

文章编号:

基于分写增强字符向量和 LSTM-CRF 的朝鲜语自动分写方法*

金国哲, 崔荣一

(延边大学计算机科学与技术学科, 吉林 延吉 133002)

摘要: 朝鲜语自动分写问题类似于中文分词问题, 属于朝鲜语自然语言处理中的基本问题。首先, 针对传统的朝鲜语自动分写方法中依赖人工特征的问题, 本文提出一种朝鲜语分写增强字符向量训练模型 KWSE, 用于获取包含语义及分写倾向性信息的字符向量。其次, 将朝鲜语分写增强字符向量与 LSTM-CRF 模型结合完成朝鲜语自动分写任务。实验结果表明本文提出的方法其单词级分写 F1 值为 92.86%, 优于其他方法。

关键词: 朝鲜语; 自动分写; 分写增强字符向量; LSTM-CRF

中图分类号: TP391

文献标识码: A

Automatic Korean Word Spacing Using Spacing-Enhanced Character

Embedding and LSTM-CRF

JIN Guozhe, CUI Rongyi

(Dept. of Computer Science and Technology, Yanbian University, Yanji, Jilin 133002)

Abstract: Automatic Korean word spacing which is similar to Chinese word segmentation problem belong to fundamental problem in Korean natural language processing. First of all, to overcome the disadvantage of traditional method which dependent on manual extracted feature, we proposed a Korean spacing-enhanced character embedding model KWSE. Through this model, we can obtain the character embedding containing semantic and spacing polarity information. Secondly, we combine Korean spacing-enhanced character embedding with LSTM-CRF to achieve Korean spacing task. The experimental result shows that our method achieved 92.86% F1-score, which is better than other methods.

Key words: Korean; automatic word spacing; spacing-enhanced character embedding; LSTM-CRF

1 引言

朝鲜语分写法(也被称作隔写法), 是朝鲜语语法中最基本的原则。例如中文“我喜欢读书”的正确分写方式为:

나는 책읽기를 좋아합니다.

正确的朝鲜语分写有助于快速、准确地理解文章的含义。反观, 糟糕的分写方式直接影响到句义。例如:

아버지가 방에 들어가셨다.

아버지 가방에 들어가셨다.

这两句由相同的字符序列构成, 而且两种分写方式在语法角度上都是合法的, 但是根据分写方式的不同表现出不同的含义。第一句表示“爸爸进屋了”, 第二句则是“爸爸进包里了”。上述例子虽然有些极端, 但也可以反映出朝鲜语分写的重要性。

正规的书籍、报纸、期刊等出版物中, 由于进行细致的人工校对, 朝鲜语分写错误相对

* 收稿日期:

定稿日期:

基金项目: 吉林省教育厅重点项目(吉教科合字[2016]第250号)

作者简介: 金国哲(1983—), 男, 讲师, 自然语言处理; 崔荣一(1962年—), 男, 教授, 智能信息处理;

较少。然而网络环境下使用的朝鲜语中存在大量的分写错误,这些错误不仅影响文章的理解,而且不利于朝鲜语规范化使用。因此,有必要引入朝鲜语自动分写系统,帮助用户纠正分写错误。朝鲜语自动分写系统读入含有错误分写的句子或未分写的句子,输出准确分写的朝鲜语句子。

朝鲜语自动分写系统的主要用途有:

- 1) 文档的自动化分写纠错
- 2) 词性标注,命名实体识别等其他朝鲜语自然语言处理的预处理模块
- 3) 朝鲜语 OCR 或语音识别系统的后处理模块

朝鲜语自动分写是典型的序列标注问题,本文将分写增强字符向量(KWSE)与双向长短时记忆循环神经网络-条件随机场(LSTM-CRF)模型结合解决这一问题。后续章节中统一使用 KWSE-LSTM-CRF 代表这种方法。章节安排如下:第二章介绍朝鲜语自动分写相关的研究,第三章详细描述基于 KWSE-LSTM-CRF 的朝鲜语自动分写模型,第四章是实验过程及实验结果分析,第五章是结论。

