基于子词级别词向量和指针网络的朝鲜语句子排序

闫晓东
中央民族大学
信息工程学院
国家语言资源监测
与少数民族语言中心
yanxd3244@sina.com

解晓庆
中央民族大学
信息工程学院
国家语言资源监测
与少数民族语言中心
xqplex@yeah.net

摘要

句子排序是多文档摘要系统和机器阅读理解中重要的任务之一，排序的质量将直接影响摘要和答案的连贯性与可读性。因此，本文采用在中英文上大规模使用的深度学习方法，同时结合朝鲜语词语形态变化丰富的特点，提出了一种基于子词级别词向量和指针网络的朝鲜语句子排序模型，其目的是解决传统方法无法挖掘深层语义信息问题。本文提出基于形态素拆分的词向量训练方法(MorV)，同时对比子词n元词向量训练方法(SG)，得到朝鲜语词向量；采用了两种句向量方法:基于卷积神经网络(CNN)、基于长短时记忆网络(LSTM)，结合指针网络分别进行实验。结果表明本文采用MorV和LSTM的句向量结合方法可以更好地捕获句子间的语义逻辑关系，提升句子排序的效果。

关键词： 词向量；形态素拆分；指针网络；句子排序

Korean Sentence Ordering Based on Sub Word Level Word Vector and Pointer Network

Xiaodong Yan
Minzu University of China
National language resource
monitoring &Research Center
Minority Languages Branch
yanxd3244@sina.com

Xiaoqing Xie
Minzu University of China
National language resource
monitoring &Research Center
Minority Languages Branch
xqplex@yeah.net

Abstract

Sentence sorting is one of the most important tasks in multi document summarization system and machine reading comprehension. The quality of sorting will directly affect the coherence and readability of abstracts and answers. Therefore, this paper adopts the deep learning method which is widely used in both Chinese and English, combined with the characteristics of the rich morphological changes of Korean words, puts forward a Korean sentence ordering model based on the sub word level word vector and pointer network, the purpose of which is to solve the problem that traditional methods can not mine deep semantic information. In this paper, a morpheme split based word vector training method (morv) is proposed, and the Korean word vector is obtained by

第十九届中国计算语言学大会论文集，第468页-第479页，海南，中国，2020年10月30日至11月1日。

comparing the sub word n-ary word vector training method (SG). Two sentence vector methods are used: convolution neural network (CNN) and long-term memory network (LSTM), combined with pointer network. The results show that the combination of morv and LSTM can better capture the semantic logic relationship between sentences and improve the effect of sentence ordering.

Keywords: Word vector , Morpheme split , Pointer network , Senternce ordering

1 引言

句子排序是多文档自动摘要任务和阅读理解答案融合的关键技术。在多文档自动摘要任务中，对文摘句子进行排序是一项关键任务，其效果直接影响最后生成的摘要的可读性。在阅读理解的答案排序过程中，也涉及到句子排序问题，其最终结果也会决定答案的可读性。

朝鲜语是我国具有文字的少数民族语言之一，在朝鲜语信息化处理的过程中(Bi, 2011)，同样也有多文档自动摘要和阅读理解答案融合任务。因此朝鲜语句子排序也是一个值得关注的问题。本文结合朝鲜语的特点，提出了基于子词级别词向量的朝鲜语句子排序模型，可以增强句子语义逻辑关系的捕获能力，进而获取句子的合理排序。为后续的朝鲜语多文档自动摘要、朝鲜语机器阅读理解等任务提供一些基础。

通常，在一个文本段落中，语义的连贯性是通过句子的顺序来保证的。对于句子排序问题，前人已经做了大量的工作：徐永东提出了一种多文档摘要中基于时间信息的句子排序方法，利用基于规则的时间信息抽取、语义计算及时序推理方法来解决句子排序问题(Xu et al., 2009)；姚超提出了一种基于内聚度的多文档文摘的句子排序方法，通过将相同话题的句子聚合到一起，避免话题中断，改善文摘可读性(Yao et al., 2006)；薛涛将条件熵引入到句子排序工作中，通过在源文档中计算句子对的转移信息量来衡量句子的关联程度，同时提出上下文对比算法来加强句子邻近度学习的准确性(Xue and Wang, 2017)；郭红建将潜在语义分析聚类算法引入文摘句子排序过程中，将话题聚类之后采用模板对文摘句子进行两趟排序(Guo and Huang, 2013)……

