瓶颈，挑战，与转机：中文分词研究的新思维

黄居仁
香港理工大学/台湾中央研究院
E-mail: churen.huang@inet.polyu.edu.hk

1 前言

中文切分词研究已有起码 30 年的历史。在算法的创新，在论文发表，在把中文计算语言学研究推到国际舞台等方面，都有非常好的成绩。但在解决中文语言处理瓶颈的语言工程基本要求方面，却可以说没有实质的进展。1960 年代，机器翻译的研究方兴未艾，学者们想把计算机处理应用到中文，但马上发现有预期之外的基本问题，就是中文文本并不标示词的边界。换句话说，在计算机处理前，中文文本需要先经过分词（或称断词）的步骤。在语言工程上，这个步骤造成了中文语言技术的门楣效应(英文为 ceiling)。所有中文自然语言处理工作，到这里就得低头。后续的分析与应用成绩，必须以分词成绩为上限。这对中文语言技术的发展，一直是最大的瓶颈。

欲求中文自然语言处理技术永续发展，并使中文自然语言处理技术与英语自然语言处理技术并驾齐驱，必须解除分词造成的门楣效应。换句话说，就是任何特定领域或特定语言工程工作的结果不能受分词结果的控制。实际上,就是需要能对任意文本，在没有训练语料的前提下，都能得到可靠的高水平分词结果。近年来，在各种中文分词评比与竞赛中([3], [7],[8])，最佳成绩可以到召回/准确平均分数（f-score）97-98。因此，同行们开始觉得分词的研究问题已解决，剩下的是工具化与商品化的问题。但是，实际上，这些系统的本质，还是研究系统；强调的是不计运算时间与资源成本，如何得到最高的准确与召回率。事实是最好的系统都要求大量训练语料与训练时间。没有一个系统能实时对任意文本作出可靠而可供下一步处理的分词结果。[18]李寿山等（2009）做的仿真实验显示，以目前成绩最好的 CRF 及汉字位置分类算法，使用 SIGHAN 评测数据，当训练数据源和评测数据源不同时，其结果 f-score 降低约 10 个百分点。这个结果的意义，就是说目前最好的系统的最佳成绩，必须仰赖大量同一来源语料训练。在实验室中，特别是在实验新算法时，要求测试数据与训练数据为同一来源，还算合理。但在实际应用分词时，这个前提合理吗？

现实世界中需要分词的文本，绝大部分并未带有来源，主题等后设标记（meta-data）数据。即便是文本有完备的后设数据，分词程序也不可能预先把所有可能的文本种类都训练好。更重要的，分词能贡献最高加值，发挥最高效益时，是针对新主题，新来源，带有许多未知词的文本。但是，这种文本，当然不可能有大量已有黄金标准的标记语料可以用来训练。也就是说，在各项分词评比得到好成绩的系统，其分词结果与目标文本及训练文本间的相似度有绝对关系。但是在实际应用上，强健系统不应该受限于训练文本；而且分词的应用价值，与目标文本与训练文本的相似度，有负相关性。总言之，分词成绩受限于训练语料类型与规模的分词模式，是目前的研究

主流，却是与实用分词的需求背道而驰。

研究超过30年，在研究与评比成绩上有具体的进展，但是面对语言科技的实际应用的基本挑战，则尚未突破。这是当今中文分词研究的瓶颈。面对瓶颈，应该是重新思考研究模式，寻求转机的恰当时间。本文将检讨分词研究的可能发展模式，并提出更具实用意义的分词评比模式。

2 分词运算模式的演进与前景

由运算模式的观点， 分词研究经过几个重要的演变阶段。早期的运算模式简单，运算需求低，但成绩也不理想。随着软硬件的进步，运算量与训练语料提高，成绩也有进步。这时的研究发现，其实改变运算模式，降低运算量，可以提升成绩。因此分词运算模式的选择，成为重要研究议题。这是目前研究面对的问题。