2 相关研究

现有的朝鲜语分写方法可以归类到以下两大类:第一类是基于规则的方法,第二类是基于统计的方法。

基于规则的方法主要是利用语言学家建立的专家级规则库与目标句子进行匹配,进而完成朝鲜语自动分写。Kim&Lee(1998)等人基于规则的方法根据启发式规则进行句子分写,并用句子形态分析的方法验证分写结果^[1]。Kang(2000)等人利用双向最大匹配的启发式规则进行了朝鲜语自动分写^[2];该方法在分写过程中使用了语言学家手工构建的分写规则,因此获得了较高的分写准确率。这种方式虽然可以在局部获得相当高的分写性能,但缺点是泛化能力差,分写准确率严重依赖于规则库。

基于统计的方法将分写问题转化为在未分写句子的适当位置插入空格的问题,其基本思路是通过分析原始语料库中的字符级 n-gram 信息获取字符间分写概率,并通过这些概率值决定是否进行分写(Chung & Lee^[3], 1999; Jeon & Park^[4], 2000; Kang & Woo^[5], 2001)。Lee, Rim et al^[6](2007)等人将朝鲜语自动分写问题归类到序列标注问题,并提出用隐马尔科夫模型 HMM 解决该问题。Shim^[7](2011)则提出基于条件随机场(CRF)的朝鲜语自动分写方法。条件随机场被认为是解决序列标注问题的一个有效模型,在中文分词等领域同样表现出了优异的性能。Lee 和 Kim^[8](2013)等人尝试用 Structured SVM 解决朝鲜语分写问题。Lee C, Choi E^[9](2014)等人提出了 BWSM 方法,该方法在 Structured SVM 的基础上融入人工分写标注信息,用于提高单词级分写准确率。Hwang^[10](2016)等人提出了基于 GRU-CRF 的朝鲜语分写方法,该方法利用循环神经网络(采用 GRU 单元)计算朝鲜语分写标注序列,同时利用 CRF 求全局最优的分写标注序列。

上述基于统计的方法无需人工构建分写规则库,且该方法从原始语料库中获取分写概率,因此相比于基于规则的方法具有更好的泛化能力,并且对未知语句进行分写时具有更好的鲁棒性。然而,上述基于统计的方法存在一些问题:基于 HMM 的方法需要大量的训练参数,CRF 和 Structured SVM 方法则依赖于特定的特征,并且未能充分利用好句子的全局上下文信息。为了克服上述问题,本文将尝试用 KWSE-LSTM-CRF 解决朝鲜语自动分写问题。

3 朝鲜语自动分写方法

3.1 朝鲜语分写标注

本文将朝鲜语自动分写问题归类为序列标注问题。标注集的选择上借鉴了汉语分词领域中常用的四词位标注法,即标注集定义为 $T = \{B, M, E, S\}$,其中 B 表示单词的起始字符, M 表示单词中间字符, E 表示单词的结束字符, S 表示单字符单词。例如,朝鲜语:

그는 이 대회에서 일등을 따냈다.
(他在这次大会中取得了第一名)

对应的标注如表 1 所示。

表 1 朝鲜语句子分写标注实例

X	그	는	이	대	회	에	서	일	등	을	따	냈	다
Y	B	E	S	B	M	M	E	B	M	E	B	M	E

本文用 $X = \langle x_1, x_2, \dots, x_n \rangle$ 表示一个句子序列，其中 x_i 表示句子中的字符， $Y = \langle y_1, y_2, \dots, y_n \rangle$ 表示该句对应的分写标注序列，其中 y_i 表示 x_i 对应的分写标注。

3.2 分写增强字符向量

本节提出朝鲜语分写增强字符向量的训练模型 KWSE。KWSE 模型在 Mikolov 等人的 CBOW^[11]模型基础上，增加了分写标注信息，目的是让训练得到的字符向量同时具有语义和分写倾向信息。CBOW 模型借鉴了 C&W^[12]模型中以上下文中间单词作为预测单词的做法，同时简化了神经网络语言模型，去掉了隐层。CBOW 模型的优化目标为最大化：

$$\sum_{t=1}^V \sum_{s \in \text{context}(t)} \log p(t|s) \quad (1)$$

其中 V 表示词表大小， $\text{context}(t)$ 为单词 t 的上下文窗口，通常单词 t 的前后各 c 个单词作为 $\text{context}(t)$ 。CBOW 模型用 softmax 函数定义概率 $p(t|s)$ ：

$$p(t|s) = \frac{\exp(w_t^T v_s)}{\sum_{j \in V, j \neq t} \exp(w_j^T v_s)} \quad (2)$$

