d$,经过后向 LSTM 的 Cell 后的输出结果为 $\overleftarrow{h_i} \in \mathbb{R}^d$,模型将这两个向量拼接(concatenate)成一个向量 $\boldsymbol{h}_i \in \mathbb{R}^{d\times 2}$。模型中 $\boldsymbol{h}_i$ 表示在考虑了第 i 个字符前后上下文的基础上,输入音节 x_i 的编码。

模型的最后一层通过 CRF 预测全局最优的分写—词性标注序列,计算如式(10)、式(11)所示。

$$s_{\text{syllable}}(i) = f(\boldsymbol{W}_{\text{out}} \boldsymbol{h}_i + \boldsymbol{b}_{\text{out}}) \tag{10}$$

$$s(X, Y', \theta) = \sum_{i=1}^{n} (\boldsymbol{A}_{y'_{i-1}, y'_i} + s_{\text{syllable}}(i)) \tag{11}$$

其中 $s_{\text{syllable}}(i)$ 表示输入音节 x_i 经过双向 LSTM 网络得到的分写—词性标注的概率分布,$\boldsymbol{W}_{\text{out}}$ 和 $\boldsymbol{b}_{\text{out}}$ 为全连接层的映射矩阵及偏置向量,f 为 softmax 函数。$\boldsymbol{A}$ 表示分写—词性标注状态的转移矩阵,例如,$\boldsymbol{A}_{y'_{i-1}, y'_i}$ 表示从标注状态 y'_{i-1} 到 y'_i 的转移概率。$s(X, Y', \theta)$ 代表同时考虑标注状态转移概率和双向 LSTM 预测的分写—词性标注概率时,输入音节序列 X 对应的一种候选标注序列 Y' 的分值,其中 θ 表示模型参数。模型从所有候选标注路径中取分值 $s(X, Y', \theta)$ 的路径作为最后的输出标注序列,这个过程可以通过标准的维特比算法有效地求解。

基于 LSTM-CRF 的朝鲜语词性及分写标注模型如图 4 所示。

3 实验

3.1 实验数据集

本文采用了“世宗 21 世纪词性标注语料库”,其中包括原始句文件和对应的词性标注句文件,共计 803 043 条句对。

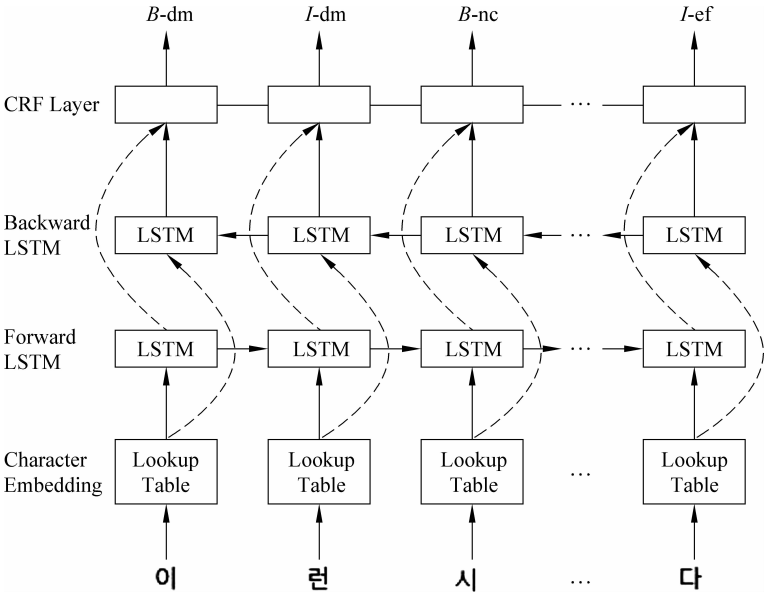

图 4　基于 LSTM-CRF 的朝鲜语词性及分写标注模型

针对形态素原形转换模型的预处理：

(1) 利用原始语料库的句子分写信息，将 86 万个句子拆分成 1 000 万个左右的语节，用于形态素原形转换模型。同时在标准的词性标注句中加入分写标注，用于训练 LSTM-CRF 模型。

(2) 生成字典：按照字符频率从高到低进行排序，取前 6 000 个字符作为字典，未出现在字典中的字符用＜UNK＞代替。

(3) 索引化：根据字典将第一步中的字符序列转化成对应字符的整型数字序列。另外，本文中涉及的 RNN 结构均采用 dynamic RNN，因此训练数据按照序列长度进行排序，生成各个 batch 后，再以 batch 为单位打乱 batch 间的顺序。

