基于生语料、最大匹配切分语料以及熟语料的中文词频估计方法

乔维，孙茂松

清华大学智能技术与系统国家重点实验室 100084

摘 要：词频估计在NLP的各个领域中都有着重要的应用，中文的特点使得中文词频估计对我们来说依然是一个严峻的挑战。其中一个主要因素就是缺少一个供我们作词频估计的“完美的”语料库。我们现有的语料库有：规模可以任意大的生语料库；由生语料库通过自动分词得到的已切分语料库；一些规模较小，由不同机构根据不同的分词标准开发的熟语料库。基于以上所有因素及已有的语料库，本文提出了一种基于折中的思想，综合利用已有信息来进行中文词频估计的方法。实验表明这一策略在多数情况下能够显著提高词频估计的准确度，但在某些情况下，这一方法的性能仍不够满意。

关键字：词频，生语料，MM-语料，熟语料

Word Frequency Approximation for Chinese Using Raw MM-Segmented Manually Segmented Corpora

QIAO Wei, SUN Maosong

(State Key Lab of Intelligent Technology and Systems, Tsinghua University, 100084)

Abstract: Word frequencies play important roles in NLP-related applications. Word frequency estimation for Chinese remains a big challenge due to the characteristics of Chinese. An essential factor is that a perfect Chinese corpus never exists. Currently we only have raw corpora which can be arbitrary large in size, automatically word segmented corpora derived from raw corpora and a variety of manually word segmented corpora which are developed by different institutions under different segmentation standards which are relatively small in size. A scheme to do word frequency approximation by combining the above factors is proposed in this paper. Experiments indicate that in most cases this strategy can significantly benefit the word frequency estimation, though in some cases its performance is still not very satisfactory.

Keywords: word frequency, raw corpus, MM-segmented corpus, manually segmented corpus.

1. 引言

词频估计在自然语言处理中有着重要的应用，比如信息检索领域的TF（term frequency）。对于英文，我们可以在一个大规模语料库中简单地通过统计词的出现次数来得到精确的词频估计结果。然而由于在中文的文本里，词与词之间没有天然的隔断，从而造成了这一简单任务在中文中变得困难和复杂。

一般来说，要进行中文词频估计，我们需要有一个完全正确的经人工切分的中文语料库。然而我们面临两个

基金支持：本文得到国家自然科学基金项目（NO.60573187） 和（NO.60321002），清华-ALVIS 项目以及国家自然科学基金项目NO.60520130299 和EU FP6的共同资助。

作者简介：乔维，硕士研究生，E-mail：qiaow04@mails.tsinghua,.edu.cn

困难：首先，在几个人工切分的语料库之间存在着严重的不一致性。原因是，由于中文词独特的构词特点，尽管“词”的定义在语言学家的角度看来非常清晰，然而中文里存在一批词，我们既可以把它们看作是一个复合词，比如“猪肉”，也可以把它们看作是一个由两个一字词“猪”和“肉”组成的短语。因此当我们统计“猪肉”这个词的词频时，如果按照前一种理解，那么该词的词频会非常高，然而如果按照后者理解，那么“猪肉”在语料库中出现次数就是零。其次，根据 Zipf 定律，我们可以知道，如果想要得到一个中等规模的词表的词频估计，需要上亿字，而不仅仅是上百万字的语料库。如此大规模的人工标注的语料库费时费力，目前也是不可能得到的。

基于上述因素，我们这里考虑用以下几个语料库进行词频估计：

一、用一个“完美”分词器对一个语料库进行自动切分，从而得到词频的估计。理论上讲，这个方法无疑是最好的。然而这样的一个“完美”分词器目前是不存在的。尽管在过去的二十年里，研究者们在这方面作了很多的努力，然而中文自动分词的性能依然不尽如人意。SIGHAN 在 2003 年举办了第一届中文分词比赛(Sproat and Emerson 2003)，在四个小范围的语料集上作的开放测试中，F-scores 最高分别达到 95.9%, 95.6%, 90.4% 和 91.2%。而在第二届分词比赛中(Emerson 2005)，分词的性能虽然在小范围内有了一定的提高，但并未从本质上解决这一难题。分词的性能依然不能满足实际的需要，尤其是未登录词出现时，分词的性能受到严重的影响。