但是，随着大数据、云计算等技术的发展，深度学习方法在自然语言处理任务中广泛应用，很多深度学习方法被引入到句子排序中。康世泽利用神经网络将几种前人提出的句子排序方法融合，并在此基础上提出了一种基于马尔科夫随机游走模型的句子排序算法(Kang et al., 2016)。Chen尝试了基于卷积神经网络(convolutional neural netw-orks, CNNs)，长短期记忆网络(long short-term memory network, LSTM)的句子排序方法，使用CNN、LSTM等模型判断句子的前后句关系，并利用集束搜索算法求解句子的最优排序(Chen et al., 2016)。Logeswaran提出了一种基于循环神经网络的句子排序方法，通过判断句子在每个位置的可能性，求得最优排序结果(Logeswaran et al., 2016)。Gong提出了一种基于端到端的指针网络的句子排序方法，通过端到端的指针网络判断每个位置上的句子的可能性，求得较优排序结果(Gong et al., 2016)。

本文的主要贡献如下：

1）对朝鲜语句子排序问题进行研究；

2）将同形异义词信息融入到朝鲜语词向量的训练；

3）使用形态素和子词级别n元进行词向量训练，并对比效果；

4）使用两种词向量训练方法得到词向量，再使用两种不同的句向量训练方法得到句向量，最后进行句子排序实验，并对比效果。

2 朝鲜语句子排序模型

2.1 任务描述

在机器阅读理解的答案融合任务和多文档自动文摘任务中，候选的句子集是从不同的文档中抽取的，因此，无法根据句子在原文中的位置或者一些显式的连接词对乱序的句子集合进行排序。句子排序任务要解决的问题就是把一组乱序的句子，排列成连贯、通顺的段落。设给定一组乱序的句子$S = s_1, s_2, \cdots, s_n$，句子排序的任务目标是将其排列成顺序$o^*$，对于顺序$o^*$有：

$$s_{o_1^*} > s_{o_2^*} > \cdots > s_{o_n^*} \tag{1}$$

在给定句子集S的情况下，顺序o^*的概率$P(o^*|S)$大于其他任何顺序的概率，可以表示为式(2)。其中o表示句子集S的任一种排序，而Ψ表示句子集S的所有可能的排序的集合。

$$P(o^*|S) > P(o|S), \forall o \in \Psi \tag{2}$$

2.2 模型架构

我们采用指针网络模型(Pointer Network)(Vinyals et al., 2015)对句子集S进行排序。指针网络(Pointer Network)是Nallapati等(Nallapati et al., 2016)提出的基于注意力机制的序列到序列模型的一个变种。它不是把一个序列转换成另一个序列，而是产生一系列指向输入序列元素的指针。最基础的用法是对可变长度序列或集合的元素进行排序，也适用于句子排序问题。

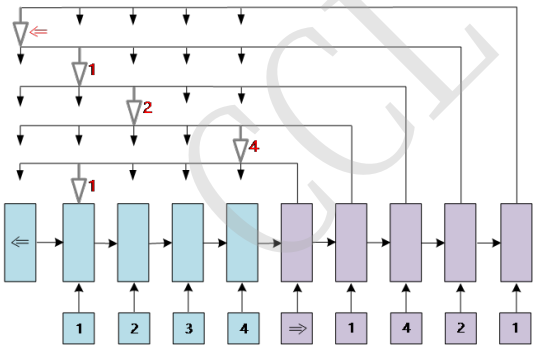

图 1: 指针网络模型结构

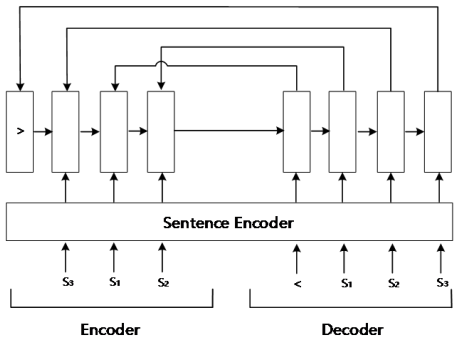

图 2: 基于指针网络的句子排序模型

指针网络模型非常简洁如图1所示，结构是基本的seq2seq+attention。基于指针网络的句子排序模型的架构如图2所示。以顺序o为集合S排序的概率$P(o|S)$的计算公式为(3)。

$$P(o|s) = \prod_{i=1}^{n} P = (o_i|o_{i-1}, \cdots, o_1, s) \tag{3}$$

概率$P = (o_i|o_{i-1}, \cdots, o_1, s)$可以通过指针网络计算，为式(5)，(6)，其中$e_j$，$d_i$分别是指针网络编码端和解码端的输出。