其中 w_t 为单词 t 的输出向量， v_s 为 $\text{context}(t)$ 中单词的平均向量：

$$v_s = \frac{1}{2c} \sum_{t-c \leq j \leq t+c, j \neq t} w_j \quad (3)$$

CBOW 模型中训练出的词向量富含语义信息，即语义相似的单词向量之间的距离也会较近。而在朝鲜语分写问题中，我们希望字符向量含有一部分分写倾向性信息，具体如下：

1. 同一字符向量在不同的上下文环境中表现出不同的分写倾向性。

그는 이 대회에서 일등을 따냈다.
그는 독서를 즐긴다.

上述两句中朝鲜语字符“서”在第一句中是助词，因此它的后面应该是一个分写。而字符“서”在第二句中是单词“독서”（读书）的组成部分，而且后面跟着一个助词“를”，因此无需进行分写处理。

2. 不同的字符向量，如若它们的适用场景类似，则分写倾向性也应较为接近。

例如分写上下文相似的“은”和“는”，其字符向量的分写倾向应该较为接近，向量空间中的距离应该较短。这种字符向量有助于提高后续分写标注任务的准确率。

基于以上分析，我们提出了一种朝鲜语分写增强字符向量的训练模型。该模型以目标字符的上下文字符及其标注信息作为输入，输出目标字符的分写标注。模型结构如图 1 所示。

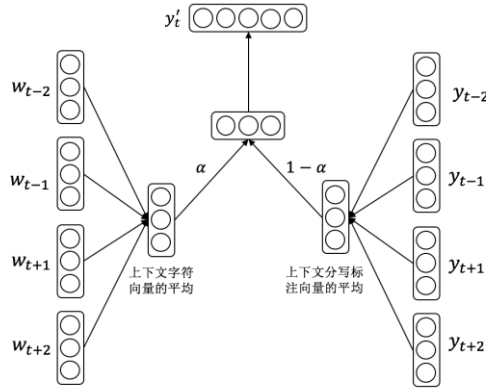


图 1 KWSE 模型

其中 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ 为中心字符 w_t 的上下文字符, $y_{t-2}, y_{t-1}, y_{t+1}, y_{t+2}$ 为上下文字符对应的分写标注, y'_t 为模型预测的分写标注。

模型首先通过查询表把上下文字符转化成低维实数向量, 查询表大小为 $|V| \times d$, 其中 $|V|$ 为字典大小, d 为字符向量的维数。另外, 分写标注通过类似的查询表转化成 d 维实数向量, 分写标注查询表大小为 $|M| \times d$, 其中 $|M|$ 为分写标注集大小, d 为分写标注向量维数。

下一步计算上下文字符的平均向量 v_s , 计算方法与公式(3)相同, 而上下文分写标注的平均向量 u_s 的计算公式如下:

$$u_s = \frac{1}{2c} \sum_{t-c \leq j \leq t+c, j \neq t} y_j \quad (4)$$

下一步是求 v_s 和 u_s 的加权和, 并通过 softmax 函数输出模型预测的分写标注。

$$p(y'_t | u_s, v_s) = f(w_{out}(\alpha \cdot v_s + (1 - \alpha) \cdot u_s) + b_{out}) \quad (5)$$

其中 α 为超参, 表示语义信息和分写倾向信息的比例, 实验中设置为 0.6 (通过多次的对比实验得到的经验值), f 表示 softmax 函数, $w_{out} \in d \times |M|, b_{out} \in |M|$ 为全连接映射权值和偏置。

模型采用标准的 BP 算法进行训练, 并在训练完成后提取字符查询表, 作为 LSTM-CRF 朝鲜语分写模型的字符查询表的初始化参数。

3.3 LSTM

循环神经网络 (RNN^[13]) (Elman1990) 可以记住序列数据的历史信息, 并根据历史信息和当前的输入预测当前的输出, 因此适合对序列标注问题进行建模。然而, 传统的 RNN 在实际训练过程中存在梯度消失和梯度爆炸的问题。

LSTM^[14] 属于改进版的循环神经网络 (RNN), 相比于传统的 RNN 模型, LSTM 可以更好地对长距离依赖关系进行建模, 同时可以很好地解决梯度消失和梯度爆炸问题。图 2 是一个典型的 LSTM 结构。

