文章编号：2095-2163(2019)06-0188-05 中图分类号：TP391.1 文献标志码：A

基于自注意力机制的口语文本顺滑算法

吴双志，张冬冬，周 明
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：口语文本顺滑技术是语音翻译系统中的重要组成部分。其目标是识别并删除语音识别文本中所包含的重复、停顿、修正、冗余等口语现象，进而使口语文本更加书面化，增加文本的可读性和可理解性，有助于提高后续语言处理任务的准确率。本文针对口语文本顺滑问题提出一种基于自注意力机制的识别技术。该技术利用了深度学习中的自注意力神经网络。自注意力神经网络具有很强的序列建模能力，本文首先利用自注意力网络对口语文本进行编码，在此基础之上识别文本中的不流畅因素。在公开数据集上的测试结果表明本文提出的方法可以有效地识别口语中的不流畅因素。

关键词：语音翻译；口语顺滑；神经网络

Disfluency detection with self-attention network

WU Shuangzhi, ZHANG Dongdong, ZHOU Ming
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Disfluency detection is a very important technique in simultaneously speech translation. It aims at detecting the repetition, redundancy, pauses in the spoken languages to make the text more fluent and readability. In this paper, we propose a self-attention-based detection method. We leverage the self-attention to encode the spoken sentences, based on which disfluent words are detected. We conduct experiments on the public dataset and the results show that our method can effectively detect disfluencies in spoken language.

【Key words】 speech-to-speech translation; disfluency detectio; neural network

0 引 言

随着移动互联网技术的飞速发展，以互联网为基础的应用层出不穷。这些应用大多都不再局限于以文本作为输入信号，语音作为一种更加方便快捷的输入备受推崇。以机器翻译为例，实时的语音翻译逐渐成为人们出行的必备工具。自动语音翻译是指让机器完成从某一种语言的语音翻译到另一种语言的语音过程。简而言之，语音翻译通常由 3 个模块组成，分别是语音识别(ASR)、文本翻译(MT)和语音合成(TTS)。其中语音识别能够将语音转化为文本，是语音翻译的重要组成部分。通常情况下，用户说出来的语音信号都是实时发生的，以口语风格为主。口语语音识别文本与正规书面文本有很大差别，这主要体现在口语文本和书面文本相比含有很多不流利因素。所谓不流利因素是指人们在讲话时所出现的重复、停顿、修正、冗余等现象。这些现象会导致口语本文中出现很多噪音单词，本文称之为不流畅单词。下面列出了口语文本不同于书面文本的一些例子：

(1) I want a fight to ***Denver I mean*** to Boston.

(2) ***Um and are these like*** do these programs.

(3) **他 这边 他 那个** 他 不想 干 嘛。

上例中斜体加黑代表这些词对于整个句子是多余的，是不流畅单词。在句子(1)中，由于讲话者对其先前的表述有所修改，因此“***to Denver I mean***”这一片段对该句子的意思没有贡献是冗余的。又例如在句子(3)中，由于讲话者的犹豫或者停顿，使句子中多出了一些不必要的单词“那”和“那个”，这些单词同样对句子的意思没有贡献。从这些例子中可以看出，口语文本中的不流利因素会使句子的整体可读性变差，机器翻译在处理这些文本的时候会出现很多翻译错误。口语顺滑任务就是为了识别并去掉口语中的这些不流畅现象，使口语文本书面化。这对提高语音翻译质量有很大的帮助。

本文针对语音翻译中的口语顺滑问题进行建模，并提出一种基于自注意力机制的识别算法。该方法利用深度学习中的自注意力神经网络对口语文本进行编码，为每一个口语单词生成一个上下文向量，然后将这个上下文向量输入到一个分类器中，去识别口语文本中的每一个单词是否为不流畅单词。自注意力神经网络具有很强的句子建模能力，能够