$$P = (o_i|o_{i-1}, \cdots, o_1, s) = softmaxt(u^i)_{o_i} \tag{4}$$

$$u_j^i = v^T \tanh(W_1 e_j + W_2 d_i) \tag{5}$$

2.2.1 编码端

指针网络的编码器模型可以表示为式(6)，其中，$Enc(s_{o_j})$表示句子so_j的编码。

$$e_j = LSTM(Enc(s_{o_j}, e_{j-1}), j = (1, \cdots, n) \tag{6}$$

2.2.2 解码端

指针网络的解码器模型可以表示为式(7)，其中，$Enc(s_{o_i})$表示句子s_{o_i}的编码。

$$d_i = LSTM(Enc(s_{o_i}, d_{i-1}), i = (1, \cdots, n) \tag{7}$$

2.3 句子顺序概率

我们将句子集的顺序表示为：$P(o|s)$，将最佳句子顺序表示为$\hat{o}$:

$$\hat{o} = \underset{o}{argmax}\, P(o|s) \tag{8}$$

找到句子集s的最佳顺序是一个NP问题，有两种策略可以用来解决这个问题：贪心算法和集束搜索算法。

2.3.1 贪心算法

贪心算法（Greedy Algorithm）的思想是指，在对问题求解时，总是做出在当前看来是最好的选择，也就是说，不从整体最优上加以考虑，它所做出的选择是在某种意义上的局部最优解。在指针网络的解码阶段，用贪心算法表示顺序$\hat{o} = \hat{o}_1, \cdots, \hat{o}_n$的生成过程可以表示为式(9)。

$$\hat{o}_i = \underset{o_i}{argmax}\, P(o_i|\hat{o}_{i-1}, \cdots, \hat{o}_1, s) \tag{9}$$

2.3.2 集束搜索算法

集束搜索(Beam Search)是一种启发式图搜索算法，通常用在图的解空间比较大的情况下，为了减少搜索所占用的空间和时间，在每一步深度扩展的时候，剪掉一些质量比较差的结点，保留下一些质量较高的结点。这样减少了空间消耗，并提高了时间效率。在求解最优解时，集束搜索算法的每一步总是保留最优的b个候选项。对于第t步来说，每个候选解可以表示为$\hat{o}_1^t = \hat{o}_1, \cdots, \hat{o}_t$，其概率为式(10)。其中概率最靠前的$b$个候选项将会在第$t$步被保留。

$$P(\hat{o}_1^t|s) = \prod_{i=1}^{t} P(\hat{o}_i|\hat{o}_{i-1}, \cdots, \hat{o}_1, s) \tag{10}$$

3 模型训练

3.1 词向量训练

在自然语言处理的发展过程中，单词的分布式表示不断发展。世界各国的研究学者提出了许多模型。这些模型大多数应用于英语，通过不同的向量来表示词汇表中的每个单词，但会忽略单词的内部结构的变化。不同于英语，对于形态丰富的语言，例如朝鲜语，很多词语在训练语料库中很少出现（或根本没有出现），这使得学得的词向量语义捕获能力差。

朝鲜语句子由多个语节构成，而每个语节(eojeol)由一个或多个形态素组成。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3的句子中共有5个语节，其中每个语节由一个或多个形态素构成，图中以“+”作为形态素的分隔符。若仅仅通过语节来训练词向量，那么由于朝鲜语的词尾形态变化丰富，使得训练得到的词向量的语义表示能力不足。为了解决这一问题，本文将采取以下两种朝鲜语的词向量训练方法：1）先将语节拆分成多个形态素（变换原形）的组成形式，再对拆分好的形态素进行词向量训练；2）以朝鲜语子词（音节和字母）为单位，用skip-gram模型训练词向量。上述两种方法都考虑了朝鲜语的形态信息，训练得到的词向量语义表达能力更强。

3.1.1 形态素词向量(Morpheme Vector, MorV)

在朝鲜语中，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）。可以看到，在形态素分析过程中，“했”形态素转化为“하”，“었”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，存在处理不好的未登录词等问题。针对这一问题，本文采用结合词性信息的多任务seq2seq模型来解决这一问题。在朝鲜语中，由空格分开的单元是语节。由于朝鲜语所有的语节数量非常庞大，实验中用到的语料中有624,655个不同的语节，不太适合直接作为seq2seq模型的输入。本文考虑将一个语节看作一个音节序列。例如语节‘개발했다’是4个音节‘개’，‘발’，‘했’，‘다’组成，可以看作一个音节序列。同样的，形态素也是由音节组成的，也可以看作一个音节序列。本文中实验用到的语料中的音节数量为5,245个，远小于语节的数量。因此，文本以音节为单元作为seq2seq模型的输入，进行朝鲜语形态素拆分模型训练。

