doi:10.3969/j.issn.1672-4348.2016.04.014

基于单词分类的归一化神经网络语言模型研究

陈铬亮[1], 徐佳[2]

(1.清华大学 交叉信息研究院, 北京 100084; 2.中国科学院计算技术研究所, 北京 100190)

摘要: 提出了一种基于单词分类的神经网络语言模型,以解决归一化问题。实验方法为,在基础翻译系统中加入模型参数,然后利用开发集调整参数,再对测试集进行翻译,对比加入模型参数前后的翻译质量以及训练模型和翻译过程所需时间。实验结果表明,在保证归一化的前提下,该模型的性能优于 Vaswani 等人的模型,且翻译质量与 Vaswani 等人的模型相当。

关键词: 机器翻译; 语言模型; 单词分类

中图分类号: TP391.2 **文献标志码**: A **文章编号**: 1672-4348(2016)04-0382-04

Research on word classification-based normalized neural network language model

Chen Geliang[1], Xu Jia[2]

(1. IIIS, Tsinghua University, Beijing 100084, China; 2. ICT, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: A word classification-based neural network language model was proposed to resolve normalization problems. Model parameters were introduced to the basic translation system, which were adjusted by development sets. The test sets were translated. The translation quality and training model and the time taken by the translation were compared. The results indicate that the model is superior to that of Vasvani in performance with its translation quality being similar to that of Vasvani.

Keywords: machine translation; language model; word classification

自然语言处理是人工智能研究的一个重要领域,该领域的研究目的是让计算机能够理解并自动处理人类的自然语言。语言模型是自然语言处理研究中的一个重要模型,它的作用是衡量一段语料的通顺程度。当前,语言模型被广泛运用于自然语言处理的各个方面,如语音识别,机器翻译,输入法和自动拼写纠错。

上世纪 50 年代,Shannon 提出了 n 元文法模型[1]。该模型用一段词语序列出现的概率来衡量这段文字的通顺程度。进一步,该模型假定词语序列是一个时齐马氏链,出于实际应用的需要,假定每一个词在给定它之前至多 $n-1$ 个词(称为该词的历史)的前提下与其他词无关。虽然距今已经六十多年,但 n 元文法模型依旧是最经典的语言模型。

传统的 n 元文法模型利用 n 元组的相对频率来估计每个词给定其历史的条件概率。随着计算机性能的提升,神经网络方法开始广泛运用于人工智能的各个领域,包括自然语言处理。2013 年,Vaswani 等人提出了一种基于神经网络的语言模型[2]。与传统的 n 元文法模型不同,Vaswani 等人从语料中提取出 n 元组,然后以每个 n 元组的前 $n-1$ 个词为输入,第 n 个词为输出来训练神经网络,得到一个神经网络语言模型。该网络的输入层为 $n-1$ 个词,输出层的结点个数等于字典大小,其输出正比于以这 $n-1$ 个词为历史的单词

收稿日期: 2016-07-22
基金项目: 国家自然科学基金(61033001);国家自然科学基金(61361136003)
第一作者简介: 陈铬亮(1990-),男,福建福州人,硕士研究生,研究方向:机器翻译,自然语言处理,人工智能。

条件概率分布。

一个概率模型，从理论上来说是需要归一化的。然而，一旦字典变大，将网络输出归一化就十分费时，这是实际应用所不允许的。Vaswani 等人引用了 Min 和 Teh 在研究中发现的一种高效的获得近似归一化结果的方法[2-3]，巧妙地回避了这个问题。然而，这并不代表归一化问题不存在。

另一方面，Kneser 和 Ney 在进行语音识别的研究时，提出了利用单词分类来提高效率的方法[4]。这启发我们从另一个角度考虑归一化问题的解决方法：产生归一化问题的根本原因是字典太大，如果字典不大，那么归一化就不需要太多时间，也就不存在效率问题了。于是，不考虑每个词给定前 $n-1$ 个词的条件概率，而是先将单词分类，然后考虑每个词的类别给定前 $n-1$ 个词的类别的条件概率。这样，输出层的结点数就从字典大小降低为类别个数，可以在实际应用允许的条件下进行归一化。而且，采用单词分类方法也可以减少语料的稀疏性对模型带来的影响[4]。本文提出了一种基于单词分类的神经网络语言模型，以解决归一化问题。

1　模型描述

Kneser 和 Ney 提出了基于单词分类的 n 元文法模型[4]：

设有一个单词序列 $w_1^n = w_1, w_2, \cdots, w_n$，语言模型的目标是估计其概率 $P(w_1^n)$。根据 n 元文法模型，有

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i \mid h_i)$$

其中，h_i 表示 w_i 的历史。设已经将单词唯一确定地分成了 J 类 C_1^J，第 i 类共有 k_i 个单词，单词 w_i 属于类 C_{w_i}，该类有 k_{w_i} 个单词。假定一个单词的概率只与它的历史所属的类别（而非历史本身）有关，则

