词汇相似度计算 实验报告

本实验实现了 5 种词汇相似度的计算方法并在 WordSimilarity-353[1]数据集上进行了评价。

- 评价标准与参考基线

本文采用 Spearman 相关系数来评价相似度计算，在计算排名时，如有打分相同的情况则对排名进行平均。

$$\rho = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}$$

作为参考，对 Set1 和 Set2 人工标注的结果进行了交叉验证，即计算每个人的打分与其余人打分均值之间的相关系数，得到的均值和范围如表 1，这可以作为这个数据集上的效果上限。

	Human
Set1	.7988 (.6790 ~ .8627)
Set2	.7286 (.5016 ~ .8128)

表 1. 人工标注之间的相关性系数

- 基于语义词典（WordNet）的词汇相似度计算

本实验使用了 nltk 提供的 WordNet 工具[2]。使用了 6 种不同的相似度计算指标：path length (path), Leacock-Chodorow Similarity (lch), Wu-Palmer Similarity(wup), Resnik Similarity (res), Jiang-Conrath Similarity (jcs), Lin Similarity(lin)。其中后三种需要额外的语料库数据来获得一些统计信息，故分别尝试了 nltk 提供的 brown, semcor, genesis 三个语料库。并且尝试了将两个词所属词集(Synset)之间最大一对或所有对的均值的相似度作为词之间的语义相似度。实验结果如表 2（简洁起见，后 3 中相似度计算指标只写了在效果最好的额外语料库上的结果。）。

从实验结果上可以看出，一般基于路径距离的方法比较适合用 max 的方式进行转换，而基于互信息的方法比较适合用 average 的方式进行转换。整体而言，基于路径距离的方法效果较差。最好一组结果出

[1] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[2] http://www.nltk.org/howto/wordnet.html

现在 res 指标+genesis 语料库时，整体的 Spearman 相关系数达到了 0.3733。但是借助 WordNet 得到的词汇相似度效果总体而言并不太好。

MAX	All	Set1	Set2	**AVE**	All	Set1	Set2
path	.2927	.3117	.2424		.2443	.2374	.2200
lch	.3035	.3337	.2431		.1739	.2847	. 0761
wup	.3312	.3590	**.2589**		.2965	.3073	.2431
res (semcor)	**.3335**	**.3823**	.2372	(genesis)	**.3733**	.4176	**.2844**
jcn (brown)	.2835	.3648	.1587	(brown)	.2926	**.4435**	.1214
lin (brown)	.2993	.3600	.1977	(brown)	.3290	.3673	.2532

表 2. 基于 WordNet 的词汇相似度计算结果

- 基于 Bing 查询结果的词汇相似度计算

这一方法根据两个词分别与联合在 Bing 中查到的页面数来计算两个词的语义相似度[3]。有多种指标可以使用，公式如下。其中，估算 Bing 总页面数为 N=100M(依据 Bing 中“a”能检索到页面 83.5M 个，不过实际上经验性结果差不多也是这里最好)。

$$\mathrm{WebJaccard}(P,Q)=\frac{H(P,Q)}{H(P)+H(Q)-H(P,Q)}$$

$$\mathrm{WebOverlap}(P,Q)=\frac{H(P,Q)}{\min\big(H(P),\ H(Q)\big)} \qquad \mathrm{WebDice}(P,Q)=\frac{2H(P,Q)}{H(P)+H(Q)}$$

$$\mathrm{WebPMI}(P,Q)=\log\frac{N*H(P,Q)}{H(P)*H(Q)} \qquad -\mathrm{NGD}(P,Q)=-\frac{\max\big(H(P),\ H(Q)\big)-H(P,Q)}{\log N-\min\big(H(P),\ H(Q)\big)}$$

实验结果见表 3。在具体实现时，有些检索词被屏蔽，该对的相似度视为所有对相似度的均值。

	All	Set1	Set2
WebJaccard	.4575	.4579	.4451
WebOverlap	.4722	.4936	.4484
WebDice	.4573	.4577	.4451
WebPMI	.4868	.5084	.4643
-NGD	**.5101**	**.5386**	**.4704**