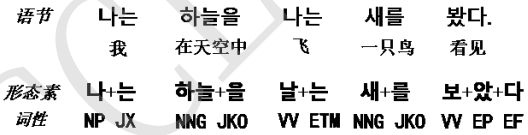

图 3: 朝鲜语句子中的语节和形态素

由于朝鲜语的形态素和词性息息相关(Song and Park, 2019)，例如语节“하늘을（在天空中）”可以被分为两个形态素“하늘（名词）”和“을（宾格助词）”，通过词性就可以把这个语节拆分成两个形态素。本文利用21世纪世宗计划语料库，在训练模型的过程中加入词性信息，可以更准确地进行形态素的拆分。此外，朝鲜语中还存在同形异义词。例如，在图2中语节“나는”出现了两次，意思截然不同，但通过词性可以识别出这是两个不同的词语。本文通过词性的不同将表达两种意思的“나는”，分别记为“나는__01”，“나는__02”，如果还有其他意思就顺次编号：“나는__03”，“나는__04”……这样可以把同形异义词当作不同的词来进行训练。用通过这种方法得到的形态素来训练词向量，可以得到语义表示能力更强的词向量。模型的示意图如图4所示。

本文使用的seq2seq模型是基于2018年Anastasopoulos和Chiang(Anastasopoulos and Chiang, 2018)提出的triangle多任务学习模型，将相互关联的任务放在同一个神经网络中同时训练，可以有效提升训练效果。本文将词性标记和形态素拆分这两个任务同时训练，并将同形异义词

进行编号。通过此模型拆分得到的形态素将更加准确且考虑同形异义信息。

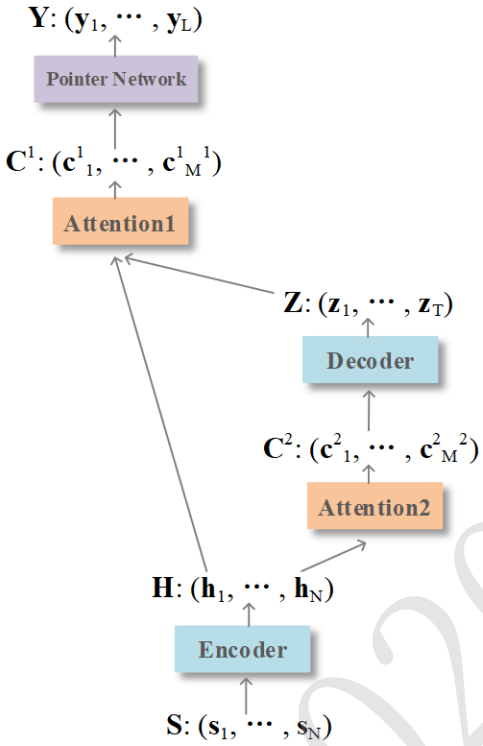

图 4: 朝鲜语句子中的语节和形态素

模型有4个部分：编码模块、解码模块、注意力机制模块1、注意力机制模块2、指针网络模块(Nallapati et al., 2016)。图4所示的是模型的整体框架。编码模块(Encoder)的作用是将输入的音节序列S ：$(s_1, \cdots, s_N)$转化成隐藏状态序列H ：$(h_1, \cdots, h_N)$；注意力机制模块1(Attention1)的作用是将隐藏状态序列H：$(h_1, \cdots, h_N)$转化成考虑上下文信息的隐藏状态序列C^1：$(c_1^1, \cdots, c_{M^1}^1)$，并输入给解码模块(Decoder)；解码模块(Decoder)的作用是将隐藏状态序列C^1：$(c_1^1, \cdots, c_{M^1}^1)$转化成词性标记序列$Z$：$(z_1, \cdots, z_T)$；注意力机制模块2(Attention2)的作用是考虑词性信息的同时，将隐藏状态序列H：$(h_1, \cdots, h_N)$转化成考虑上下文信息的隐藏状态序列C^2：$(c_1^2, \cdots, c_{M^2}^2)$，并输入到指针网络(Pointer Network)；指针网络模块(Pointer Network)的作用是通过softmax函数形成指针，从输入音节序列或给定的音节表中选择音节（或空格），生成形态素序列Y：$(y_1, \cdots, y_L)$。给定输入序列S：$(s_1, \cdots, s_N)$，S表示的是输入的语节，将语节拆分成音节组成的序列，其中s_i表示的是第i个音节，输出序列是Y：$(y_1, \cdots, y_L)$，Y表示的是拆分好的形态素序列，其中Y_i表示的是第i个形态素。