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i \mid C_{h_i})$$

其中，C_{h_i} 为 h_i 所属类的序列。又假定在给定单词 w 的所属类 C_w 的情况下 w 与 C_h 独立，即

$$P(w \mid C_w, C_h) = P(w \mid C_w)$$

假如只知道一个词的所属类和该类的单词个数，而对其他信息一无所知的话，没有理由去假定这个词在该类中比其他词更频繁或更不频繁出现。所以，在上述模型的基础上进一步假定，每一类中的单词给定其所属类的条件概率相等，即

$$P(w \mid C_w) = \frac{1}{k_w}$$

于是有

$$P(w_1^n) =$$

$$\prod_{i=1}^{n} P(w_i \mid C_{h_i}) = \prod_{i=1}^{n} P(w_i, C_{w_i} \mid C_{h_i}) =$$

$$\prod_{i=1}^{n} P(w_i \mid C_{w_i}, C_{h_i}) P(C_{w_i} \mid C_{h_i}) =$$

$$\prod_{i=1}^{n} P(w_i \mid C_{w_i}) P(C_{w_i} \mid C_{h_i}) =$$

$$\prod_{i=1}^{n} \frac{1}{k_{w_i}} P(C_{w_i} \mid C_{h_i}) =$$

$$\prod_{i=1}^{n} \frac{1}{k_{w_i}} \prod_{i=1}^{n} P(C_{w_i} \mid C_{h_i}) \qquad (1)$$

(1)式由两部分相乘，左边是每个单词所属类的单词个数倒数之积，在给定每个单词及其所属类的情况下无需训练；右边是每个单词所属类的 n 元文法模型，可以用多种方法训练，比如传统的基于词频的方法和神经网络方法。本研究使用神经网络方法来训练每一个 $P(C_i \mid C_h)$ 并归一化，保证 $\sum_{i=1}^{J} P(C_i \mid C_h) = 1$。在这个神经网络的归一化性得到保证之下，这个模型也是归一化的（即 $\sum_{w_1^n} P(w_1^n) = 1$）。这是因为对于任意的历史 h，

$$\sum_{w} P(w \mid h) =$$

$$\sum_{w} P(w \mid C_w) P(C_w \mid C_h) =$$

$$\sum_{i=1}^{J} \sum_{w \in C_i} P(w \mid C_w) P(C_w \mid C_h) =$$

$$\sum_{i=1}^{J} \sum_{w \in C_i} \frac{1}{k_i} P(C_i \mid C_h) =$$

$$\sum_{i=1}^{J} k_i \frac{1}{k_i} P(C_i \mid C_h) =$$

$$\sum_{i=1}^{J} P(C_i \mid C_h)$$

单一的单词分类方法不能保证取得良好的翻译质量。为此，可采用多种单词分类方法，对每种方法得到一个语言模型，最后将这些模型合并，得

到最终的语言模型：

$$P(w \mid h) = \prod_{m=1}^{M} (P_m(w \mid h))^{\lambda_m}$$

其中，λ_m 是权重系数，可利用开发集来调整。

2 模型的实现

2.1 单词分类

采用 Kneser 和 Ney 提出的统计学习方法[4]，利用 Och 和 Ney 发布的工具 GIZA++中的 mkcls 组件[5]将训练集的单词分为 100、200、300 和 400 类。

2.2 训练语言模型

得到单词分类表后，先选择每类中词频最高的单词作为该类的代表，然后将训练集和开发集当中的所有单词都替换为该单词所在类的代表，最后采用 Vaswani 等人的方法训练神经网络语言模型，参数与 Vaswani 等人文章中的参数基本一致。最后共得到 4 个神经网络模型 LM1-4。

2.3 测试集概率的计算

根据单词分类表将每个单词替换为该单词所在类的代表，然后代入 2.2 中训练所得模型，并且乘以 $\prod_{i=1}^{n} \frac{1}{k_{w_i}}$。对 Vaswani 等人发布的工具作了一些修改，使之能够自动实现上述步骤。

2.4 权重系数 λ_m 的确定

权重系数根据模型在开发集上的翻译质量来调整。

3 实验及结果

用机器翻译实验检验模型，实验内容是汉语到英语的翻译任务。

训练集、开发集和测试集均取自 IWSLT2014[6]的汉语-英语语料，使用双语训练集的英语部分作为语言模型训练语料。语料的统计数据见表 1，为方便仅列英文部分的统计数据。翻译工具使用 Moses[7]，为 Moses 的默认设置，采用短语翻译模型[8]，对齐工具用 GIZA++[5]，使用 MERT[9]方法调整各模型的权重系数。