表 3. 基于 Bing 查询结果的词汇相似度计算结果

[3] 查询的 URL 为：http://cn.bing.com/search?q="%s"&qs=bs&form=QBLH，%s 为 word 或 word1+’+’+word2

- 基于维基百科页面词频的词汇相似度计算

这一方法统计词在维基百科页面[4]出现的概率作为词汇的相似度。具体来说，有：

$$\mathrm{Sim}(w_1, w_2) = P(w_1|w_2) + P(w_2|w_1), \quad P(w_1|w_2) = Cnt\big(w_1\ in\ \mathrm{Page}(w_2)\big)/Size\big(\mathrm{Page}(w_2)\big)$$

而在计算$Size\big(\mathrm{Page}(w)\big)$时，有三种方式，分别为 Page(w)中的字符数、空格数(以估算词数)和 w 在页面中出现的次数。实验结果如表 4：

	All	Set1	Set2
Blank	**.5826**	**.6350**	**.4816**
Character	.5824	.6349	.4815
word_w	.5670	.6196	.4697

表 4. 基于维基百科页面词频的词汇相似度计算结果

- 基于 GloVe 词向量的词汇相似度计算

使用 glove.840B.300d[5]预训练好的词向量来计算词汇相似度。使用欧氏距离和余弦距离 2 中方式，结果如表 5。可以看到，相对而言，使用词向量的效果是最好的，而余弦距离的效果比欧氏距离要好。

	All	Set1	Set2
Cosine	**.7379**	**.7454**	**.6685**
Euclidean	.5882	.6909	.4623

表 5. 基于 GloVe 词向量的词汇相似度计算结果

- 基于上述方法结果进行特征训练的词汇相似度计算

我们使用 SVM-rank[6]工具，以前面的：基于 WordNet 的方法： res_ave(genesis)，基于 Bing 查询结果的-NGD、基于维基百科页面词频的 Blank 以及机遇 GloVe 词向量的 Cosine 这四个特征进行训练。特征做归一化处理。分别尝试了两组互为验证及组内分别的 5 折交叉验证及全部数据上的 5 折交叉验证等多种验证模式。结果如表 6 所示。在训练时，调整参数对结果影响较小，故使用线性核函数，参数 c 统一设置成 1。

---

[4] 查询的 URL 为：https://en.wikipedia.org/wiki/+word，有时有可能是并不是实际的页面

[5] http://nlp.stanford.edu/projects/glove/

[6] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

	All	Set1	Set2
set1-set2	.7514	**.7723**	**.6928**
in_set_5-cv	.6313	.7631	.6715
all_5-cv	.7610	--	--

表 6. 基于特征训练的词汇相似度计算结果

在调试时发现交叉验证组的效果比较不稳定，主要应该是折间有差异，将多折结果合并计算排名并不可靠。只看两组互为验证的实验效果，发现较单一模型的效果有显著地提升。分析 svm-rank 给各个特征打出的权重（0.52，0.90， 3.28，4.41 与 1.46，1.94， 2.17，5.10）可以发现，这四个特征都是比较有用的，且靠后的特征重要性更高。

- 总结

本实验实现了 5 种词汇相似度的计算方法并在 WordSimilarity-353 数据集上进行了评价。五种方法最好的参数配置下的效果如表 7。5 种方法的效果依次提升，机遇大量数据的效果要优于基于规则和语义网的效果，在数据的基础上获得的词向量的效果由比简单的词频统计的效果要好。最后特征组合的效果已经比较接近人类标注的效果了。且从实验效果上可以看出，Set1 与 Set2 的分布并不太一致，比如 Bing 和 Wikipedia 的方法在 Set2 上效果相近，而在 Set1 上效果却有较大差异。

	All	Set1	Set2
WordNet	.3733	.4176	.2844
Bing	.5101	.5386	.4704
Wikipedia	.5826	.6350	.4816
GloVe	.7379	.7454	.6685
SVM-rank	.7514	**.7723**	**.6928**
Human	--	.7988	.7286

表 7. 各种计算方法的效果总结

GitHub 地址：

https://github.com/Erutan-pku/WordSense

其它参考资料：

http://blog.csdn.net/wsywl/article/details/5859751