(4) 将预处理的数据按照 9∶1 的比例分成训练集和测试集。

实验中采用的数据集结构如表 3 所示。

表 3　数据集结构

模型	数据单位	训练集	测试集
形态素原形转换	语节	910 5591	1 011 732
LSTM-CRF 分写-词性标注	句子	722 739	80 304

3.2　实验设置

实验中采用了 tensorflow1.2 框架，并用 NVIDIA 的 1070GPU 进行了加速。

具体的模型参数配置如下：

(1) 形态素原形转换模型：编码器和解码器均采用了 4 层 LSTM 叠加的纵向结构和动态 RNN 横向结构，LSTM Cell 的大小为 256，batch size 设置为 128，学习率为 0.001，采用了 Adam 优化算法，经过 5 个 epoch 的训练最终得到朝鲜语形态素原形转换器。

(2) 训练 LSTM-CRF 模型：模型中双向 LSTM 网络的输入是大小为(128×None×128)的张量，其中第一维代表 batch size，第二维 None(每个 batch 的长度都不同)表示 LSTM 网络的步长(一个 batch 内序列长度均等于 batch 内最长序列的长度)，第三维表示音节向量的大小。LSTM 网络的输出部分将生成(128×None×256)的张量，其中 256 是前向和后向两个 LSTM 的 Cell 拼接而成的向量大小。最后通过全连接及 softmax 函数得到(128×None×90)的张量，其中 $\boldsymbol{W}_{\text{out}}$ 大小为 256×90，$\boldsymbol{b}_{\text{out}}$ 大小则是 90。其中数字 90 的解释如下：语料库中的词性标注集合共有 45 种，这些词性集合与分写标注集合(B，I 两种)组合形成 90 种输出标注集合。

3.3　实验结果及分析

首先，实验中复现了 Shim 等人提出的基于音节的形态素原形恢复词典的方法，并与本文提出的基于 seq2seq 的方法进行了对比实验。另外，为了验证朝鲜语字母向量的有效性，实现了两种 seq2seq 模型：音节嵌入＋seq2seq，音节嵌入＋字母嵌入＋seq2seq。表 4 中 P_{syllable} 为以音节为单位的原形恢复准确率，P_{eojeol} 为以语节为单位的形态素原形恢复准确率。

表4 形态素原形恢复准确率

方法	$P_{syllable}$/%	P_{eojeol}/%
基于词典的方法	96.25	94.32
音节嵌入+seq2seq	98.13	97.19
音节嵌入+字母嵌入+seq2seq	99.34	98.82

从实验结果中可以看到,相比于基于词典的形态素原形方法,基于seq2seq模型方法将音节单位准确率提高了2～3个百分点,将语节单位准确率提高了3～4个百分点。同时,由于取得了较高的形态素恢复准确率,最大限度地降低了本阶段误差传递到词性标注阶段,进而影响词性标注准确率的风险。

从错误分析中发现,基于词典的方法中90%的错误是由于未登录形态素引起的。音节嵌入+seq2seq方法中,以语节为上下文,对音节进行建模,因此即使遇到未登录形态素,也可以进行准确的预测。另外,与第二种方法相比,第三种方法中还引入了字母嵌入,因此可以更好地对形态素实际写法到原形映射关系进行建模,同时提高模型的泛化能力。例如,"당연한->당연하ㄴ"的映射中模型学习到了"한->하ㄴ"的形态素映射关系。当在测试集中遇到未登录词干"대견"时,仍能根据字母嵌入给出准确的结果"대견하ㄴ"。

其次,在相同的"世宗21世纪词性标注语料库"条件下,实验中复现了相关研究中的几种典型的朝鲜语词性标注方法,分别是Lee等人提出的基于隐马尔科夫模型的方法、Han等人2004年提出的形态素原形词典结合统计模型(利用马尔科夫假设)的方法、Shim等人2013年提出的以音节为单位的CRF模型,本文提出的基于seq2seq模型的形态素原形转换+基于LSTM-CRF的词性标注方法。表5给出了各个模型的实验结果。

表5 实验结果

模型	$P_{syllable}$ /%	$P_{morpheme}$ /%	$R_{morpheme}$ /%	$F1_{morpheme}$ /%
HMM	93.12	88.45	87.86	88.15
形态素原形词典+统计模型	94.53	91.23	89.58	90.40
音节单位的CRF模型	96.05	93.34	92.72	93.03
基于seq2seq模型的形态素原形转换+基于LSTM-CRF的词性标注	96.79	95.64	93.88	94.75