2.1. 序列匹配算法

计算机自动分词研究开始时，受到有限状态自动机（finite-state automata），以及乔姆斯基阶层（Chomsky Hierarchy）对语句定义模式的影响大部分的研究都把分词当成是输入字符串的序列匹配问题。乔姆斯基及早期的有限状态机理论把句子剖析（合法度判断）看成是如何正确走到终止状态（final state）。因此分词也可看成把字符串的终止状态。当然，因为词的长度太短，大部分的作法，还是把句点当成终止状态，而字符串辨识为词，当成前进到下一状态(shift)的条件。换句话说，这时的运算，输入字符串，输出标有词终点标志的字符串。这个阶段的研究议题，是使否采用长词优先，还是双字词优先；字符串阅读应该是由左到右，还是由右到左；是否可以预览（look-ahead）等等。由于这是研究的萌芽阶段，并没有可靠的评测标准，因此没有留下可与目前研究比较的成绩。现在回顾，这个运算法是相当简洁有效率的。只是当自然语言研究范典（paradigm）转到统计运算法时，并不相容，很自然的被遗忘了。这个算法，另一个重要的缺点，所有决定都是线性局部的，相同的局部字符串只可有一个固定解，无法有效解决如抢词等歧义的问题

这个运算法可以和辞典结合，如[11]把辞典转换成带权重的有限状态转换器（finite-state transducer）。有限状态自动转换器是自然语言处理方法中，各语言处理构词最有效的算法..但这个方法由于运算量大，对辞典质量与涵盖率的依存度太高。换句话说，如果目标文本来源或主题等差异较大时，不容易有强健性调整，会影响成绩。

2.2. 辞典匹配算法

1980年代以后，中文电子辞典构建完成，如台湾中研院的CKIP辞典，分词研究转换成以辞典为本的研究。这些研究中,如[1]，分词的最重要模块是辞典（dictionary lookup）查找模块。分词的成绩，取决于两大因素，辞典的涵盖率是否够高，发生抢词的问题时，是否有好的解法，以及是否能有效解决未知词问题等。这是规律式分词的主要模式。辞典查找虽然看来是很简单的运算，但其实包含了约十万个不同字符串的比对；因此有效的归纳成规律比对，而非序列比对。非常重要。这个算法的另一个瓶颈，就是未登录词（OOV, Out-of-vocabulary word）的分词，因

为没有辞典可以参考，成绩会比较差。而且处理非登录词与歧义抢词现象的规律，不但要费很多人力撰写，而且更多新规律的增加，不一定能提高系统的成绩，有时反而会因与现有规律冲突，而降低成绩。也就是说，这个运算法不能保证渐进改善（incremental）。

2.3.统计算法-互见讯息与 N-连字

统计法分词与辞典查找法分词几乎同时发展。统计分词法有两个基本理论模式，一个方向是利用字符串中字与字间结合紧密度，来判定哪些字符串应该结合成词。这个研究方向由词的定义出发(如[4], [10])，使用的基本统计工具是互见讯息（MI）。这个方法的好处，是不需要辞典，对辞典登录词与未登录词的解法相同。但对于长于二字的词的处理延伸，没有很有效的方法。

另一个方向是计算字符串本身在语料中分布的强度，包括重复出现的机率其出现的环境，来判定该字符串是否该被分成词。这个模式，主要是利用连字的分布，找出最可能成词的 N—连字(N-gram)。这类模式，如[2]的优点，是可以利用标记语料库做训练，训练结果适用于已登录与未登录词；而且统计模式对同一字符串在不同语境中可以得到不同分析，也就是说对歧义字符串可以有两个以上的分析。更重要的是，当目标文本的特性改变，也可以藉不同训练语料来调整，提升好成绩。在第一届 SIGHAN 分词评比中[3]，大部分的系统，采用的是 N—连字的统计模式。

