机器学习的兴起
现代方法

- 最大熵用于词性标注
  - Adwait Ratnaparkhi, Jeffrey C. Reynar, Salim Roukos. A Maximum Entropy Model for Prepositional Phrase Attachment. HLT 1994

- 机器翻译
  - Franz Josef Och, Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In ACL 2002: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (**best paper award**), pp. 295-302, Philadelphia, PA, July 2002.

- CoNLL
  - The Conference on Computational Language Learning (CoNLL-97) was held on July 11, 1997 in Madrid, Spain.
  - Named entity recognition, chunking, semantic role labeling, dependency parsing, joint learning of syntactic and semantic dependencies, etc

机器学习方法的兴起
中文处理

- Bakeoff-1: 2003
  - 分词
- Bakeoff-2: 2005
  - 分词，统一的机器学习方法
- Bakeoff-3: 2006
  - 分词，命名实体识别
- Bakeoff-4: 2007,2008
  - 分词，命名实体识别，词性标注

为什么要机器学习

- 样本比规则好定义
- 规则会忽略低频情形
- 语言的解释涉及的因素过多

– Fernando Pereira

– Machine Learning in Natural Language Processing

– University of Pennsylvania

– NASSLLI, June 2002

为什么要机器学习

- 机器学习降低了知识表示的难度！

机器学习方法的特征

- 标注数据：语料
  - 知识表示
- 学习方法
  - 知识获取

机器学习方法的特征

- 机器学习针对于传统的人工智能。
  - 知识表示和获取的分离
  - 语料构建：专注于知识表示
  - 机器学习：专注于知识获取

- 对比：专家系统
  - 规则的获取和表示是同步的。
  - 规则的管理是低效率的，困难的。

机器学习和知识源

- 从知识工程看待机器学习
  - 规则1
    - 学习模型本身/特征体系
  - 规则2-n
    - 标注语料

学习模型

- 学习模型的三要素
  - 目标函数：知识源
  - 特征体系：部分的知识源
  - 参数估计算法：与知识源基本无关

机器学习：数据

- 假定已有数据合理近似现实世界？
- 拥有数据
  - 训练数据集（training set data）：训练
  - 测试数据（testing data）：评估
  - 验证集[validation set]：避免过拟合[overfitting]。
- 真实数据（real data）：最终的检验

学习模型并不重要
定理：没有免费的午餐

- 结论描述 by David Wolpert and William G. Macready
  - 由于对所有可能函数的相互补偿，最优化算法的性能是等价的。
- 没有其它任何算法能够比搜索空间的线性列举或者纯随机搜索算法更优。
- 该定理只是定义在有限的搜索空间，对无限搜索空间结论是否成立尚不清楚。

- 参考文献

  - Wolpert, D.H., Macready, W.G. (1995), No Free Lunch Theorems for Search, Technical Report SFI-TR-95-02-010 (Santa Fe Institute).
  - Wolpert, David (1996), ""The Lack of A Priori Distinctions between Learning Algorithms," *Neural Computation*, pp. 1341-1390.
  - Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation* **1**, 67.

算法的人工倾向

- 任何学习算法都需要一些“倾向性”，用来区分可能的结果。
- 回到知识源的观点
  - 学习模型的三要素
    - 目标函数：知识源
    - 特征体系：部分的知识源
    - 参数估计算法：与知识源基本无关

不拒绝个别优化

- 机器学习的最优是依赖于案例特性的！
- 算法可能特别适应于某个特定任务
- 存在一般的优越算法吗？
  - 不存在

学习模型 vs. 特征工程

- 我们给出的一个没有免费午餐定理的直观的强化描述
  - **给定任何一个学习模型，如果进行充分的特征工程，则在此意义下，没有一个学习模型能够给出更优的性能。**

- 举例：
  - 我们在依存句法分析上的实践
    - Nivre验证SVM提供了最强的性能
    - 我们用最大熵在同样的学习框架下给出了更强的结果。
    - 而通常认为SVM这样的边界最大化分类器优于最大熵。
  - 我们在语义依存分析上的实践
    - 我们同行用联合学习模型,我们使用纯粹的特征工程。
    - CoNLL-2009评测结果：我们在SRL项目总分第一。

