**1 Introduction**    针对无监督中文分词，前人做了很多工作，但是大部分包含多种启发式的规则，而且缺乏在大规模语料库中形成一种可比较的标准。我们将词汇提取看做中文分词的一种特殊情况，他们之间有一些区别：分词必须包含分词算法，而词汇提取目标则是提取出一个词汇表就好了。

**2 Generalized Framework**    本文提供了一种一般化的框架，其中包括：一个词串成词的可能性；分词算法。C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\a4c0d3375a7945caab1f877d05d08199\clipboard.png 表示每个词一个分数，该分数是通过不同的指标计算得来。分词算法则包括两种：维特比算法以及另一种动态规划的算法。有人认为算法二是算法一的特殊情形，但是算法二也有后向版本。

**3 Goodness Measurement**    前人的工作有基于互信息（MI）判断一个字串是否成词的，原则上，我们计算的各个指标应该遵循一个基本的规律：某字串的指标越大，则其成为一个词的概率则越大。

    然后本文列举了以下四种指标用来评估一个字串独立成词的概率大小。

1.字串下降频率（FSR）。

C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\d6780282851248ef9963be7650e90309\clipboard.png，其中C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\3616534505b942509da0a8c47af8047b\clipboard.png代表一个字串w的频率，且只考虑出现次数大于0次的字串

2.描述长度增益(DLG)

    他的主要思想为，将语料中的某个字串替换为一个字符，然后计算替换前和替换后描述长度（DL）的增加情况，称为描述长度增益（DLG）

C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\78e13bdbc644499b80c2e4757698ad88\clipboard.png，其中正的信息增益值的字串才会被加入到候选词汇列表中。

3.抵达多样性（AV）

    该指标仅考虑一个词汇左右两侧的不重复分支词汇的数目，且AV定义为左右两侧分支词汇数目较少的一个。

C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\d192cea91c644d12b30c6a04d159d731\clipboard.png

4.分支熵（BE）

    与我们定义的左右分支熵一致，左右分支熵的融合，本文采用的是取左右分支熵中的较小值。且只考虑分支熵大于0的词汇作为候选词汇。

**4 Evaluation**  
    本文采用的评估指标为F值，而且是基于Bakeoff-3的四种标注语料进行测试。然而，分词算法会遇到某些评估方法无法计算一元词汇的指标，所以提出了一种解决方法，即为将无法计算的指标设置为某一个默认值，未登录词的平滑方法也是采用这种方式。

**4.1 Comparison**    本文分别应用上面所说的四个指标，应用“Viterbi分词算法”在四种语料库上进行了测试，而且考虑了词语的最大长度这个因素。结果显示DLG的效果最好，FSR效果最差，而AV和BE的效果相当。

    然后本文又将AV和BE应用不同的分词算法，不同的词汇最大长度进行了比较。结论为：AV在短词汇上面表现较好，而在长词汇上表现则不如BE。

    然后本文还以词汇长度为自变量，将DLG和BE进行了对比，找出了二者表现最好时的词汇长度。

    这部分工作，我认为作者的目的在于找到每种指标最适合的切分情形，方便后面进行指标的融合时考虑这些因素。

**4.2 Word Candidate Pruning**  
     因为每种指标生成的候选词的数目差别比较大，但是现在默认的裁剪阈值（即某指标大于该值才会被保留加入词汇表中）一般为0，这样每个指标的候选词表大小不一，该部分所做的工作就是为每个指标确定一个合适的裁剪率，将分数较低的词汇剔除出候选词表。这样做的目的就是提升整体候选词表的质量，而不仅仅是考虑每个词汇的成词的概率大小。

    本文做了几个实验，分别尝试不同的裁剪率，然后观察整体F值的变化，最终得到了每个指标的最优的裁剪率，且经过对比，效果确实有显著提升。且基本规律不变，即DLG在二元词汇上表现较好，而AV/BE在更高元数据上表现较好。

**4.3 Ensemble Segmentation**  
    上面的方法也是有局限性的，在缺少标注语料的非监督学习过程中，无法获得每个指标的最优裁剪率，这样就导致每个指标获得的词汇表规模不一，而词汇表的质量对分词结果有很重要的影响。本文提出了一种提高单词表质量的方法，即对上述四种指标得出的单词表取交集，作为最终切分算法的词汇表。本文在四种指标上做了实验，证明该方法优于原先的单一指标阈值取默认值的方法，且算法一优于算法二，DLG仍是表现最好的指标。

    最后本文又将DLG的切分结果分别和AV、BE的切分结果进行了融合，做了实验比对。文章中说是将AV/BE的切分结果强加到了DLG的切分结果上（我的理解是：比如说这两种指标分别对一个句子切分，在一个切分点，DLG切开了，而BE没有切开，则这个切分点的结果遵从BE的结果，即为“不切分”）

**4.4 Yet Another Decoding Algorithm**

    本段介绍了另一种切分算法，基本原理类似老师当初跟我说的意思，例如“澳大利亚”分支熵一定大于"澳大利"，若是遇到这种情况，则需要在“澳大利亚”后面进行切分。公式表达则为C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\38e81de3a5bd407e8896464f14c03023\clipboard.png时，在C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\0c3c55b87c2543baaad90323896b2875\clipboard.png后面切分。这个算法有前向和后向两种方式，分别进行，最终将切分结果合在一起即为最终结果。

    然后本文基于《人民日报》的语料库和测试集进行了测试，词汇的元数最高考虑到了6元，分别针对词汇的召回率和边界的召回率进行分析，结果表明，边界召回率较高，且算法3效果并不好。

**4.5 Comparison against Supervised Segmentation**  
    本文中评价非监督学习系统性能的方法，就是黄昌宁和赵海提出的“各自训练，交叉检验”，利用不同语料之间的词汇认同度评价系统的性能。系统上界和下界的定义同上一篇论文定义相同。另外有一点，系统根据指标得分提取词汇表时，不光是提取训练语料，还包括测试语料。

    本文通过实验对比了各个分词指标以及融合指标，分别利用不同的分词算法下的效果，虽然无监督的学习方法相对有监督方法稍逊一筹，但是在没有任何先验知识的条件下，仍能达到0.65--0.7的F-score。

**5 Discussion: How Things Happen**

      本文分析了DLG和AV/BE的不同。DLG本质是压缩编码，目的是获得最大的压缩效率，考虑的是整个语料，而且会有“长词歧视(bias)”的问题。而AV/BE则不同，他只是考虑一个词串周围局部的情况，而不是全局，所以适合采用贪心法求解，而DLG则适合采用Viterbi算法。

    这部分从根源上解释了前面的实验现象。

**6 Conclusion and Future Work**  
    该部分总结了上面的各大部分，在将来的工作方面他建议其他人采用他的指标去进行有监督的中文分词。

论文特色或借鉴：

1.本文不光提出了四种指标，而且考虑了候选词阈值的问题，而且证明了这个因素对分词效果有很大影响。而我们的工作未考虑这个问题；

2.在无监督学习时无法确定最优的阈值情况下，提出了取各个指标的候选词交集作为最终候选词表的方法，提升了整体候选词表的质量；

3.提出了两种分词算法，而且他们对于不同的分词指标，效果各不相同；

4.论文实际只是融合了两个指标，而且融合方法一句带过，不知道前面我的理解对不对。