用训练好的模型进行形态素原形转换拆分，朝鲜语的最小单位是形态素。通过形态素原形转换，去除了朝鲜语词尾形态变化丰富这一干扰因素。采用Word2vec进行形态素向量（即词向量）训练，得到关于形态素的分布式表示，具有较强的语义表示能力。

3.1.2 融入子词级别信息(Subword Gram, SG)

形态素拆分过程比较复杂，容易出现错误，提出了基于字母和音节的词向量表示方法(Park et al., 2018)。将一个语节拆分成字母序列，再进行音节级别和字母级别的n元划分。

音节拆分规则:

每个朝鲜语音节可拆分成由3个字母组成的序列，例如“달”可拆分成{ㄷ,ㅏ,ㄹ}。如果有的音节只有两个字母组成，那么就用一个占位符“e”代替第三个字母，例如“해”拆分成{ㅎ,ㅐ,e}。使用“<”作为语节的开始标志，“>”作为语节的结束标志，这样语节“강아지”可以拆分成字母序列{<,ㄱ,ㅏ,ㅇ,ㅇ,ㅏ,e,ㅈ,ㅣ,e,>}。

音节级别的n元划分：

语节“강아지”的一元划分可以表示为：{ㄱ,ㅏ,ㅇ}, {ㅇ,ㅏ,e}, {ㅈ,ㅣ,e}；二元划分可以表示为：{ㄱ,ㅏ,ㅇ,ㅇ,ㅏ,e}, {ㅇ,ㅏ,e,ㅈ,ㅣ,e}；三元划分可以表示为：{ㄱ,ㅏ,ㅇ,ㅇ,ㅏ,e, ㅈ,ㅣ,e}。

字母级别的n元划分：

由于朝鲜语的粘着性，只考虑音节级别的n元，无法捕捉到形态变化信息，因此还需要考虑字母级别。

关于语节“강아지”，字母级别的三元划分可以表示为：{<,ㄱ,ㅏ}, {ㄱ,ㅏ,ㅇ}, {ㅏ,ㅇ,ㅇ}, {ㅇ,ㅇ,ㅏ}, {ㅇ,ㅏ,e}, {ㅏ,e,ㅈ}, {e,ㅈ,ㅣ},{ㅈ,ㅣ,e}, {ㅣ,e,>}。

然后用这两个级别的n元，通过skip-gram方法(Bojanowski et al., 2017; Mikolov et al., 2013)进行词向量训练。我们使用该方法与上文提出的方法均用来训练词向量，将训练得到的词向量再进行下一步处理，最后观察句子排序的结果。

3.2 句向量表示

句向量又可以称为句嵌入(Cer et al., 2018)，句嵌入模型的输入为词向量，输出为表示句子的向量，该向量可以作为具体任务的输入进行预测和训练。自然语言处理的任务大多数都是序列化的信息，序列化信息的特点就是不同时间步上的信息会有交叉作用，如何发掘序列化输入的信息是自然语言处理任务的关键。在当前研究成果中，主要分为两大解决方法：一是以循环神经网络为基础的解决方案；二是以卷积神经网络为基础的解决方案。本文将采用这两种方案对句子进行向量化，并对比不同的句向量训练方法对句子排序结果的影响。

3.2.1 卷积神经网络模型

卷积神经网络(Convolutional neural networks, CNN)(Simard et al., 2003)仿造生物的视知觉机制，包含卷积计算且具有深度结构的前馈神经网络，是深度学习的代表算法之一。将包含n_w个单词的句子s通过卷积神经网络编码的过程可以表示为公式(11)，(12)。其中$W_{cov} \in R^{(dl_f)d_f}$和$b_{cov} \in R^{d_f}$是可训练的参数，其中$\phi(\cdot)$是tanh函数。$k = 1, \cdots, n_w - l_f + 1$。其中的$l_f$和$d_f$都是卷积神经网络模型中的超参数，分别是过滤器(filter)的长度和特征图(feature map)的个数。

$$cov_k = \phi(W_{cov}^T(\oplus_{u=0}^{l_f-1} w_{k+u}) + b_{cov}) \tag{11}$$

$$Enc(s) = \max_k cov_k \tag{12}$$

3.2.2 长短时记忆网络模型

长短时记忆网络(Long short term memory, LSTM)(Hochreiter and Schmidhuber, 1997)是一种特殊的RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。LSTM的存储单元$c \in R^{d_r}$由三种门控制：输入门$i \in R^{d_r}$、遗忘门$f \in R^{d_r}$输出门$o \in R^{d_r}$，表示为公式(13)-(15)。其中，$W_g \in R^{(d+d_r)4d_r}$和$b_g \in R^{4d_r}$是可训练的参数，d_r是表示存储单元和门控单元的维