2.4. 字分类算法与 CRF 机器学习

[15]提出的字分类算法，与[13]采用的 CRF 机器学习运算模式，近年来蔚为中文分词研究的主流。近来中文分词评比([8],[9])，成绩最好的，大概都是采用这种算法(如[13],[17])。这个算法的基本动机，在于降低预算模型的复杂度，从而提升运算效率与成绩，并降低对训练语料的依存度。这个算法的核心概念，是把分词由词的辨识与分类问题，转换成字的辨识与分类问题。[15]最早把训练语料中的每个字加上位置标记：如 LL 代表该字是多字词的开始的第一个字，RR 表示该字是多字词结束的最后一个字，LR 表示该字是单字词，MM 表示是多字词中不在首尾的任何其他位置等。这个转换，把分词由十万以上不同词种（word-type）的分类问题，化简成了将五千余汉字标记为四个不同类的简单标记与分类问题。[13]主张采用的 CRF 机器学习算法，则提供了运算能力高的适当分类器来解上述分类问题。

这个模式最明显的问题，就是训练语料量的要求非常高，大概要几百万字以上。训练时间，也通常需要几个小时。对于任意新的，而且没有确定来源或主题分类标记的文本，不容易马上适用。

2.5. 词界判定(WBD, word boundary decision WBD)模式

我们如果回到人类的分词能力，基本上是在语言处理的最早阶段进行的。理论上是先有分词的动作，分成词的结果才能用于查询心理辞典，提取词汇讯息。换句话说，分词的能力，必须独立且不能依附于辞典讯息。以上的几个分词算法，除了最早的互见讯息法外，都需要辞典讯息。就连字分类法，虽然不直接引用词表，但事实训练语料的位置标记，是在已知某些字符串为词的前提下完成的。也就是说，已参考到一个所有词的清单。但互见讯息算法，不能得到符合目前语言技术要求的准确分词结果。而任何引用辞典知识的算法，又难免受辞典与其领域的影响，对未

知词与未知领域处理时，成绩会降低。

[5]提出了一个完全不用辞典讯息的分词算法，称之为词界判定(WBD, word boundary decision WBD)模式。顾名思义，这个算法把分词简化成判定字与字的边界是否同时为词的边界，一个简单的二选一问题。这个算法把分词的标的，看成是字(c)与字间空白(I)交夹的字符串：

$$c_1,\ I_1,\ c_2,\ I_2,\ \ldots,\ c_{n-1},\ I_{n-1},\ c_n$$

而分词的目标，简化成判定字间空白 I 是否是词界。字只被视为视作这个分类判断时需要的语境讯息。这个模式把分类简化成二，把分类标的与环境分开，大大的降低了分词需要的运算量。[5]经过实验后，提出判定最有效的环境，是个五维向量。对任何处在 abIcd 字符串中的字间空白 I，取两个五维向量 <ab, b, bc, c, cd, 1> 与<ab, b, bc, c, cd, 0>，分别表示在这些语境中，I 被判定为词界或非词界的机率。这些机率，可由训练语料事先计算。[6]的实验显示，使用不同的机器学习分类器，把上述向量作为训练数据，只要随机选取 1000 个向量，就可以把分类成绩优化。[19]的实验则在相同运算环境下测试，发现 CRF 算法需要 1-2 小时的训练时间，WBD 只要 2-3 分钟。更重要的是，当分词程序遇见不熟悉文本时，WBD 可以采用[19]提出的以标的文本作训练与调适的算法，提高分词成绩。CRF 算法则必须有大量训练语料才能调整。

3. 分词运算与分词评测的未来发展

上一节的讨论中显示，分词运算法的理论演变，是由规律法到统计法，而且是由一般统计算法到机器学习。然而，在机器学习的运算模式中，最近的两个创新，都是要把分类问题简化，以降低训练语料量，以加快机器学习速度，提高成绩。但是近 30 年来的研究，并没有真正改善中文自然语言与信息处理最基本的瓶颈问题。中文自然语言应用在分词的第一步就远远落后英文或其他先进语言，在技术尚不能达到大规模应用。其基本原因就是没有在应用上可行(不需大量训练，不论任何来源或文体均有强健性结果)的切分词工具。如何面对这哥挑战，突破瓶颈呢？

