**2 State of the Art**  
分支熵应该随着元数的增加而逐渐变小，但是从某些情况看，并非如此。以往基于分支熵的分词系统，切分与否的主要依据是设定好的阈值，且会考虑分支熵的变化趋势，但这种方法阈值的设定需要训练语料的协助，而且会涉及一些参数的训练调优等。先前部分基于分支熵的工作也未曾考虑归一化的问题，导致不同gram的分支熵不具有可比性。

**3.Evaluation**

        评估无监督学习系统是个比较困难的事情，在有监督学习系统中，为了能够评估系统的性能，消除不同分词标准带来的影响，黄昌宁和赵海提出了一种“各自训练，交叉检验”的方式，即在各自的训练集训练好模型以后，分别用该模型去切分四种测试集，最终将在其他测试集获得的F-score与自身测试集的F-score相除，该比值就代表不同分词系统之间的词语认同率。他们二人做的实验中，这个认同率最低为0.85，所以本文就将这个0.85作为了无监督学习评价指标的上限，而他的下界0.35我没看懂是怎么得出的0.35。

        我认为本文对无监督学习采用的评价方法，也是这个“不同分词系统之间的词语认同率”，但是文章中提到了per word-length evalution，我的理解是因为该评价指标依赖的是F值（F值为准确率和召回率的算术平均值），这个F值是不是基于单个词汇计算的，所以称为per word-length evalution。

        本文分别用一元，二元和三元进行了无监督的实验，结果表明一元和二元评估指标较好，且二元指标评估结果甚至优于一元，但是基于三元指标的实验效果不佳，本文分析原因，因为一元词和二元词是高频词汇，但是数量却远远少于三元词，三元词虽多但是出现频率很低，故而三元词对系统的影响也非常有限。

        所谓无监督学习，并不是完全的无监督，并不排斥进行一些初步的预处理工作，比如中文语料中的标点，拉丁字符以及阿拉伯数字等。

**4 Normalized Variation of Branching** **Entropy (nVBE)**

论文中的分支熵计算方法与我们的方法一致，不过他是在未切分语料上进行的各项统计，而我们的目前所做的统计工作和分支熵等指标计算，都是在经过中科院分词的语料上进行的。标准化方法（以右分支熵标准化方法为例）C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\c21263deb8ed43aa99e5fab12badc680\clipboard.png

其中的C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\7cb7b655060e4ae49ac5260869c42b8b\clipboard.png为原始分支熵；C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\fd32b63f65c248eea9d4e64e506dfd12\clipboard.png为与x字串长度相等的所有字串的分支熵（右分支熵）的平均值，所以说这个值可能为负数，而且采用这种标准化方式就不再遵循Harris的理论，即C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\ee1707125e284de3b7d8aa2129b9aafa\clipboard.png这个等式并不一定永远成立。以往仅仅基于原始分支熵数据进行分词效果不佳，采用该标准化方法后，基于分支熵的分词系统取得了较大的提升。

**5 Decoding algorithm**  
        通过观察“词串”和“非词串”的分支熵，本文发现一个词串的边界处的分支熵比较具有识别性，而词串内部的分支熵则识别词的作用较小，所以本文选择了词边界的分支熵作为整个词的分支熵。该方法有两个优点：一是分支熵可以被提前计算出来而不需考虑上下文；二是，分词算法可以采用动态规划的方法去实现。

        进而本文定义了一个词的“自主性（或称独立性）”C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\0c3c4907d943495b952e80ec0d135329\clipboard.png，该值与一个词的独立成词的概率成正相关关系。（而我们的实验，在融合左右分支熵方面尝试了“相加、相乘、取最小值三个方式”）

        采用上面计算的VBE，则分词问题就转化为了在所有的切分方式中，寻找各切分词串分支熵加和最大的那个切分方式，其中考虑到了词串的不同长度对分支熵的影响，这实际是在解决不同gram的词汇分支熵大小比较问题。即为：

C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\b068bf08688b46bda836fd99a9a2f7f1\clipboard.png

**6  Results and discussion**

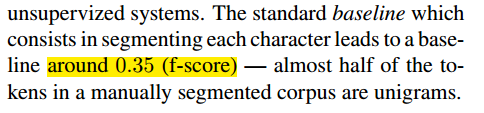
        该方法相比ESA，准确率可以达到ESA的水平，却不需要多次迭代以及训练参数，该方法仅仅基于左右分支熵相加这一种简单的指标评估一个词汇的紧密程度（或者是独立程度），且不需要参数去平衡左右参数的影响。

        然后本文列举了几项系统的评估，效果稍逊于ESA参数最优时的结果，但是已经非常接近。本文还分析了分词出错的原因，其中一部分是与日期或者中文数字相关，另一部分则是由于频繁出现的语法词素（我认为应该是说的类似：“在这”，“是在”，“中的”这种类型的词汇）以及功能性很强的词缀（例如：“你的”，“我的”里面的“的”？）要消除这些错误，第一部分可以通过充分的预处理解决，第二部分则可能需要在系统中引入一些语法的先验知识才能解决。

        通过系统评估可以看到，在PKU和MSR上效果较好，这是因为PKU的语料全部来自人民日报，数据量相对其他语料较少，而且语法结构和表达方式相对统一，而其他香港和台湾的语料来源十分多样。这也说明了，语料内部的一致性与语料库的大小同样重要。

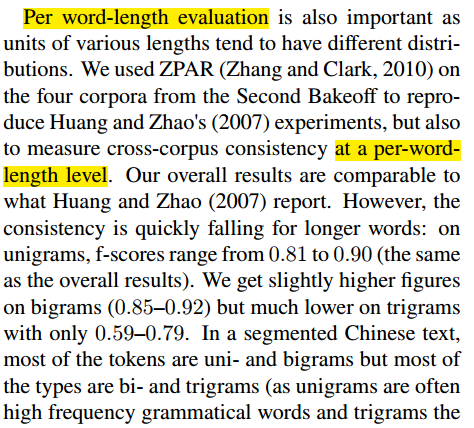
**阅读遇到的问题**

① pp.384,line 25,para.2。Part3



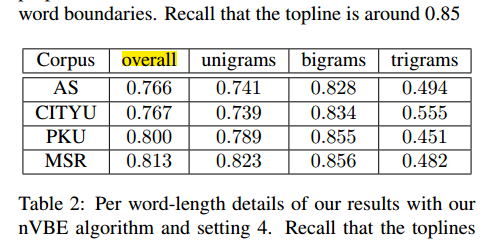
我没太看懂baseline是怎么得到的0.35，文中说是从切分每个字符中得出，不太明白什么意思。

②pp.384,line 1,para.3,Part3



 这里的per word-length evalution指的是什么方法？上文中我说明了下我的理解，不知是否正确。

③pp.386,Part6



这里的overall是指的在实验中，unigrams，bigrams，trigrams全部使用，而后面的实验仅仅只使用一种指标（unigrams或者bigrams或者trigrams）的意思吗？