二、第二种是用 MM 切分语料作词频估计。我们用一种最基本的分词方法——最大匹配（MM）分词法对语料进行切分。然后从经 MM 切分后的语料中得到我们所需的词频估计值。Liu and Liang (1986) 第一次将 MM 切分方法用于处理大规模的文本。根据扫描顺序的不同，MM 可以进一步分为前向最大匹配（FMM）和后向最大匹配（BMM）两种。在 Liang (1987) 的工作中，结论显示 MM 是一种既有效而且高效的分词方法(速度快而且易用)。Sun and T'sou (1995) 的工作表明基于 MM 的机制进行词频估计是非常有效的。使用基于 MM 自动分词机制的另一个优点是在分词中较高的一致性。这种方法的弱点在于，和其他分词方法一样 MM 法不可避免的同样的存在分词错误，而且当未登录词出现时，其性能下降很厉害。

三、第三种是利用生语料进行词频的估计。这里我们考虑用“串频”近似地代替词频 (Sun, Shen, and T'sou 1998)。“串频”信息可以直接从任意一个未经人工处理的生语料库中统计得到。很明显，对任一个词，统计出的“串频”值都会高于该词在语料中真实的词频值。对某些词，这一方法会出现“过估计”现象。对于单音节词，这一现象尤其严重。但是这种机制、有两个优点：首先该方法避免了切分错误的影响。其次，生语料非常易得，而且理论上来说其规模可以任意大。

综上所述，对于中文词频估计这一任务，完美切分语料是进行词频估计的最理想的语料。然而目前是无法得到的。其余的方法各有其优缺点。没有一个可以有效准确地独立完成词频估计这一任务。因此，这里我们考虑用一种折中的方法：综合利用人工标注语料、MM 切分语料和生语料进行中文词频估计。后面的文章按以下内容组织：第二部分介绍在本文中使用的各种数据集。第三部分将详细介绍提出的词频估计的新机制。第四部分展示基于新机制进行词频估计的实验结果。第五部分对我们的工作进行总结。

2. 数据集

在本文的实验中，我们共用到两个人工标注的语料库：第一个是由清华大学开发的华语语料库 HUAYU（1,040,190 词 1,763,762 字）。第二个是由北大开发的，这里我们称 BEIDA 语料库（5,659,831 词 15,839,323 字）。这样我们就有一个总计 17,603,085 字的经人工标注的语料库。

为了作对比测试，我们需要有一个标准语料库作为测试标准。国家语委语料库是一个经人工校对规模较大的，具有权威性的语料库。因此我们用 YUWEI 语料库（25,000,309 词 51,311,659 字）作为标准。 从 YUWEI 语料库里我们可以从中抽取出一个词表，并统计出相应的词频信息。去除频度小于 4 次的词后得到一个包含 99,660 个词的标准词表，这里我们记为 YWL

本文中我们选用一个规模 447,079,112 字的生语料库。我们将这个生语料库记作 RC。用 YWL 作为词表，可以从生语料中提取相应词的串频信息。

最后，是 MM 切分语料库，用 YWL 作为词表，我们分别用 FMM 和 BMM 分词器对生语料进行切分。从而得到两个 MM 切分语料。我们记作 RC_FMM 和 RC_BMM。

综上，我们有两个中等规模的人工标注语料库，(HUAYU 和 BEIDA)；一个规模较大的生语料库(RC)；两个经 MM 切分的语料(RC_FMM 和 RC_BMM)；一个用于做测试的标准语料库 (YUWEI)。

3. 词频估计的新机制

这一章里介绍中文词频估计的新机制。为了恰当结合已有的五个语料库进行中文词频估计，我们将合并步骤分为三步进行：首先合并生语料及 MM 切分语料；其次，合并熟语料；最后合并前面两步得到的结果。