可以看到,本文提出的方法在音节级准确率、形态素级准确率、形态素级召回率以及F1值均高于其他现有的方法,其中F1值相比于现有最好的音节单位的CRF模型提高了1.72个百分点。与音节单位的CRF模型相比本文提出的方法在音节标注准确率上较为接近。然而本文中提出的基于seq2seq的形态素转换方法提供了较高的形态素原形恢复准确率,因此在形态素级的词性标注任务中由于形态素本身的原形错误导致的词性标注错误极少,这也帮助我们提高了词性标注准确率。

4 结束语

本文提出了一种新的朝鲜语词性标注方法。该方法将朝鲜语词性标注过程分为三步:第一步利用seq2seq模型将朝鲜语形态素以语节为单位转化成原形;第二步利用LSTM-CRF模型以句子为单位进行音节级词性标注;第三步根据音节级分写及词性标注进行合并,得到最终的形态素级的词性标注结果。相比于现有最好的音节单位的CRF模型,本文提出的方法将F1值提高了1.72个百分点。未来工作中我们希望尝试最近较为流行的端到端的训练模型,并进一步挖掘朝鲜语本身的语言特征,用于提高朝鲜语词性标注的准确性。

参考文献

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of Advances in Neural Information Processing Systems,2014: 3104-3112.

[2] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv: 1508.01991, 2015.

[3] Ahn Y M, Seo Y H. Korean part-of-speech tagging using disambiguation rules for ambiguous word and statistical information[C]//Proceedings of International Conference on IEEE, 2007: 1598-1601.

[4] Lee S Z, Lim H S, Rim H C. Two-level part-of-speech tagging for Korean text using hidden markov model[C]//Proceedings of Annual Conference on Human and Language Technology. Human and Language Technology, 1994.

[5] Kang I H, Kim J H, Kim G C. Korean part-of-speech tagging based on maximum entropy model[C]//Proceedings of Annual Conference on Human and Language Technology. Human and Language Technology, 1999.

(下转第68页)

metric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(2): 2319-2323.

[18] Li Y F, Kwok J T, Zhou Z H, Cost-sensitive semi-supervised support vector machine [C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10), 2010: 500-505.

[19] UCI database [J/OL]. http://www.ics.uci.edu/%20mlearn/MLRepository.html.

[20] Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifying imbalanced data [J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[21] Silva J, Bacao F, Caetano M, Specific land cover class mapping by semi-supervised weighted support vector machines [J]. Remote Sensing, 2017, 9(2): 181-196.

[22] David H A, Gunnink J L. The paired t test under artificial pairing [J]. The American Statistician, 1997, 51 (1): 9-12.

周国华(1977—),硕士,讲师,主要研究领域为智能学习、模式识别。
E-mail: tiddyddd@sina.com.cn

宋洁(1981—),硕士,讲师,主要研究领域为智能识别。
E-mail:songjie@czili.edu.cn

殷新春(1962—),教授,博士生导师,主要研究领域为人工智能、密码学。
E-mail: xcyin@yzu.edu.cn

(上接第 58 页)

[6] Lee S, Tsujii J, Rim H C. Hidden Markov model-based Korean part-of-speech tagging considering high agglutinativity, word-spacing, and lexical correlativity [C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000: 384-391.

[7] Han C H, Palmer M. A morphological tagger for Korean: Statistical tagging combined with corpus-based morphological rule application[J]. Machine Translation, 2004, 18(4): 275-297.

[8] Shim K S. Syllable-based POS tagging without Korean morphological analysis[J]. Korean Journal of Cognitive Science, 2011, 22(3): 327-345.

[9] Shim K. Morpheme restoration for syllable-based Korean POS tagging[J]. Journal of KIISE: Software and Applications, 2013, 40(3): 182-189.

[10] Na S H, Yang S I, Kim C H, et al. CRFs for Korean morpheme segmentation and POS tagging[C]//Proceedings of 24th Annual Conference on Human and Cognitive Language Technology, 2012: 12-15.

金国哲(1983—),通信作者,硕士研究生,讲师,主要研究领域为自然语言处理。
E-mail:34200519@qq.com

崔荣一(1962—),博士研究生,教授,主要研究领域为智能计算、机器学习、模式识别。
E-mail:cuirongyi@ybu.edu.cn