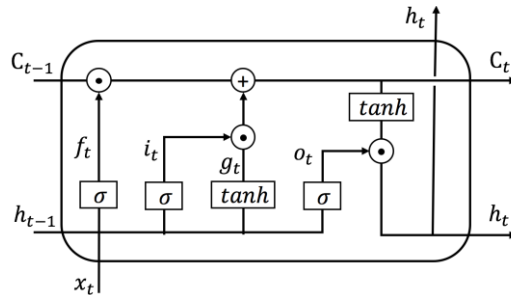


图 2 LSTM 单元

LSTM 的隐层由特殊构建的记忆单元 Cell 构成。每个 Cell 由以下四个部分组成: (1)

循环连接的 Cell，(2) 用于控制输入信号流量的输入控制门*i*，(3) 用于控制流向下一个单元的信号强度的输出门*o*，(4) 用于控制遗忘之前 Cell 状态的遗忘门*f*。下面给出每个时刻*t*，各个单元的计算公式。

$$i_t = \sigma(W_i \cdot [h_{t-1}; x_t] + b_i) \quad (6)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}; x_t] + b_f) \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}; x_t] + b_o) \quad (8)$$

$$g_t = \tanh(W_C \cdot [h_{t-1}; x_t] + b_C) \quad (9)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (10)$$

$$h_t = o_t \odot \tanh(c_t) \quad (11)$$

其中 \odot 表示元素级乘法计算， σ 表示 sigmoid 函数， W_i ， W_f ， W_o ， b_i ， b_f ， b_o 分别为输入门、遗忘门、输出门的权值矩阵。

3.4 基于 LSTM-CRF^[15]的朝鲜语分写方法

我们用 $X = \langle x_1, x_2, \dots, x_n \rangle$ 表示一条朝鲜语句子，其中 x_i 为代表第 *i* 个字符的索引值， $Y = \langle y_1, y_2, \dots, y_3 \rangle$ 为一个句子的分写标注序列。模型首先把 X 输入到字符查询表 (Lookup Table)，通过查询将每个字符 x_i 将转化成固定长度的低维实数向量。本文采用通过 KWSE 模型预先训练好的分写增强字符向量作为 Lookup Table 的初始值，训练过程中将 Lookup Table 当做可训练参数，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，为了更好地捕获字符前后上下文信息，将 $LX(X)$ 输入到双向 LSTM 网络中。假设输入字符 x_i 经过前向 LSTM 的 Cell 后的输出结果为 $\vec{h}_i \in \mathbb{R}^d$ ，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 $h_i \in \mathbb{R}^{d \times 2}$ 。模型中 h_i 表示考虑了第 *i* 个字符前后上下文的基础上，输入字符 x_i 的编码。

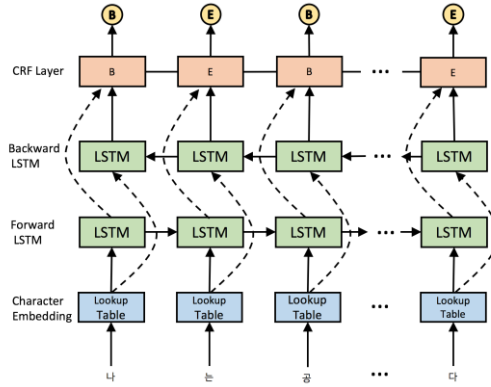


图3 基于 KWSE-LSTM-CRF 的朝鲜语分写模型

模型的最后一层通过 CRF 预测全局最优的分写标注序列，计算公式如下：

$$s_{char}(i) = f(W_{out} h_i + b_{out}) \quad (12)$$

$$s(X, Y', \theta) = \sum_{i=1}^n \left(A_{y'_{i-1}, y'_i} + s_{char}(i) \right) \quad (13)$$

其中 $s_{char}(i)$ 表示输入字符 x_i 经过双向 LSTM 网络得到的分写标注的概率分布， W_{out} 和 b_{out} 为全连接层的映射矩阵及偏置向量， f 为 softmax 函数。 A 表示分写标注状态的转移矩阵，例如 $A_{y'_{i-1}, y'_i}$ 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 预测的分写标注概率时，输入字符序列 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 $s(X, Y', \theta)$ 最大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。