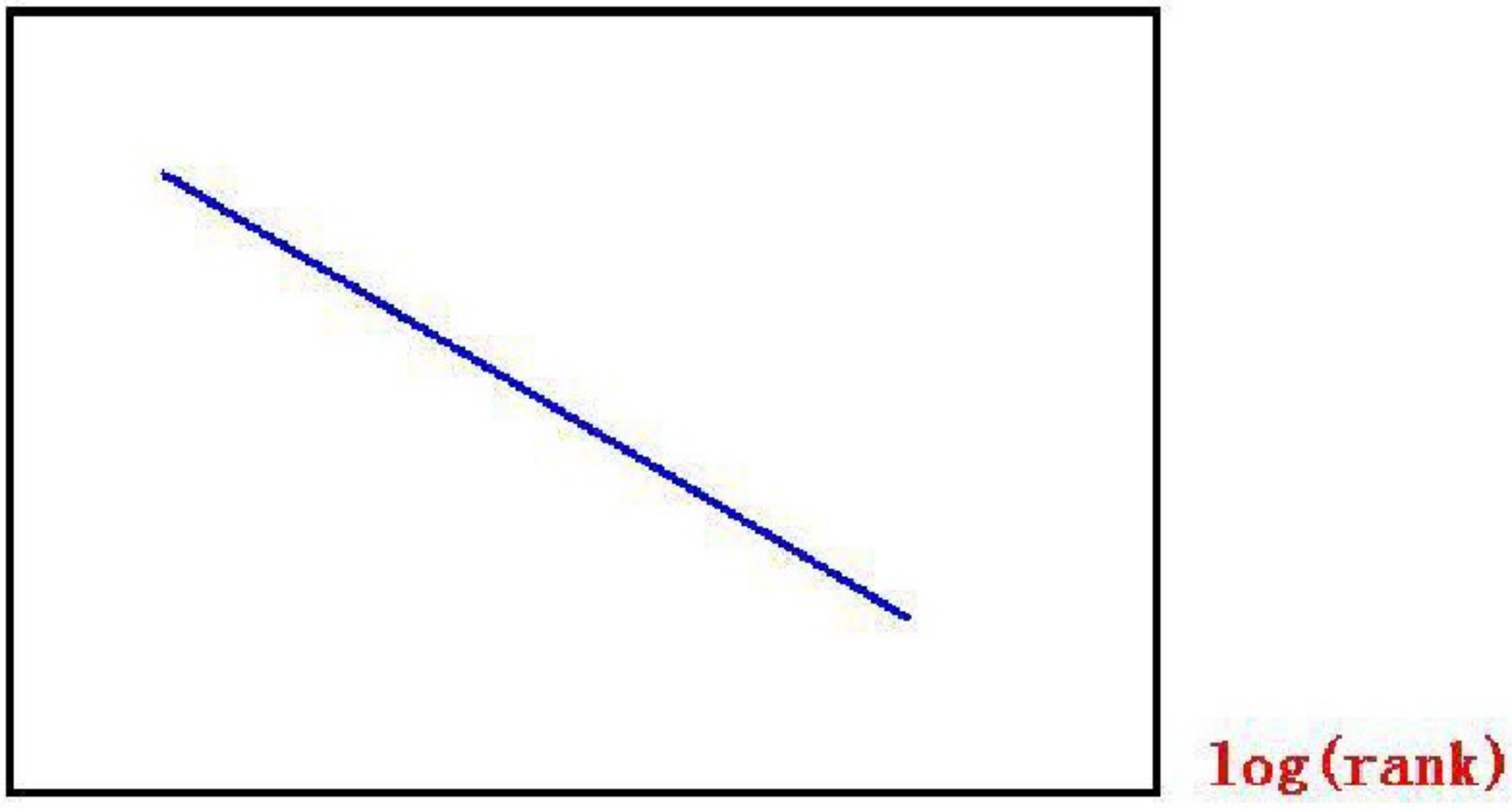
丑小鸭原理

- 20世纪60年代美籍日裔模式识别专家渡边慧证明了“丑小鸭定理”。该定理认为“丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大”。
- 世 界上不存在分类的客观标准，一切分类标准都是**主观**的。
- 渡边慧举了鲸鱼的例子说明该定理：按照生物学分类方法，鲸鱼属于哺乳类偶蹄目，和牛是一 类；但在产业界，捕鲸与捕鱼都要行船出海，鲸和鱼同属水产业，而不属于包括牛的畜牧业。
- 分类结果取决于选择什么特征作为分类标准，而特征的选择又依存于人 的目的或价值观。
- 丑小鸭是白天鹅的幼雏，在画家眼里，丑小鸭和白天鹅的区别大于两只白天鹅的区别；但在遗传学家眼里，丑小鸭与其父亲或母亲的差别小于父母 之间的差别。

- 参考文献
  - Watanabe, Satosi (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: Wiley. pp. 376–377.

Zipf's Law

- 数据稀疏的严重性



严重的问题

- 统计方法如何克服严重的稀疏性？
  - 不断增大标注数据

研究者的通常做法

- 忙于把各种最新的机器学习方法移植到所有的自然语言处理任务上，并企图证明某个最新机器学习模型的移植是最有效的。
  - 忘了没有免费的午餐？
- 少有人考虑特征工程/语料构建

不能脱离人的主观性的机器学习

- 小结：
  - 从语料中自动获得表达知识的规则
  - 依赖于人的主观定义下的启发式规则确定特征和目标函数
  - 知识的流动：从语料到学习获得的模型
  - 大量的标注数据的获得并不容易，但是必须

一个简单的机器学习任务：中文分词

- Bakeoff切分语料
- CRF学习模型
- 字标注框架

分词信息的知识源

- 切分语料
  - 辅助切分器作为特征
- 词典
  - 最大匹配结果作为特征

- 参考文献
  - Hai Zhao, Chang-Ning Huang, Mu Li (2006). An Improved Chinese Word Segmentation System with Conditional Random Field, SIGHAN-2006