我认为分词分类理论，与在线调适（online adaption）的技术结合，是必然的趋势。但是更重要的，目前习用的分词评比方式，必须改弦更张。目前的评比的方式，是固定训练语料，保留小部分作评测语料。结果是鼓励「过优化」（over-fitting）,而不考虑同一系统，处理不同类语料时的成绩。但是在分词实际运用中，就是需要处理不熟悉，没有大量已标记语料的文本。如果要真正找出对中文分词技术有实用强健性的系统，评测应该是模拟实用状态。也就是说，不提供训练语料，只提供 10-20 个不同来源，文体，主题的短文本。然后评测那个系统可以不用学习调适，或经过在线调适，在这些差异性大的文本中，得到最佳结果。我们也可以考虑评测处理时间，以评估系统的可行与实用性。

4. 结语

中文自然语言研究的愿景，是希望中文语言工程技术能达到和英文语言与知识工程相比或甚更高的水平。但是作为入门瓶颈的分词技术，一直未能达到实用阶段。这是中文自然语言处理

研究的危机。本文由这个问题研究演进过程探讨，提出几个可能的方向，以就教于大家。

参考文献

[1] Chen, Keh-jiann and Shing-huan Liu. 1992. Word Identification for Chinese Sentences. Proceedings of COILING 1992. 101-107.

[2] Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su. 1992. Statistical Models for Word Segmentation and Unknown Word Resolution. In Proceedings of Rocling V, pages 123–146, Taiwan.

[3] Emerson, Tom. 2005. The Second International Chinese Word Segmentation. Bakeoff. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 123–133, Jeju Island, Korea.

[4] Huang, Chu-Ren. 1995. The Morpho-Lexical Meaning of Mutual Information: A Corpus-based Approach Towards a Definition of Mandarin Words. Presented at the 1995 Linguistic Society of America Annual Meeting, New Orleans.

[5] Huang, Chu-Ren, Petr Simon, Shu-Kai Hsieh, and Laurent Prevot. 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In Proceedings of the Association of Computational Linguistics Annual Meeting, pages 69–72, Prague, Czech.

[6] Huang, Chu-Ren, Ting-Shuo You, Petr Simon, and Shu-Kai Hsieh. 2008. A Realistic and Robust Model for Chinese Word Segmentation. In Proceedings of ROCLING-2008, Taiwan.

[7] Jin, Guangjin and Xiao Chen. 2008. The Fourth International Chinese Word Segmentation Bakeoff. In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, Hyderabad,India.

[8] Levow, Gina-Anne. 2006. The Third International Chinese Word Segmentation Bakeoff: Word Segmentation and Named Entity Recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia.

[9] Low, Jin Kiat, Hwee Tou Ng, and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 161–164, Jeju Island, Korea.

[10] Sproat, R. and C. L. Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages, 4(4):336–351.

[11] Sproat, R., Chilin Shih, William Gale, and Nancy Chang. 1996. A Stochastic Finite-State Word-segmentation Algorithm for Chinese. Computational Linguistics, 22(3):377–404.

[12] Sproat, Richard. 2000. A Computational Theory of Writing Systems. Studies in Natural Language Processing, Cambridge University Press.

[13] Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Chris Manning. 2005. A Conditional Random Field Word Segmentater. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea.

[14] Wu, Andi and Zixin Jiang. 1998. Statistically Enhanced New Word Identification in a Rule-based Chinese System. In Proceedings of the Second Workshop on Chinese Language Processing, pages 46–51, Hong Kong, China.

[15] Xue, Nianwen. 2003. Chinese Word Segmentation as Character Tagging. International Journal of Computational Linguistics and Chinese Language Processing, 8(1):29–48.

[16] Zhao, Hai, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of the 20th Pacific Asia Conference on

Language, Information and Computation (PACLIC-20), pages 87–94, Wuhan, China.

[17] Zhao, Hai and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, pages 106–111, Hyderabad, India.

[18] Li, Shoushan and Huang, Chu-Ren. 2009. 基于词边界分类的中文分词方法. To be presented in Proceedings of CNCCL-2009.