度的一个超参数。$t = 1, \cdots, n_w$其中$\sigma(\cdot)$是sigmoid函数，$\phi(\cdot)$是tanh函数。

$$\begin{bmatrix} i_t \\ o_t \\ f_t \\ \widetilde{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \phi \end{bmatrix} (W_g^T \begin{bmatrix} w_t \\ h_{t-1} \end{bmatrix} + b_g) \tag{13}$$

$$c_t = c_{t-1} \odot f_t + \widetilde{c}_t \odot i_t \tag{14}$$

$$h_t = o_t \odot \phi(c_t) \tag{15}$$

我们将通过长短时记忆网络编码的句子向量表示为：

$$Enc(s) = h_{n_w} \tag{16}$$

3.3 目标函数训练

设有m个训练样本$(x_i, y_i)_{i=1}^m$，x_i表示的是一个句子集合，这个句子集合有一个唯一特定的排序序列y_i，y_i的句子顺序是最优顺序o^*。为了得到更多的训练数据，本文在训练模型的过程时，在每个epoch中为句子集合x_i随机生成新的排序。目标函数可以表示为公式(17)。其中，$P(y_i|x_i;\theta) = P(o^*|S = x_i;\theta)$，$\lambda$是正则项的超参数。$\theta$表示所有可训练的参数。此外，本文采用AdaGrad(Duchi et al., 2011)结合小批量梯度下降(Turian et al., 2010)优化算法来训练模型。

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} logP(y_i|x_i;\theta) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{17}$$

4 实验

4.1 数据集

本文从延边日报朝鲜语版、人民网朝鲜语版等新闻网站爬取了20000篇朝鲜语新闻作为语料。将每篇新闻进行语段分隔，选取句子数目大于2的语段作为一个数据单元，将每个数据单元的句子进行打乱编号。例如将语段[s1，s2，s3，s4]编码为[4,1,2,3]，然后再对该语段编码随机打乱为[3,2,4,1]。这样我们就得到一个训练样本([句1,句2,句3,句4]，[4,1,2,3]，[3,2,4,1])，第一项为顺序句子集合，第二项为正确顺序，第三项为乱序顺序。按照上述形式对所有数据单元进行编码再打乱，得到样本集合。对这些样本集合进行训练集、验证集和测试集的划分。划分结果如表1所示：

Models	PM	LSR	PMR
新闻类型	训练集	验证集	测试集
经济	19,223	2,465	2,497
政治	15,495	1,943	1,866
科技	821,795	102,584	102,892
体育	84,689	10,624	10,453
教育	13,273	1,619	1,695
娱乐	5,201	708	670
法律	216,153	26,819	26,854

表 1: 实验所用语料

Initial learning rate	$\alpha = 0.5$
Regularization	$\lambda = 10^{-5}$
Hidden layer size of Ptr-Net	h=200
Filter length of CNN	l_f=3,4,5
Number of features maps	d_f=128
Hidden size of LSTM	d_r=200
Size of embedding	d_e=100
Beam size	b=64
Batch size	128

表 2: 超参数设置

4.2 超参数设置

表2展示了上述模型中的超参数的设置。卷积神经网络的句子编码模型使用了3种不同长度l_f的过滤器(Kim et al., 2016)。

4.3 评测方法

本文采用了3中不同的模型评测方法：(1)成对度量法；(2)最长序列比法；(3)最佳匹配比法。

4.3.1 成对度量法

成对度量法(Pairwise metrics, PM)指的是，预测的相对顺序与原本真正顺序相同的句子对的分数越高越好。成对度量法可以表示为三个量化分数：精确值P、召回率R和F值，如公式(18)-(20)所示。其中，函数$S(\cdot)$表示一段文本中所有句子对的集合，绝对值符号表示的是集合的大小。

$$P = \frac{1}{m}\sum_{i=1} m\frac{|S(\hat{o}_i) \cap S(o_i^*)|}{|S(\hat{o}_i)|} \tag{18}$$

$$R = \frac{1}{m}\sum_{i=1} m\frac{|S(\hat{o}_i) \cap S(o_i^*)|}{|S(\hat{o}_i^*)|} \tag{19}$$

$$F = \frac{2 * P * R}{P + R} \tag{20}$$

设$\{\hat{o} = (2,3,1,4), o^* = (1,3,4)\}$，其中第二句话是一个噪声项。对于这个例子，成对度量分数可以表示为：$P = 1/6, R = 1/3, F = 2/9$。