  - Low, Jin Kiat, & Ng, Hwee Tou, & Guo, Wenyuan (2005). A Maximum Entropy Approach to Chinese Word Segmentation, SIGHAN-2005

数据

Bakeoff-2006	AS	CityU	CTB	MSRA
Training(M)	8.44	2.71	0.83	2.17
Test(K)	146.3	364.5	256.5	172.6

- **方法**
  - 字标注学习方法 CRFs
  - 前向最大匹配算法

- 参考文献
  - Hai Zhao, Yan Song and Chunyu Kit, How Large a Corpus do We Need: Statistical Method vs. Rule-based Method, LREC-2010

实验1: 统计方法给出的结果

等效于有效知识源的扩大

- 辅助分类器

	A	B	C	D	E	F
AS	MSRSeg	MSRSegNE	MSRA2005	PKU2003	PKU2005	CTB2006
	G	H	I	J	K	L
AS	AS2003	AS2005	CityU2003	CityU2005	CityU2006	AS2006

实验1: 结果

- CTB2006 MSRA2006

baseline	+Ext.Dict	+C	+D+E	+G+H	+I+J+K	+A	+B(Final)
0.927	0.9423	0.9468	0.9475	0.9515	0.9518	0.9522	0.9531

Baseline	+Ext.Dict.	+E+G+H+K	+A	+B(Final)	+C
0.961	0.9694	0.9704	0.9823	0.9826	0.9702

实验1: 为什么附加语料提升性能

- 作为机器学习的解释
  - 学习模型记住了引入的新的字搭配模式，改进了*Foov*
  - 我们需要多少附加语料？
- 有效知识源扩大：只要知识源规模不断扩大，性能就能提升？
- 学习模型的贡献在哪里？

实验1: 谨慎的结论

- 开放测试问题是否可以转换为一个单一的可供集成的语言资源的扩大。
  - 我们部分做到了这一点!