3.1 合并生语料以及 MM 切分语料

从生语料以及两个 MM 切分语料中，我们可以分别得到每一个词 $w_i(i=1,2...,99660)$ 的词频估计。为描述方便，我们用以下符号表示：

$f_{\mathrm{FMM}}(w_i)$：从 RC_FMM 得到的每个词 w_i 的词频估计值。

$f_{BMM}(w_i)$：从 RC_BMM 得到的每个词 w_i 的词频估计值。

$f_{RAW}(w_i)$：从 RC 得到的每个词 w_i 的词频估计值。

已有的研究工作 Sun and Zhang (2006)显示，对 1-4 字词，取 $f_{\mathrm{FMM}}(w_i)$ 和 $f_{BMM}(w_i)$ 的平均值得到最好的近似结果。对 5 字词，用 $f_{BMM}(w_i)$ 得到的结果最好。而对于 5 字及以上词，串频 $f_{RAW}(w_i)$ 估计的效果最好。因此，对于不同字长的词，我们应用不同的策略进行处理。

综上，用 $F_{RFB}(w_i)$ 来表示由 RC, RC_FMM 和 RC_BMM 三个语料库得到的最终词频估计结果。则有：

对 1-4 字词：

$$F_{RFB}(w_i)=\frac{1}{2}[f_{FMM}(w_i)+f_{BMM}(w_i)] \tag{1}$$

对 5 字词：

$$F_{RFB}(w_i)=f_{BMM}(w_i) \tag{2}$$

对 6 字及 6 字以上词：

$$F_{RFB}(w_i)=f_{RAW}(w_i) \tag{3}$$

3.2 合并熟语料

目前已有两个经人工标注的中等规模的语料库 HUAYU 和 BEIDA，我们可以从中分别得到 YWL 中每个词的词频。我们记为 $f_{HUA}(w_i)$ 和 $f_{BEI}(w_i)$。对于熟语料的处理，我们运用简单地加和的方法进行处理。我们用 $F_{HB}(w_i)$ 来标记从熟语料中得到的词频估计的结果，则有：

$$F_{HB}(w_i)=f_{HUA}(w_i)+f_{BEI}(w_i) \tag{4}$$

3.3 合并 $F_{RFB}(w_i)$ 和 $F_{HB}(w_i)$

已有以上两个词频估计的结果，$F_{RFB}(w_i)$ 和 $F_{HB}(w_i)$。我们面临两个问题：一是这两个结果分别来自不同大小的语料库。语料库规模悬殊非常大，(HUAYU+NEIDA)是 17,603,085 字而生语料(RC) 是 447,079,112 字。为了平衡语料库大小，我们引入 α 因子。对于 α 的取值，直观上讲我们可以取两个语料库大小的比值。也就是 α=25.4。后面我们会在实验中给出 α 参数的调整过程。

这里我们用 C_0 表示人工标注语料(HUAYU+BEIDA)的总字数，用 C_1 表示生语料（RC）的总字数。考虑把两个语料库综合成为大小为 $2C_0$ 的一个语料库。那么生语料的规模将被压缩至 C_1/α。相应的，生语料所估计的词频 $F_{RFB}(w_i)$ 也将变为 $F'_{RFB}(w_i)$：

$$F'_{RFB}(w_i) = F_{RFB}(w_i)/\alpha \tag{5}$$

为了保持整合后整个语料库大小为$2C_0$，则最终人工标注语料库大小C_0将变成C'_0：

$$C_0' = 2C_0 - C_1/\alpha \tag{6}$$

这样由人工标注语料得到的词频估计值$F_{HB}(w_i)$也将随着变为$F'_{HB}(w_i)$：

$$F'_{HB}(w_i) = F_{HB}(w_i) \times \frac{2C_0 - C_1/\alpha}{C_0} \tag{7}$$

第二个问题是：在中文里，存在这样的现象，也就是越短的词，用熟语料估计的词频准确度越高，即更加可靠。因此，我们在这里希望增加熟语料估计值$F_{HB}(w_i)$的权重。这里引入β作为权重因子。从经验可得β的取值：

$$\beta = \begin{cases} 7 & \text{对1字词} \\ 6 & \text{对2字词} \\ 3 & \text{对3字词} \\ 0 & \text{其它} \end{cases}$$