作者简介：吴双志(1991—)，男，博士研究生，主要研究方向：口语文本处理、机器翻译。

收稿日期：2019-04-18

充分地编码句子中的上下文信息，有助于识别句子中的冗余成分。

1　相关工作

近年来，语音翻译技术蓬勃发展，其应用场景也逐渐变多。因此如何提高语音翻译的性能备受关注。在语音翻译中，语音识别文本是整个翻译系统的基础，对整个系统的性能有至关重要的影响。而口语和书面语的诸多不同，导致直接识别出来的口语文本大大降低了后续自然语言处理任务的性能。因此口语文本顺滑任务逐渐被很多从事语音翻译的研究人员所重视。

解决口语文本顺滑任务的方法大体上可以分为 4 类：基于噪声信道模型的口语顺滑算法[1-2]、基于序列标注的口语顺滑方法[3-5]、基于句法的口语顺滑算法[6-9]、基于神经网络的方法[10-11]。在具体的特征选择方面，一般分为文本特征和与声学有关的特征，比如音律特征等[12-13]。下面将简述以上几类方法。

(1)基于噪声信道的口语顺滑方法。是最早被提出来的检测方法。该方法将顺滑任务视为一个信号去噪的过程。与统计机器翻译相似，采用对数线性模型处理口语噪音，取得了不错的效果。之后语言模型和重排序模型相继被引入到 TAG 模型中[14]。

(2)基于序列标注的识别算法。是将口语顺滑任务当做序列标注问题。Liu(2006)[4]等人提出了基于条件随机场(CRF)的口语顺滑模型，CRF 模型对口语文本的每一个单词进行标注，利用丰富的上下文相关特征来检测不流畅单词。之后，更多的特征被用来识别不流利单词都取得了很好的效果。除了 CRF 模型之外，Qian 和 Liu(2013)[5]等人提出了基于最大间隔马尔科夫随机场的口语顺滑模型(M^3N)，而后 Wang(2014)[15]等人在 Qian 的工作基础上，提出了基于柱搜索的解码算法，取得了很好的效果。

(3)基于句法的顺滑算法。Lease(2006)[7]等人提出了基于 PCFG 句法分析器的口语顺滑模型，该方法认为口语中很多噪音单词从结构上是与整句话多的语法结构有冲突的，因此可以通过句法结构来辅助识别不流畅单词。实验证明句法的确有助于识别口语噪音。随后 Miller、Rasooli、Honnibal 等人相继提出了不同的基于句法的口语文本顺滑算法[6,8-9]。

(4)近年来深度学习技术在自然语言处理领域展现了很强的建模能力，因此也将深度学习技术用于口语顺滑任务。Wang Shaolei 等人[11]率先提出基于端到端神经网络的口语顺滑模型，以序列到序列的模型对口语顺滑任务建模取得了很好的实验效果。此后 Wang Shaolei 等人[10]将神经网络模型和句法模型相结合，提出了一种新型的基于循环神经网络的口语顺滑建模方法，该方法利用了长短期记忆网络的序列建模能力，同时结合了句法结构，取得了非常好的识别效果。

2　本文算法设计研究

本节提出一种基于自注意力网络的口语顺滑技术，该算法利用自注意力神经网络对口语文本进行编码，然后在编码结果基础上识别口语中的不流畅单词。

2.1　自注意力神经网络

自注意力神经网络源于 Vaswani 等人在文献[16]中提出的 Transformer 模型。该模型是一个端到端的神经网络模型，由编码器和解码器组成。其中编码器和解码器均由自注意力神经网络组成。下面以编码器为例重点介绍自注意力神经网络。

Transformer 的编码器由 N 个同构的网络层堆叠而成。每一个网络层包含两个子网络层：第一个子网络层称为分组自注意网络，用于将同层的源语言句子里的其它词的信息通过自关注网络考虑进来，以生成当前词的上下文向量；第二个子网络层是一个全联通的前馈神经网络，该网络的作用是将自关注网络生成的源语言句子内的上下文向量同当前词的信息进行整合，从而生成考虑了整个句子上下文的当前时刻的隐含状态。为提高模型的训练速度，残差链接(Residual Connection)和层规范化(Layer Normalization)被用于这两个子网络层，即图中的 AddNorm 层，定义为 $\mathrm{Norm}(x+\mathrm{SubLayer}(x))$，其中 x 为子网络的输入；SubLayer 为该子网络的处理函数；Norm 为层规范化函数。通过对 N 个这样的网络层堆叠可以对信息进一步进行抽象和融合。一般情况，自注意力网络由多路注意力机制实现，如图 1 所示(该图源自文献[16])。