4 实验

4.1 实验数据集

在训练模型之前，首先针对原始语料库进行了预处理，其过程如下：

1. 朝鲜语句子的 **tokenize**：根据语料库中的句子生成字符序列，同时用<NUM>（表示数字），<FOREIGN>（表示其他语言字符）等 **token** 替换掉非朝鲜语字符。另外，根据语料库中包含的分写信息生成每一句对应的分写标注序列。
2. 生成字典：按照字符频率从高到低进行排序，取前 1000 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，为了训练固定步长的 LSTM 网络，将每一条句子的长度截断为 40 个字符，小于 40 个字符的句子用<UNK>补齐。

本文采用了 HANTEC-2.0 语料库，其中包含社会科学、自然科学、一般综合等三大分类的共 12000 篇文档。该语料库的文档均为带标签的结构化数据，因此首先抽取了带<TEXT>标签的正文部分，之后经过分句，过滤掉朝鲜语字符比例低于 50%的无效句，最终得到 197723 条句子构成的原始语料库。为了跟过往研究方法做对比实验，参与实验的所有朝鲜语分写模型均采用如下的数据集结构。

表 2 数据集结构

	训练集	测试集
句子数	173723	19303
单词数	3267856	362252
字符数	15127884	1680176
非分写字符百分比	79.55%	79.59%
分写字符百分比	20.45%	20.41%

4.2 实验设置

实验中采用了 tensorflow1.0 框架，并用 NVIDIA 的 1080GPU 进行了加速。

具体的模型参数配置如下：

1. 获取朝鲜语分写增强字符向量：将训练集和测试集中 193026 条句子及对应的分写标注作为输入，训练朝鲜语分写增强字符向量。朝鲜语单词大部分都由 4 个以内的字符构成，因此模型中参数 c 设置为 2，即目标字符的左右各取 2 各字符作为上下文窗口， $batch\ size$ 设置为 128，学习率为 0.001，字符向量的维度 d 设置为 128。朝鲜语分写标注法采用了四位标注法，即输出向量 y_t 的大小为 4。模型通过随机梯度下降法进行优化，经过 50 个 epoch 的训练，最终得到大小为 1000×128 的朝鲜语分写增强字符向量。

2. 训练 Bi-LSTM-CRF 模型：模型的查询表初始化为上述第一步中预训练得到的字符向量。其他参数均采用均匀分布的随机函数初始化成较小的实数。模型中双向 LSTM 网络的输入是大小为 $(128 \times 40 \times 128)$ 的张量，其中第一维代表 $batch\ size$ ，第二维表示 LSTM 网络的长度，第三维表示字符向量的大小。LSTM 网络的输出部分将生成 $(128 \times 40 \times 256)$ 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 $(128 \times 40 \times 4)$ 的张量，其中 W_{out} 大小为 256×4 ， b_{out} 大小则是 4。

4.3 实验结果及分析

本文采用了如下四种评估指标：字符级准确率(P_{char})、单词级准确率(P_{word})、单词级召回率(R_{word})、单词级 F1 值($F1_{word}$)。

$$P_{char} = \frac{\text{正确标注的字符数}}{\text{测试集中的字符数}}$$

$$R_{word} = \frac{\text{正确分写的单词数}}{\text{测试集中的单词数}}$$

$$P_{word} = \frac{\text{正确分写的单词数}}{\text{模型生成的单词数}}$$

$$F1_{word} = \frac{2 \times P_{word} \times R_{word}}{(P_{word} + R_{word})}$$

实验中复现了相关研究中的几种典型的朝鲜语分写模型，分别是 Lee, Rim 2007 年提出的基于 HMM 的分写模型、Shim 等人 2011 年提出的基于 CRF 的模型、Lee 和 Kim(2013) 等人提出的基于 Structured SVM 的分写模型。另外，为了验证本文提出的朝鲜语分写增强字符向量的有效性，实现了三种模型：随机初始化的查询表+LSTM+CRF(LSTM-CRF)，用 CBOW 模型预训练的字符向量+LSTM+CRF(CBOW-LSTM-CRF)，朝鲜语分写增强字符向量+LSTM+CRF(KWSE-LSTM-CRF)。表 3 给出了各个模型的实验结果。