综合考虑以上两个问题，由（5）式和（7）式可得最终由 RC, RC_FMM 和 RC_BMM 所估计的词频为：

$$F''_{RFB}(w_i) = F'_{RFB}(w_i)/(1+\beta) = F_{RFB}(w_i)/(\alpha(1+\beta)) \tag{8}$$

同样的，由熟语料 HUAYU 和 BEIDA 得到的词频估计为：

$$F''_{HB}(w_i) = F_{HB}(w_i) \times \frac{2C_0 - C_1/\alpha(1+\beta)}{C_1} \tag{9}$$

由（8）和（9）式我们可以得到最终的词频估计的折中策略：$F_{RFB+HB}(w_i)$

$$\begin{aligned} F_{RFB+HB}(w_i) &= F''_{HB}(w_i) + F''_{HB}(w_i) \\ &= F_{HB}(w_i) \times \frac{2C_0 - \frac{C1}{\alpha(1+\beta)}}{C_0} + F_{RFB}(w_i) \times \frac{1}{\alpha(1+\beta)} \\ &= F_{HB}(w_i) \times (1 + \frac{\beta}{1+\beta}) + F_{RFB}(w_i) \times \frac{1}{\alpha(1+\beta)} \end{aligned} \tag{10}$$

特殊的，对于 4 字词及 4 字以上词，（10）式退化为（11）式：

$$F_{RFB+HB}(w_i) = F_{HB}(w\ ) + F_{RFB}(w_i) \times \frac{1}{\alpha} \tag{11}$$

4. 实验与结果分析

为了进一步观察所提出的新的机制的效果，我们设计了以下实验：

4.1 Spearman 秩相关系数测试

由语委语料库统计出的标准词频，按词频由高到低，我们可以得到一个标准的词频排序序列R_{YW}。同理，我们也可以从$F_{RFB}(w_i)$，$F_{HB}(w_i)$和$F_{RFB+HB}(w_i)$中得到相应的三个序列R_{RFB}, R_{HB}和R_{RFB+HB}。我们把R_{YW}作为标准序列。保持R_{YW}里词的位置不动，可以找到每个词在这三个序列里各自的排序位置，从而得到另外三个序列，我们记为R'_{RFB}, R'_{HB}和R'_{RFB+HB}。这样我们就有四个序列$R_{YW}, R'_{RFB}, R'_{HB}$和$R'_{RFB+HB}$。把$R_{YW}$序列当作标准序列，我们可以计算序列之间的相似度。这里我们用 Spearman 秩相关系数(SRCC)来衡量序列的相似程度。SRCC 的值的计算由下式给出：

$$SRCC \equiv 1 - 6\sum \frac{d^2}{N(N^2-1)}$$

这里的 d 是相应两序列项的差值。N 是语料库大小。表 1 给出了考察所有 99，660 个词，取 α =25.4 的实验结果。

表 10. SRCC 测试结果 with α =25.4

序列	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SRCC	0.675	0.704	0.732

与其它两个序列 R'_{HB} 和 R'_{RFB} 相比，新的机制的 SRCC 值分别提高 5.7%和 2.8%。另外，用同样的测试标准，我们对因子 α 的值进行调整记为 α'。

