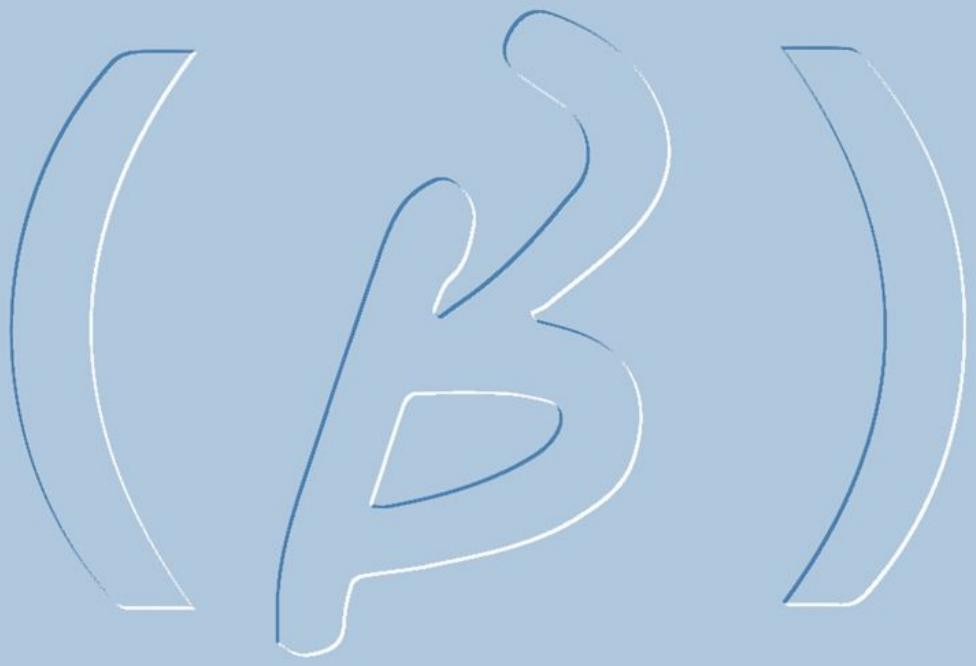

Bracket

The Mathematical Model Behind The Intelligent Cross-language Document Annotating System

Lin Zuzeng

(DRAFT)

What is Bracket?

- An Intelligent translation accessory which only provides essential phrase hints without interrupting your reading.
- Cloud-based, runs in the browser or the PDF reader.
- User's personalized language level model is generated automatically when using this software.
- The annotations are based on every user's personal language level individually, i.e. the software predicts what the user doesn't know.

Where can I use Bracket?

- Reading technical documents, especially academic journals with a lot of scientific terms.
- Surfing news websites in foreign language.
- Help you understand stories or novels.
- Providing data interface with digital dictionaries, or notbook software (Evernote, etc.) which can help you with language learning.

Demo!

Is it ready now?

- Working. It can deal with words, prhases, tenses, and plural form correctly, but sometimes its annotation is not so intelligence.
- Two major problems:

1. How to predict whether the user knows this word or not.
2. How to decide the proper translation of a word with multiple meanings.

Solution to Prob.1

- Frequent words are usually familiar to most of the people. We can decide the difficulty of a word by its probability of appearance in daily language documents.
- Noitce that everyone has his own field of interest (FOI), so the same word may have different levels of difficulty. So the model that simply analyzes the frequency can only give a global rank of the difficulty of a word and associate everyone to a certain level in that rank. (which is excatly what Bracket implements now.)
- It totally ignores the difficulty relative to different persons. So it generally works bad.

Model 1 to Prob.1

- Inspired by PageRank and Markov process model, we suggest a new model by assuming the language samples as G, every doucument in G that belongs to some specific subject as D.
- We assume that words form documents in the same subject may have some kind of connections.
- 将来自同一文档D内的每一个单词相互视为“互相具有超链接”，建立类PageRank模型，则可以反映词汇间的内在联系。
- We suppose that user gives his feedback by marking a word he knows as "easy word".
- 以用户选择的“Easy word”为起点，在该图中作random surfing，则“最终到达概率”较大节点，也可能是这个人认识的词。