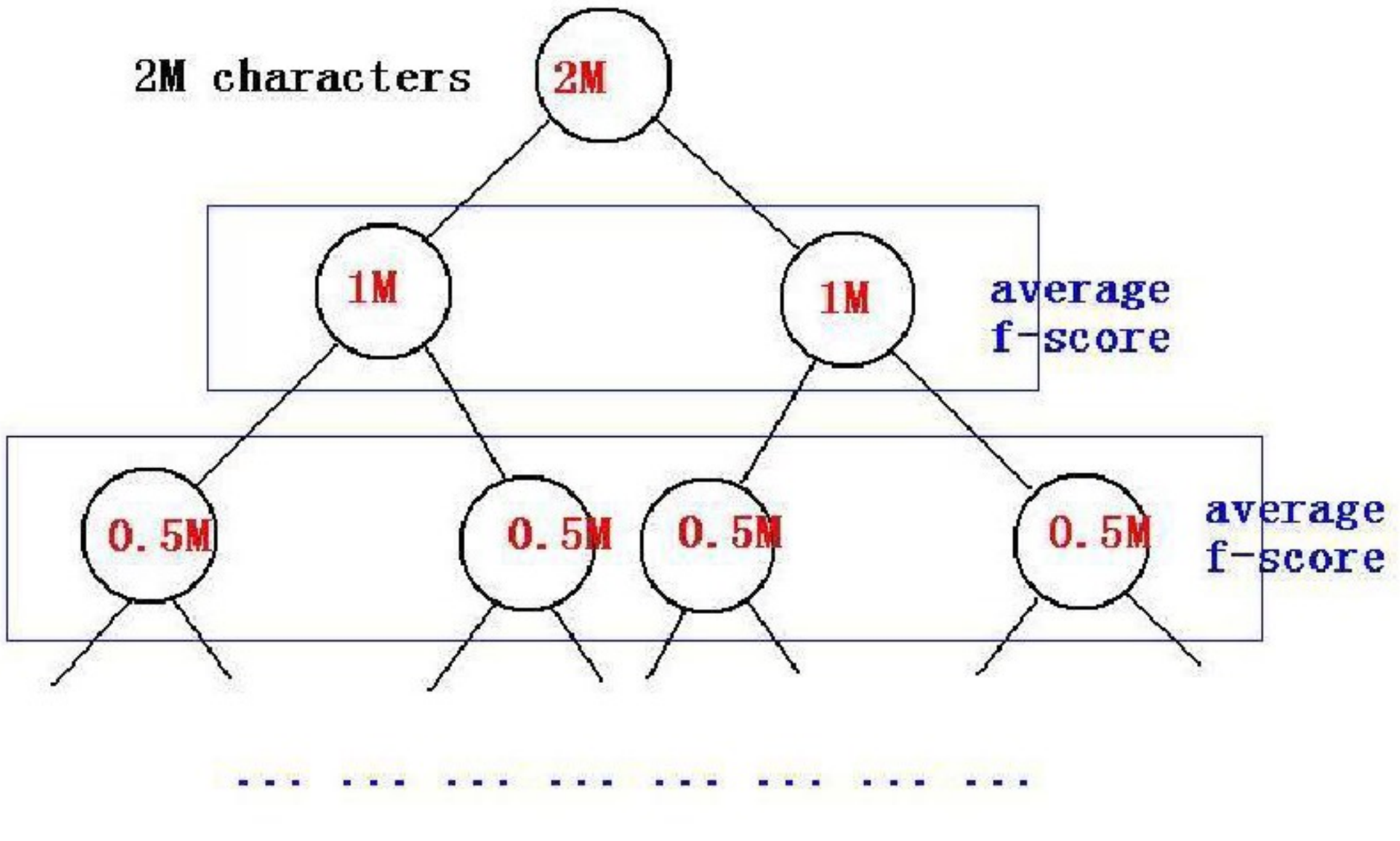
- 机器学习模型的贡献有限。

实验2: 评估语料规模对性能的影响

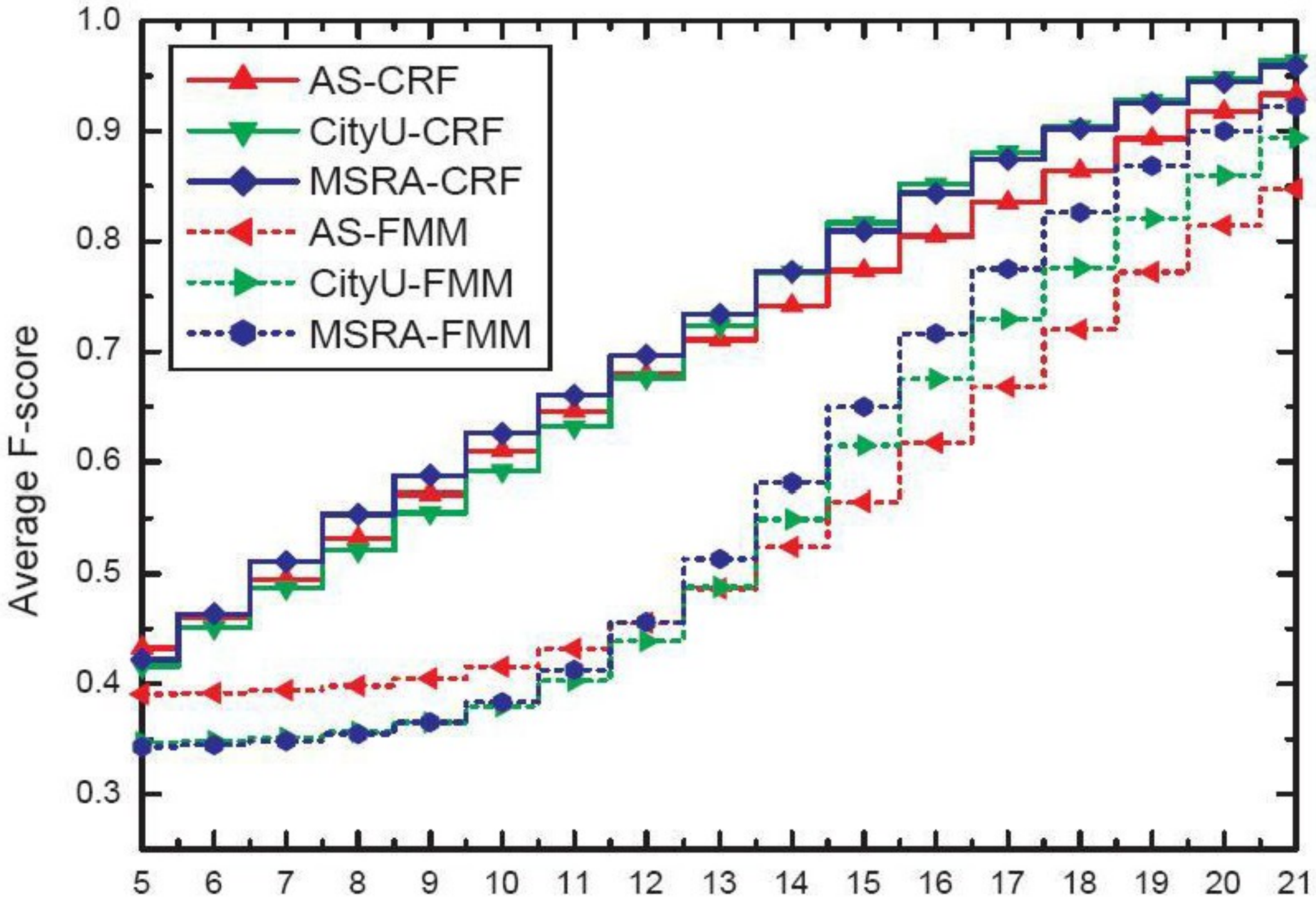
- 如果
  - **语料规模**是唯一影响性能的因素，
- 那么
  - 对于一个特定的性能度量要多大规模的语料来学习?

实验2: 数据划分

- 是用平均化策略克服过小数据集的数据稀疏性



实验2: 学习曲线: CRFs vs. FMM



实验2: CRFs 性能vs语料规模

指数增长的语料带来线形性能提升