图 1 中 Q 为检索向量；K 为键向量；V 为值向量。对于第一层自注意力网络来说，三元组(Q,K,V)由词向量和位置向量计算得到。对于其它层，(Q,K,V)均由前一层的输出计算得到。传统的注意力机制只使用一个注意力网络来生成一个上下文向量，而多路注意力网络将多个注意力网络进行拼接。首先使用不同的线性映射分别将 Q、K 和 V 映射到不同的空间，然后使用不同的注意力网络计算得

到不同空间的上下文向量,并将这些上下文向量拼接得到最后的输出。计算公式为:

$$MultiHead(Q,K,V) = Concat_i(head_i)\ W^o$$

$$head_i = Attention(Q\ W_i^Q, K\ W_i^K, V\ W_i^V)$$

$$Attention(Q,K,V) = softmax(\frac{Q\ K^T}{\sqrt{d_k}})V$$

其中, $\boldsymbol{W}$ 为权值矩阵, *Attention* 为点成注意力函数。在多路注意力机制计算完成之后,其输出结果会通过一个全连接网络得到该层的输出结果。这个全连接层包括 2 个线性映射和 1 个 RELU 激活函数,具体计算过程如下所示:

$$FFN(x) = \max(0, xW + b)\ W_2 + b_2$$

其中, W 和 W_2 是权重矩阵。对于同一层来说,每一个位置共享这两个权重矩阵,而不同层之间的权重矩阵是不同的。

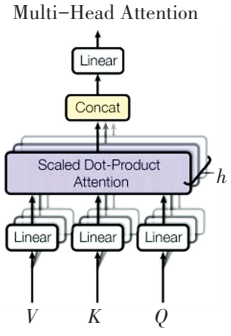

图 1 多路注意力机制[16]

Fig. 1 Multi-head Attention

2.2 基于自注意力网络的顺滑算法

本节将详细介绍基于自注意力网络的口语文本顺滑算法。首先给定一个口语句子:

I want a flight ***to Boston um*** to Denver .

居中加黑斜体的单词属于不能流畅单词,其余单词为正常单词。定义这些不流畅单词的类别为 D,即 disfluent。其它单词定义为 N 类,即 normal。因此将口语顺滑任务转化成了分类任务。对于一个口语句子,通过对句中每一个词进行分类,最终将冗余单词删去即可得到一句正规文本。图 2 给出了基于自注意力网络的口语文本顺滑模型图。

如图 2 所示,该网络口语文本作为输入句子,首先将词语转化为词向量,再与位置向量相加作为编码器的输入,其计算方法与文献[16]一致。然后利用多层自注意力网络对口语文本进行编码,同时为每一个单词生成一个隐层向量。最后将每一个词的隐层向量作为 logistic 分类器的输入来预测该词的类别。

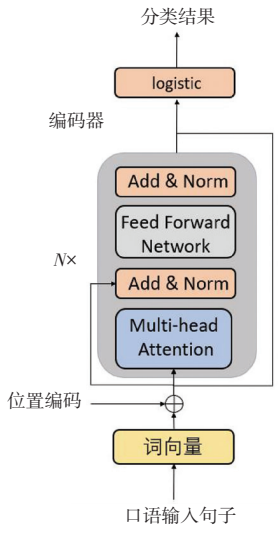

图 2 基于自注意力网络的口语文本顺滑模型

Fig. 2 The self-attention based disfluency detection model

2.3 特征设计

一般情况下,口语顺滑技术都会提取一些特征来辅助识别[5,15,10-11]。本文同样利用了两种简单直观的词法特征。这些特征对提升模型性能起到了关键作用。特征定义如下:

- $\delta_w(a,b)$:逻辑函数, 如果 a 等于 b 返回真,否则返回假。
- $\delta_P(a,b)$:逻辑函数,如果 a 和 b 的词性相同返回真,否则返回假。

文中为口语句子中的每一词设计若干特征。定义 wi 为口语句子中的一个单词,其特征包括 6 个词法特征:

$\delta_w(w_i, w_{i+1})$, $\delta_w(w_i, w_{i+2})$, $\delta_w(w_i, w_{i+3})$, $\delta_w(w_i, w_{i-1})$, $\delta_w(w_i, w_{i-2})$, $\delta_w(w_i, w_{i-3})$ 以及 6 个词性特征;

$\delta_p(w_i, w_{i+1})$, $\delta_p(w_i, w_{i+2})$, $\delta_p(w_i, w_{i+3})$, $\delta_p(w_i, w_{i-1})$, $\delta_p(w_i, w_{i-2})$, $\delta_p(w_i, w_{i-3})$ 。这些特征以向量的形式拼接到词向量中一起作为自注意力模型的输入 。

3 实验评估

3.1 实验数据和评测指标

本文主要针对英文口语文本中的不流利现象进行检测。实验数据采用公开的英文滨洲树库[17]中的 Switchboard(SWBD)数据集进行实验。这部分数据集主要包含的是电话语音数据的识别文本。这些文本是依据 Metter 等人[18]中提出的规范标注的。本文将这个数据集划分为 3 个部分,分别为训练数据集、测试数据集和开发集。划分方法遵从

Charniak(2001)[19]中给出的标准。由于该部分数据的句法结构是 PCFG 结构，因此采用斯坦福句法分析转换工具[4]，将其转化为依存结构。对于实验结构的评估，本文采用 F1 分数，具体定义如下，其中 Prec.表示准确率，*Rec.*代表召回率：

$$Prec. = \frac{\text{正确识别的数量}}{\text{识别的总量}},$$

$$Rec. = \frac{\text{正确识别的数量}}{\text{文本中不流利单词的数量}},$$

$$F1. = \frac{2 * Prec. * Rec.}{Prec. + Rec.}.$$

3.2　基线模型

对于基线模型，本文选取比较常用的基于 CRF 序列标准的识别算法。在基于 CRF 的标准算法中，定义 3 种标签：*N* 为正常单位词；D－B 为不流畅片段的起始；D－I 为不流畅段的中间以及结尾。给定一个语音识别句子 $S = W_1, W_2, \cdots, W_i, \cdots, W_n$，以及 W_i 的词性 POS_i，该词性信息由外部的句子分析器预处理获得。特征模板设计见表 1。

表 1　特征模板

Tab. 1　Feature template

词法特征	W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}, $W_{i-1} + W_i$, $W_i + W_{i+1}$
逻辑函数	$I(W_i)$
词性特征	POS_{i-2}, POS_{i-1}, POS_i, POS_{i+1}, POS_{i+2}, $POS_{i-2} + POS_{i-1}$, $POS_{i-1} + POS_i$, $POS_i + POS_{i+1}$, $POS_{i+1} + POS_{i+2}$, $POS_{i-2} + POS_{i+1} + POS_i$, $POS_{i-1} + POS_i + POS_{i+1}$, $POS_i + POS_{i+1} + POS_{i+2}$

$I(w_i)$ 为逻辑函数，定义如下：

$$I(W_i) = \begin{cases} 1, & W_i \in dis\_table \\ 0, & \text{否则} \end{cases}$$

其中，*dis_table* 是从训练数据中统计出来的词表。该词表包含了在训练数据中 5 次以上被标记为不流畅的单词集合。

3.3　模型实现

对于基于自注意力网络的算法，本文依照 Vaswani 等人与文献[16]中提出的基础模型参数来设置网络。模型维度为 512，过滤层维度设置为 2 048，query、key 和 value 维度均为 64、路数为 8。与 Vaswani 不同，本文采用 12 层编码网络。此外，本文采用文献[20]提出的 adam 进行参数更新，其参数设置为 $\beta_1 = 0.9$，$\beta_2 = 0.98$。在训练过程中，学习率 *lrate* 随着训练步数逐渐变化，如以下公式所示。

$$Lrate = d_{moder}^{-0.5} \, min(stepnumber^{-0.5}, stepnumber \times warmup^{-1.5})$$