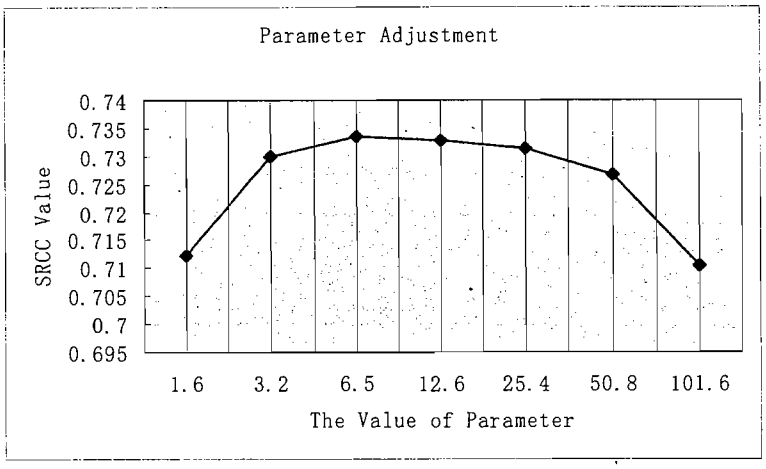

图 2.取不同 α' 值时的 SRCC 值曲线图。

图 1 显示 α'=6.5 时得到最高的 SRCC 值。所以在后面的实验中，我们取 α =6.5 作为最终的 α 值。

之后，我们选取不同频度范围的词进行测试，进一步观察新机制的性能。表 2 和表 3 分别给出了当选取词频大于 10 次以及大于 200 次时 SRCC 测试的结果：

表 11. 选取词频>=10 的词 SRCC 测试结果

Scheme	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SRCC	0.663	0.682	0.736

表 12. 选取词频>=200 的词 SRCC 测试结果

Scheme	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SRCC	0.680	0.708	0.771

以上实验比较了新机制和其它方法的 SRCC 测试结果。当我们选择不同频度范围的词时(>=4,>=10,>=200)。我们总结上面的比较结果在表 4：

表 13. 选取不同频度范围词的 SRCC 相对提高率

词的范围	词的总数目	(R'_{RFB+HB}, R'_{HB})	(R'_{RFB+HB}, R'_{RFB})
>=4	99,660	5.7%	2.8%
>=10	68,100	7.3%	5.4%
>=200	10,528	9.1%	6.3%

4.2 标准序列差 σ 测试

对每一个 YWL 里的词 w_i,都有一个标准序列号我们记为 $R_{YW}(w_i)$，以及相应的其它三个序列 $R'_{HB}(w_i), R'_{RFB}(w_i)$ 和 $R'_{RFB+HB}(w_i)$。对于一个序列 R，$\sigma = \sum |R(w_i) - R_{YW}(w_i)|$ 的值越小，则两个序列 R 和 R_{YW} 就越相似。用这个标准，我们可以测试哪一个序列与 R_{YW} 更为相似。基于这个想法，我们作了以下测试，称为 σ 测试：

用这个方法，我们可以分别得到 $\sigma_{HB}, \sigma_{RFB}$ 和 σ_{HB+RFB} 的值。$(\sigma_{HB+RFB} - \sigma_{HB})/\sigma_{HB}$ 和 $(\sigma_{HB+RFB} - \sigma_{RFB})/\sigma_{RFB}$ 分别代表了新机制得到的序列与其它序列相比较的相对下降率。表 5 给出了在不同字长情况下的实验结果：

表 14. 针对不同字长的 σ 测试

词 长	$(\sigma_{HB+RFB} - \sigma_{HB})/\sigma_{HB}$	$(\sigma_{HB+RFB} - \sigma_{RFB})/\sigma_{RFB}$	结论
1	-22.7%	-16.2%	√
2	-19.0%	-13.4%	√
3	-13.7%	-7.3%	√
4+	15.2%	17.5%	×