4.3.2 最长序列比法

最长序列比法(Longest sequence ratio, LSR)计算最长正确子序列的比（不需要连续性，越高越好）。最长序列比法可以表示为三个分数：精确值P、召回率R和F值，如公式(21)-(23)所示。其中，函数$L(\cdot)$表示的是最长正确子序列中元素的个数。那么，$L(\hat{o} = (2,3,1,4), o^* = (1,3,4))$的值就是2。

$$P = \frac{1}{m}\sum_{i=1}^{m} \frac{|L(\hat{o}_i, o_i^*)|}{|\hat{o}_i|} \tag{21}$$

$$P = \frac{1}{m}\sum_{i=1}^{m}\frac{|L(\hat{o}_i, o_i^*)|}{|o_i^*|} \tag{22}$$

$$F = \frac{2 * P * R}{P + R} \tag{23}$$

4.3.3 最佳匹配比法

最佳匹配比法(Perfect match ratio, PMR)计算的是确切的匹配项的比例，如公式(24)，(25)所示。其中，$P(\cdot)$表示$\hat{o}_i$和o_i^*的最佳匹配子序列的长度。

$$PMR = \frac{1}{m}\sum_{i=1}^{m}1\{\hat{o}_i = o_i^*\} \tag{24}$$

$$\{\hat{o}_i = o_i^*\} = \frac{P(\hat{o}_i \cap o_i^*)}{\hat{o}_i} \tag{25}$$

4.4 实验结果和分析

我们用两种不同的词向量训练方法，两种不同的句向量训练方法对句子进行编码，然后通过指针网络进行句子排序，在进行句子排序的过程中，使用两种不同的搜索算法：贪心算法和集束搜索算法，结果分别用三种评测方法进行评测。结果如下表所示： 根据表3我们可以看出，

Models	PM	LSR	PMR
+greedy algorithm			
MorV+CNN	80.21	74.33	39.12
MorV+LSTM	84.02	78.25	43.68
SG+CNN	78.35	73.28	37.21
SG+LSTM	81.37	77.92	41.69
+beam search			
MorV+CNN	80.68	76.87	40.36
MorV+LSTM	**85.13**	**79.20**	**44.32**
SG+CNN	79.28	76.97	37.49
SG+LSTM	82.63	78.51	43.56

表 3: 不同方法的句子排序结果对比

使用本文提出的形态素拆分模型(MorV)将语节拆分成形态素，再进行词向量训练，在三种评测方法下，可以使得朝鲜语句子排序效果更好。使用LSTM进行句子编码相对于CNN，句子排序效果更好。增加集束搜索(beam search)过程后，句子排序的效果也有所提升。 从图5中也可以直观得出结论：使用MorV词向量训练模型+LSTM句编码模型，句子排序效果最佳。表4给出的是句子排序的实例

5 总结

句子排序是近年来自然语言处理中多文档摘要生成和机器阅读理解答案融合任务中的一个十分重要子任务。以往的研究主要是基于传统的机器学习方法，但随着深度学习方法的不断发展，句子排序任务也可以使用一些深度学习方法来解决。