由上式可知，在 warmsteps 个训练步长中，模型的学习率呈线性增长，同时在 warmsteps 之后，以指数的速度下降。本文设置 warmsteps 大小为4 000。模型在单块 M40GPU 上训练。

3.4　实验结果

表 2 给出了在英文 SWBD 数据集上的实验结果。由表可知，CRF 模型只用词法级别特征时，准确率较低，只有 57.3%，其主要原因是词法特征的建模能力有限，对简单的不流畅单词识别能力尚可。当结构复杂时就很容易出现识别错误。当加入词性特征之后，CRF 模型的识别准确率得到了很大程度的提高，达到了 73.0%，这说明词性特征对口语顺滑任务是十分有帮助的，但由于这些特征都局限于局部特征，很难对长距离的依赖关系建模。因此模型的识别能力依然十分有限。本文提出的基于自注意力网络的算法的准确率为 80.7%，比 CRF 模型高出了 7 个百分点，说明本文提出方法比传统的序列标注模型更适合口语顺滑任务。这是由于自注意力机制可以从全局对口语本文建模，对长距离的依赖关系有很好的建模能力，对于识别复杂的口语文本噪音非常有帮助。此外，鉴于词性特征对口语顺滑任务有帮助，当本文提出的模型同时利用上词性特征后，其识别准确率达到了 82.5%。这说明本文提出的方法依然可以很好地利用传统的识别特征进一步提高模型的识别能力。

表 2　实验结果

Tab. 2　Experimental Results

模型	准确率/%
CRF(BOW)	57.3
CRF(BOW+POS)	73.0
基于自注意力网络算法	80.7
基于自注意力网络算法(+特征)	82.5

4　结束语

在移动互联网蓬勃发展的今天，人们对多元化信息的渴求越来越强烈。因此语音翻译逐渐被人们重视，其中对口语文本的处理对整个语音翻译的效果有着重要的影响。针对口语文本存在很多重复、停顿、修正、冗余等现象，本文提出了基于自注意力网络的口语顺滑算法。利用强大的自注意力网络帮助文本建模，并在此基础上识别口语中的冗余成分。实验结果表明本文提出的方法可以有效地识别口语中的不流畅单词。

参考文献

[1] Johnson M, Charniak E. A TAG-based noisy channel model of speech repairs[C]. In the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 33.

[2] Zwarts S, Johnson M. The impact of language models and loss functions on repair disfluency detection[C]. In the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 703-711.

[3] Georgila K. Using integer linear programming for detecting speech disfluencies [C]. In the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009: 109-112.

[4] Liu Y, Shriberg E, Stolcke A, et al. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2006, 14(5): 1526-1540.

[5] Qian X, Liu Y. Disfluency Detection Using Multi-step Stacked Learning[C]. In the HLT-NAACL. 2013: 820-825.

[6] Honnibal M, Johnson M. Joint incremental disfluency detection and dependency parsing [J]. Transactions of the Association for Computational Linguistics, 2014, 2: 131-142.

[7] Lease M, Johnson M. Early deletion of fillers in processing conversational speech[C]. In the Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006: 73-76.

[8] Miller T, Schuler W. A unified syntactic model for parsing fluent and disfluent speech[C]. In the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, 2008: 105-108.

[9] Rasooli M S, Tetreault J R. Joint Parsing and Disfluency Detection in Linear Time[C]. In the EMNLP. 2013: 124-129.