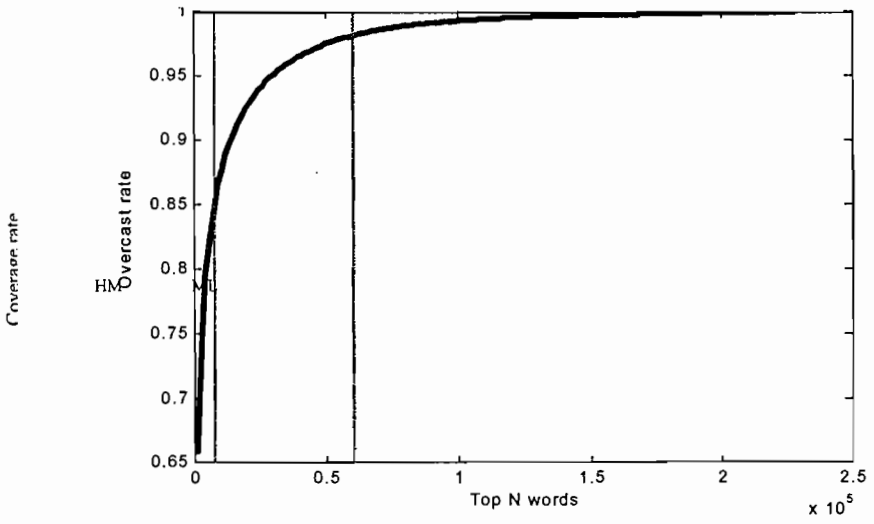

图 3. YUWEI 前 N 个词覆盖率图

从表 5 可以看到，对于所有 YWL 的词，我们的新机制在 1 到 3 字词得到了最好的结果，而在 4 字以上词，结果变差。为了进一步考察新机制的性能，我们将 YWL 里的词划分为三个部分，我们称为高频词，中频词，低频词。图 2 给出了 YWL 前 N 词覆盖率曲线。

从图 2 中我们看到，点 HM 是高和中频词的分界点。ML 是中低频词的分界点。这样，我们得到：

高频词：前 8,076 词(0~HM) 也是频度>281 的词；中频词：8,077th 到 60,224th (HM~ML) 的共 52,148 词也即

词频>12 的词；低频词：(ML~99,660) 也即词频 >3 的词。

在这一基础上，做了以下实验：

表 15. 高频σ测试值

	1 字词	2 字词	3 字词	4+ 字词
$(\sigma_{HB+RFB}-\sigma_{HB})/\sigma_{HB}$	-44.5%	-38.0%	-68.9%	-88.1%
$(\sigma_{HB+RFB}-\sigma_{RFB})/\sigma_{RFB}$	-35.9%	-31.7%	-59.0%	-81.2%
结论	√	√	√	√

表 16. 中频σ测试值

	1 字词	2 字词	3 字词	4+ 字词
$(\sigma_{HB+RFB}-\sigma_{HB})/\sigma_{HB}$	-33.0%	-14.5%	-7.5%	-13.4%
$(\sigma_{HB+RFB}-\sigma_{RFB})/\sigma_{RFB}$	-18.1%	-7.3%	-9.1%	-10.0%
结论	√	√	√	√

表 17. 低频σ测试值

	1 字词	2 字词	3 字词	4+ 字词
$(\sigma_{HB+RFB}-\sigma_{HB})/\sigma_{HB}$	27.5%	-24.0%	-17.6%	49.1%
$(\sigma_{HB+RFB}-\sigma_{RFB})/\sigma_{RFB}$	-3.1%	-10.9%	-6.2%	22.9%
结论	×	√	√	×

从以上三个表的结果，可以看到，对 1-3 字词，我们提出的新机制得到了最好的估计结果，而对于低频 1 字词及 4+ 字词，结果仍不尽人意。

另外我们观察了前 100 词σ_{HB}和σ_{HB+RFB}的比较结果，图 3 给出了对比图。其中蓝线是新机制的曲线。红线是σ_{HB}的结果曲线。黑色直线是标准值，波动幅度越小，即表示越接近标准值。