Model 2 to Prob.1

- The model 1 is 以词汇为中心，注意到了词汇间的关系，适合交叉学科的论文，比如能够标出生物学论文中出现生物信息学术语。
- 但是忽略了中国英语教育按等级划分的事实，另外，有时候，可能刚开始了解一个学科的人“掌握的单词”分布和这个词的使用频率无关。这一模型没有利用好所有用户的“Easy word”信息。
- 注意到不同用户数据的横向比较，能提供词汇难度的另一个角度的信息。

解决方案2

- Inspired by Standford CS224D, we can caculate the co-existence matrix whith users as its rows and "easy words" as its lines, uses SVD to obtain “主题维度”.
- 对每个词计算在这些主题维上的分量，对每个人也同样计算对主题维度的分量。
- 当用户Pi在某个主题维度Xi分量很大时（说明他这块很熟），对这一主题维度Xi分量很大的词汇（很多人都会的词）就不用注释了。
- 缺点，计算量很大！矩阵太稀疏，词汇和用户一多就爆炸了。

问题一总结

- 所以其实还是没有很好的解决。。。

问题一总结

- 问题一对应的最接近的学术问题是主题分析，除了上述方法还查到用k临近做聚类等。
- 另外，上述只以词为单位进行分析，而忽视了亚词单位。注意到英语词汇的特点，具有一定的结构，比如词的前缀后缀，有时可被直接用于推断词的相似性。这方面可能不易采用数学建模手段完成，而需要专业知识介入。
- 另外，词汇第一次出现到其被用户是为掌握的时间长度，可能也可作为自动难度判断的参数。初步设想这里可以对大数据进行曲线拟合，建立学习模型。

问题二

- 问题二实质是翻译问题。
- 有别于传统翻译系统，Bracket 只要求给出多义词词汇在上下文语境下的含义，而不像机器翻译系统那样要求输出句子，因此没有句子结构，语法，插入词汇的等等一大堆问题。
- 但是，这一基本操作在翻译流水线中非常密集，因此我们希望识别速度很快。

问题二

- 目前Bracket采用的是自己开脑洞想出来的“聚光灯算法”，即在翻译阶段（second-pass）每次窗口中心指向一个词汇，提交到数据库查询，返回所有可能的词组，及每个词人工标注的学科分类，在窗口内向前向后扫描匹配，对同一学科分类出现的次数求和，取最大，使得句子里每一个词尽量属于同一个学科。
- 例如 sell(商业用语:卖) goods (商业用语:货物;形容词:好)-->最大共同学科:商业用语--> 卖货物

问题二的初步想法

- 这种方法很简单，也没啥数学内涵，但是，人工标注质量太差，错误和遗漏过多！
- 注意到近年来 Deep Learning 中的word2vec模型，试图抛弃传统的人工标注，而利用输入语料库中，同一个词在不同出现地方的上下文词汇的概率分布信息，将该词编码为一个行向量。
- 也就是利用词附近的上下文来机器学习这个词的含义，从而生成可以快速计算词汇相似度的特征向量。（详见论文）
- 这一方法还没有被应用于语言翻译问题，设想对某一对齐语料的进行学习，则可能可以得到十分理想的“源语言上下文->目标语言中的词意”。

问题二的初步想法

- 上述模型也存在缺点是，对齐语料相对难以取得。某些特定领域可能缺乏对齐语料。
- 因此设想通过众包（类似Google的ReCapcha图片验证码帮助图书馆识别难以OCR的字符的思路），让用户可以提交“更好的翻译建议”来帮助Bracket。
- 这时候，需要新的模型来评估用户的提交的建议。例如，建议间的相似度衡量，可以采用切词+余弦相似度，也可以用于避免恶意提交。

问题二的初步想法

- 总之，关于同义词翻译这块的模型简直太多了，有基于概率（贝叶斯，HMM）的，有基于图论的（Random Forest），也有用到ANN，SVM，Deep Learning 学习的。都实验一遍是不可能的，所以需要更深入的方法评估。

Acquisition of traing data.

- 有网上的语料库，但是是生活用语过多。
- 专业术语方面维基百科中英对照极好。40GB，已加入语料全家桶：）
- 根据Bracket隐私政策，用户翻译的公开的，非安全链接的文档也会被加入语料。

THANK YOU

- 以上大概是关于Bracket目前我的一些初步想法。由于对这方面了解较少，所以多半是看了论文或者一些资料，生搬硬套出来的模型。可能还有更多比较好的模型，希望指导！