[10] Wang, Shaolei, et al. "Transition-based disfluency detection using LSTMs." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[11] Wang, Shaolei, Wanxiang Che, and Ting Liu. "A neural attention model for disfluency detection." Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.

[12] Kahn J G, Lease M, Charniak E, et al. Effective use of prosody in parsing conversational speech [C]. In the Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005: 233-240.

[13] Zhang Q, Weng F, Feng Z. A progressive feature selection algorithm for ultra large feature spaces[C]. In the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 561-568.

[14] Johnson M, Charniak E, Lease M. An improved model for recognizing disfluencies in conversational speech [C]. In the Proceedings of Rich Transcription Workshop. 2004.

[15] Wang X, Ng H T, Sim K C. A beam-search decoder for disfluency detection[C]. In the Proc. of COLING. 2014.

[16] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[17] Marcus M P, Marcinkiewicz M A, Santorini B. Building a large annotated corpus of English: The Penn Treebank[J]. Computational linguistics, 1993, 19(2): 313-330.

[18] Meteer M W, Taylor A A, MacIntyre R, et al. Dysfluency annotation stylebook for the switchboard corpus[M]. University of Pennsylvania, 1995.

[19] Charniak E, Johnson M. Edit detection and parsing for transcribed speech[C]. In the Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, 2001: 1-9.

[20] Kingma Diederik p, and Jimmy Ba."Adam: Amethod for stochastic optimization." arXiv perprint arXiv:1412.6980(2014).

(上接第187页)

[6] 邬文娟, 傅惠民. 仿真结果检验和状态分析方法[J]. 航空动力学报, 2010, 25(5): 1092-1096.
WU Wenjuan, FU Huimin. Method for test and state analysis of simulation results[J]. Journal of Aerospace Power, 2010, 25(5): 1092-1096.

[7] 傅惠民, 邬文娟. 确定性仿真结果检验与状态分析方法[J]. 航空动力学报, 2011, 26(5): 1124-1127.
FU Huimin, WU Wenjuan. Methods for test and state analysis of determinate simulation results[J]. Journal of Aerospace Power, 2011, 26(5): 1124-1127.

[8] 傅惠民, 陈建伟. 动态仿真结果距离检验方法[J]. 机械强度, 2007, 29(2): 206-211.
FU Huimin, CHEN Jianwei. Distance test method for dynamic simulation results. Journal of Mechanical Strength, 2007, 29(2): 206-211.

[9] 傅惠民, 林逢春, 陈建伟. 动态仿真结果相关性检验和分析方法[J]. 机械强度, 2007, 29(4): 584-588.
FU Huimin, LIN Fengchun, CHEN Jianwei. Test and analysis method for correlation of dynamic simulation results[J]. Journal of Mechanical Strength, 2007, 29(4): 584-588.

[10] 傅惠民, 陈建伟. 仿真结果距离检验方法和 TIC 方法对比分析[J]. 航空动力学报, 2009, 24(12): 2784-2788.
FU Huimin, CHEN Jianwei. Comparison between distance test method and TIC method for simulation results[J]. Journal of Aerospace Power, 2009, 24(12): 2784-2788.

[11] MEEKER W Q, ESCOBAR L A. Statistical methods for reliability data[M]. John Wiley & Sons, 2014.

[12] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1982.

[13] 王宗仁, 李毅, 刘波, 等. 一种基于压簧应力松弛测试数据的可靠度确定方法: 中国, 105300673B[P]. 2017-07-28.

[14] SCHIJVE J. Fatigue of structures and materials[M]. Springer Science & Business Media, 2001.

[15] 吴琼, 傅惠民. 橡胶密封性能多元双方差回归分析方法[J]. 航空动力学报, 2011, 26(8): 1855-1859.
WU Qiong, FU Huimin. Sealing performance multivariate two-variance regression analysis method for rubbers[J]. Journal of Aerospace Power, 2011, 26(8): 1855-1859.

[16] 国家经济贸易委员会. GB/T 3087-2001 静密封橡胶零件贮存期快速测定方法[S]. 北京: 国家经济贸易委员会, 2001.