表 3 实验结果

模型	P _{char}	P _{word}	R _{word}	F1 _{word}
HMM	93.25	88.45	87.86	88.15
CRF	93.97	91.19	89.08	90.12
Structured SVM	94.05	89.73	87.31	88.5
LSTM-CRF	93.13	91.16	90.64	90.90
CBOW-LSTM-CRF	93.29	91.23	90.87	91.05
KWSE-LSTM-CRF	94.72	93.15	92.57	92.86

表中可以看到我们提出的 KWSE-LSTM-CRF 方法在字符级准确率、单词级准确率、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 CRF 方法提高了 2.74 个百分点。

与传统的 HMM, CRF, Structured SVM 方法相比 LSTM-CRF 系列模型在字符级准确率上较为接近。然而字符级准确率有一定的片面性：如果一条句子中关键的几个分写标注有误（例如：朝鲜语中的助词分写은,는等），即使字符级准确率再高，整句的分写效果也会大打折扣。因此我们更关注单词级的准确率和召回率。后三种采用 LSTM-CRF 模型的分写方法在这两项指标上均优于其他几种模型。

在同样采用 LSTM-CRF 模型的后三种方法之间的对比中可以看到：随机初始化字符查询表的方法与 CBOW 预训练的方法相差无几，说明在朝鲜语分写标注问题上 CBOW 预训练的字符向量虽然富含字符的语义信息，但是对于分写标注帮助不大。而我们提出的分写增强字符向量由于提前融入了朝鲜语分写倾向性信息，因而有利于 LSTM-CRF 模型做出更为准确的分写预测。

结论

本文提出了一种朝鲜语分写增强字向量模型 KWSE，并将 KWSE 预训得到的字向量应用于基于 LSTM-CRF 的朝鲜语分写模型中。实验结果表明本文提出的 KWSE-LSTM-CRF 方法的字符级分写准确率优于 HMM、CRF、Structured-SVM。单词级准确率上，相比于现有最好的 CRF 模型，本文提出的 KWSE-LSTM-CRF 方法将 F1 值提高了 2.74 个百分点。未来工作中我们希望统计和分析高频分写错误，并分析出与之关联的特定字符或单词，从这方面入手，进一步提高模型的性能。

参考文献

- [1] Kim K S, Lee H J, Lee S J. Three-stage spacing system for Korean in sentence with no word boundaries[J]. Journal of the Korea Information Science Society, 1998, 25(12): 1838-1844.
- [2] Kang S S. Eojeol-block bidirectional algorithm for automatic word spacing of Hangeul sentences[J]. Journal of KIISE: Software and Applications, 2000, 27(4): 441-447.
- [3] Chung Y M, Lee J Y. Automatic word-segmentation flatline-breaks for Korean text

- processing[C]//Proceedings of the 6th Conference of Korea Society for Information Mangement. 1999: 21-24.
- [4] Jeon N Y, Park H R. Automatic word-spacing of syllable bi-gram information for Korean OCR postprocessing[C]//Annual Conference on Human and Language Technology. Human and Language Technology, 2000.
- [5] Kang S S, Woo C W. Automatic Segmentation of Words using Syllable Bigram Statistics[C]//NLPRS. 2001: 729-732.
- [6] Lee D G, Rim H C, Yook D. Automatic word spacing using probabilistic models based on character n-grams[J]. IEEE Intelligent Systems, 2007, 22(1), 28-35.
- [7] Shim K S. Automatic word spacing based on Conditional Random Fields[J]. Korean Journal of Cognitive Science, 2011, 22(2): 217-233
- [8] Lee C, Kim H. Automatic Korean word spacing using Pegasos algorithm[J]. Information Processing & Management, 2013, 49(1): 370-379.
- [9] Lee C, Choi E, Kim H. Balanced Korean Word Spacing with Structural SVM[C]//EMNLP. 2014: 875-879.
- [10] 황현선, 이창기. 딥러닝을 이용한 한국어 자동 띄어쓰기[J]. 한국정보과학회 학술발표논문집, 2016: 738-740.
- [11] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [12] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.
- [13] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Interspeech. 2010, 2: 3.
- [14] Hochreiter S, Schmidhuber J. LSTM can solve hard time lag problems[C]//Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference. 1997: 473-479.
- [15] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

金国哲，通讯作者，讲师，主要研究领域为自然语言处理
作者联系方式：吉林省延吉市延边大学工学院计算机系 244 信箱 133002(邮编)
电话：13514330983
E-mail: 34200519@qq.com