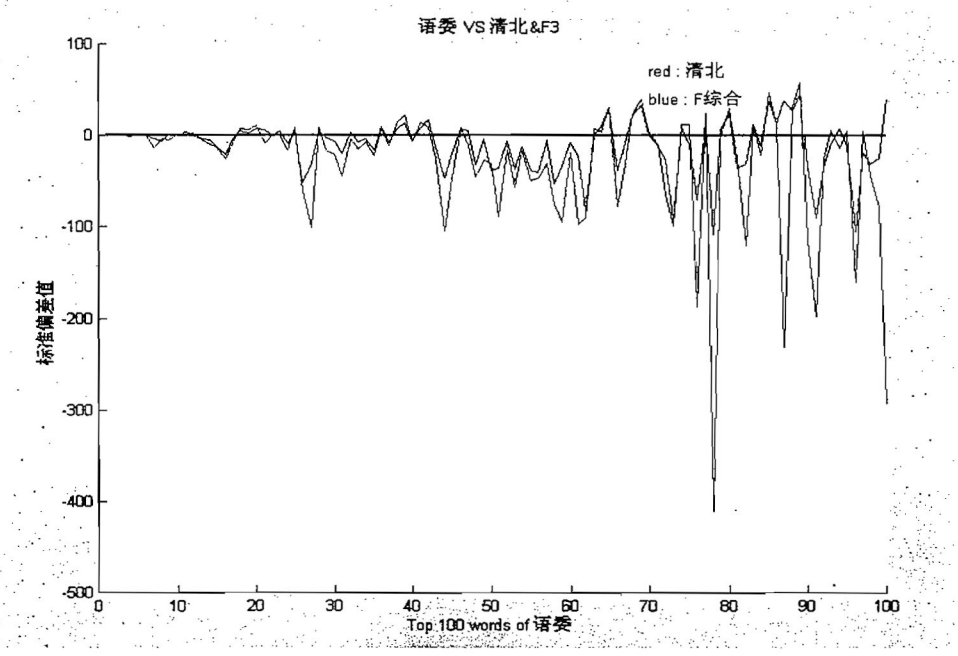

图 3. YWL 前 100 个高频词对比图

5. 总结

本文提出了一种中文词频估计的新机制，运用折中策略，实现了将生语料，熟语料，MM一切分语料结合起来进行中文词频估计的新机制。实验表明在多数情况下，该机制能够有效地提高词频估计的准确性，尤其在高频词部分表现较好。但在某些情况下，表现仍不能令人满意。今后我们将进一步考察低频词，以使得中文词频估计的性能得到提高。中文词频估计依然是一个难题，我们面前还有很长的路要走。

参考文献：

[13] Sun M.S., Zhang Zhengcao., Benjamin KYT'sou., Lu Huaming.: Word Frequency Approximation for Chinese without Using Manually Annotated Corpus. Proceeding of 7th International Conference, CICLing 2006, Mexico, 105-116, (2006)

[14] Chen G.L.: On Chinese Morphology. Xuelin Publisher, Shanghai, (1994)

[15] Dai X.L.: Chinese Morphology and its Interface with the Syntax. Ph.D Dissertation, Ohio State University, USA, (1992)

[16] Emerson T.: The Second International Chinese Word Segmentation Bakeoff. Proceedings of the Third SIHAN Workshop on Chinese Language Processing. Jeju, Korea, (2005)

[17] Liang N.Y.: CDWS: A Word Segmentation System for Written Chinese Texts. Journal of Chinese Information Processing. Vol. 1, No. 2, 44-52, (1987)

[18] Liu E.S.: Frequency Dictionary of Chinese Words. Mouton & Co N.V. Publishers, (1973)

[19] Liu K.Y.: Study on the Evaluation Technique for Word Segmentation of Contemporary Chinese. Applied Linguistics (Beijing). No. 1, 101-106, (1997)

[20] Liu Y., Liang N.Y.: Counting Word Frequencies of Contemporary Chinese – An Engineering of Chinese Processing. Journal of Chinese Information Processing. Vol. 0, No. 1, 17-25, (1986)

[21] Sproat R., Emerson T.: The First International Chinese Word Segmentation Bakeoff. Proceedings of the Second SIHAN Workshop on Chinese Language Processing. Sapporo, Japan, 133-143, (2003)

[22] Sun M.S., Shen D.Y., T'sou B.K.Y.: Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data, Proceedings of 36th ACL & 17th COLING, 1265-1271, Montreal, Canada, (1998)

[23] Sun M.S., T'sou B.K.Y.: Ambiguity Resolution in Chinese Word Segmentation. Proceedings of the 10th Pacific Asia Conference on Language, Information & Computation. Hong Kong, 121-126, (1995)

[24] Sun M.S., Wang H. J. et al.: Wordlist of Contemporary Chinese for Information Processing. Applied Linguistics (Beijing), No. 4, (2001), 84-89

[25] Tang T.C.: Chinese Morphology and Syntax: Vol. 3. Taiwan Student Publisher, Taipei, (1992)

[26] Zhu D.X.: Lectures on Grammar. The Commercial Press, Beijing, (1982)