首先以 n 元文法模型为语言模型进行翻译，以此为基准，对每个汉语句子输出其最好的 100 个候选翻译，作为基准候选翻译表，并取最好的翻译候选作为基准翻译结果。随后，对每个候选翻译分别使用 LM1-4 求出对数概率，以此为特征加入基准候选翻译表，再根据开发集调整各模型的权重系数，得到翻译结果。为了与经典的神经网络语言模型比较，按照文献[2]中的参数，训练了一个归一化的神经网络语言模型 LM0，并分别将其给出的对数概率加入基准候选翻译表，同样调整权重系数，得到翻译结果。用 BLEU[10]分数来评价翻译结果的好坏，分数越高结果越好。

表 1 实验所用语料的统计数据

Tab.1 Statistics of experimental corpus

语料	Train2014	Dev2010	Tst2010	Tst2011	Tst2013	Tst2014
句子数	181 226	886	1 570	1 450	1 299	1 092
单词数	3 594 550	20 219	32 062	27 018	28 947	21 194
词典大小	60 037	3 221	3 714	3 446	4 025	3 340

实验结果见表 2。虽然 LM1-4 只考虑了分类信息而没有考虑具体的单词信息，但加入 LM1-4 的翻译结果不逊于加入 LM0 的结果。也就是说，LM1-4 在满足归一化要求的同时不会降低翻译质量。这个结果与 Kneser 和 Ney 的研究结果[4]一致。LM1-4 翻译结果良好的原因是，基于单词分类的模型比基于具体单词的模型更加稳健，一定程度上解决了训练样本稀疏性的问题。

表 2 翻译实验结果

Tab.2 Result of translation in BLEU score

%

数据集	Dev2010	Tst2010	Tst2011	Tst2013	Tst2014
基准	9.01	10.75	12.92	14.44	12.57
LM0	9.39	10.83	12.99	14.46	12.45
LM1-4	9.33	10.92	13.14	14.50	12.53

测试了训练 LM0 和 LM1-4 以及运用它们求测试集概率的耗时。测试所用计算机的 CPU 为 Intel(R) Core(TM) i7-4700HQ 双核 2.4GHz，内存为 4GB，均以 8 线程运行。

测试结果见表 3。可以看出，LM1-4 在求测试集概率时的耗时要远短于 LM0，这是符合预期的，因为归一化所需时间与字典大小成正比，LM1-4 的字典大小要远小于 LM0，花在归一化上的时间自然就更短。但在训练时间方面 LM1-4 要长于 LM0，主要是因为单词分类需要消耗大量时间，类别越多，分类所需时间就越长，LM4 的单词分类时间甚至超过了神经网络模型的训练时间。不过，考虑到在翻译实务中需要面对远多于

测试集的数据，在翻译速度上占优的 LM1-4 显然好于 LM0。

表 3　各模型训练和解码耗时

Tab.3　Time taken for the training and decoding of models

模型	LM0	LM1	LM2	LM3	LM4
训练时间/min	82	81	99	134	155
解码时间/s	1 560	45	49	55	52

4　结语

实验证实了将单词分类的方法运用到神经网络语言模型中以解决归一化问题的可行性。在翻译实践中，归一化模型是否优于非归一化模型，还需进一步的研究；但在理论上，归一化模型的数学基础远比非归一化的模型来得扎实可靠。

参考文献：

[1] Shannon C E. Prediction and entropy of printed English[J]. Bell System Technical Journal, 1951, 30(1):50-64.

[2] Vaswani A, Zhao Y, Fossum V, et al. Decoding with large-scale neural language models improves translation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, America: Association for Computational Linguistics, 2013:1387-1392.

[3] Mnih A, Teh Y W. A fast and simple algorithm for training neural probabilistic language models[C]//Proceedings of the 29th International Conference on Machine Learning. Edinburgh: International Machine Learning Society, 2012:1751-1758.

[4] Kneser R, Ney H. Improved clustering techniques for class-based statistical language modelling[C]//Eurospeech'93. Berlin, Germany: International Speech Communication Association, 1993:973-976.

[5] Och F J, Ney H. A systematic comparison of various statistical alignment models//[J]. Computational Linguistics, 2003, 29(1):19-51.

[6] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Prague, Czech: Association for Computational Linguistics, 2007.177-180.

[7] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada: Association for Computational Linguistics, 2003.127-133.

[8] Och F J. Statistical machine translation : from single word models to alignment templates[J]. Rwth Aachen, 2002, 10(2): 65-70.

[9] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, America: Association for Computational Linguistics, 2002.311-318.

(责任编辑：陈雯)