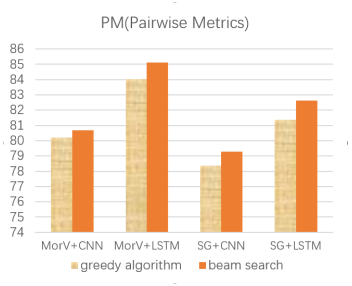

图 5: PM评测结果

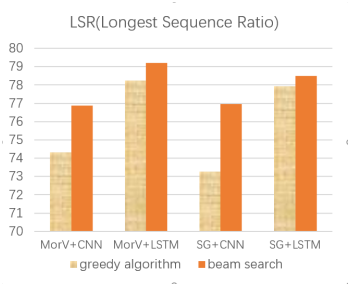

图 6: LSR评测结果

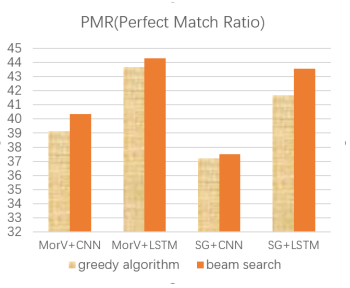

图 7: PMR评测结果

正确顺序语段	도로시는 1910 년에 당시 영국의 식민지였던 이집트의 까히라에서 태여났다. 1928 년에 옥스퍼드대학의 화학과에 진학해 생화학을 전공했다. 그녀는 1932 년에 대학을 졸업했지만 직장을 구할 수 없었다. 도로시는 소개를 통해 케임브리지대학의 화학과 교수를 알게 됐고 그 교수는 다시 버널 밑에서 공부하도록 주선해 주었다. 버널은 엑스선을 리용해 단백질을 비롯한 생물학적 결정을 연구하고 있었다. 。
顺序编码	3，2，5，4，1
乱序编码	4，5，1，2，3
排序结果	3，2，5，1，4
排序结果对应语段	도로시는 1910 년에 당시 영국의 식민지였던 이집트의 까히라에서 태여났다. 1928 년에 옥스퍼드대학의 화학과에 진학해 생화학을 전공했다. 그녀는 1932 년에 대학을 졸업했지만 직장을 구할 수 없었다. 버널은 엑스선을 리용해 단백질을 비롯한 생물학적 결정을 연구하고 있었다. 도로시는 소개를 통해 케임브리지대학의 화학과 교수를 알게 됐고 그 교수는 다시 버널 밑에서 공부하도록 주선해 주었다. 。

表 4: 句子排序示例

在朝鲜语信息化进程中，也需要跟上深度学习发展的步伐。本文将深度学习模型用于朝鲜语信息化处理，使用多任务seq2seq模型进行形态素拆分，并且将指针网络用于朝鲜语句子排序，取得了较好的效果。接下来，我们将继续结合朝鲜语本身的特点，继续提高句子排序的效果，并将其用于多文档摘要任务中。

参考文献

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. *Association for Computational Linguistics*. Proceedings of NAACL-HLT 2, New Orleans, Louisiana, 82–91.

Yude Bi. 2011. On the study of Korean natural language processing. *Journal of Chinese Information Processing*, 25(6):166–169. *In Chinese*.

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. 2018. Universal Sentence Encoder. *ArXiv*. abs/1803.11175.

Xinchi Chen, Xipeng Qiu and Xuanjing Huang. 2016. Neural sentence ordering. *ArXiv*. abs/1607.06952.

John Duchi, Elad Hazan and Yorma Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Jingjing Gong, Xinchi Chen and Xipeng Qiu. 2016. Neural sentence ordering. *ArXiv*. abs/1611.04953.

Hongjian Guo and Bing Huang. 2013. The application of latent semantic analysis clustering algorithm in abstract sentence ordering. *Application Research of Computers*, 30(11):3299–3301. *In Chinese*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Shize Kang, Hong Ma and Ruiyang Huang. 2016. A method of sentence ordering based on neural network model. *Journal of Chinese Information Processing*, 30(5):195–202. *In Chinese*.

Yoon Kim, Yacine Jernite, David Sonta and Alexander M. Rush 2016. Character-aware neural language models. *ArXiv*. abs/1508.06615.

Lajanugen Logeswaran, Honglak Lee and Dragomir Radev. 2016. Sentence ordering using recurrent nerual networks. *ArXiv*. abs/1611.02654.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ArXiv*. abs/1301.3781.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre and Bing Xiang. 2018. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. *Association for Computational Linguistics*. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 280–290.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho and Alice Oh. 2018. Subword-level Word Vector Representations for Korean. *Association for Computational Linguistics*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2429–2438.

Patrice Y. Simard, Dave Steinkraus and John C. Platt. 2003. Best Practices for Convolutional Neural Networks. Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 958–963.

Hyun Je Song and Seong Bae Park. 2019. Korean Morphological Analysis with Tied Sequence-to-Sequence Multi-Task Model . *Association for Computational Linguistics*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 1436–1441.

Joseph Turian, Lev-Arie Ratinov and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. *Association for Computational Linguistics*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 384–394.

Oriol Vinyals, Meire Fortunato and Navdeep Jaitly. 2015. Pointer Networks. Advances in Neural Information Processing Systems 28, 2692–2700.

Yongdong Xu, Yadong Wang and Yang Liu. 2009. Research on the strategy of sentence ordering based on time information in multi document summarization. *Journal of Chinese Information Processing*, 23(4):27–33. *In Chinese*.

Tao Xue and Heng Wang. 2017. Research on sentence ordering based on conditional entropy and context proximity. *Application Research of Computers*, 34(9):2680–2684. *In Chinese*.

Chao Yao, Sheng Li and Shu Zhang. 2006. Sentence ordering of multi document abstracts based on cohesion. The academic conference of the 25th anniversary of the Chinese information society. *In Chinese*.