基于改进编辑距离的中文相似句子检索①

车万翔② 刘挺 秦兵 李生
（哈尔滨工业大学计算机学院信息检索研究室 哈尔滨 150001）
{car, tliu}@ir.hit.edu.cn

摘要 中文相似句子检索的方法在基于实例的机器翻译等中文信息处理领域，具有非常广泛的应用背景。本文提出的基于改进编辑距离的中文相似句子检索方法，在使用信息检索技术提高检索效率的同时，以普通编辑距离算法为基础，加入了词汇的语义信息，使之更加符合中文句子相似度计算的要求。改进编辑距离与单纯基于语义辞典计算句子相似度的方法相比，具有便于扩展，准确率高等优点。在基于大规模双语句对检索的英文辅助写作系统中使用该算法进行中文句子检索，最后获得了 81.33%的查准率和 95.31%的查全率。

关键词：改进编辑距离、相似句子检索、英文辅助写作

0 引言

相似句子检索，在自然语言处理领域具有非常广泛的应用背景，如信息过滤技术中的句子模糊匹配，基于实例机器翻译的原语言检索，自动问答技术中常问问题集的检索以及问题与答案的匹配，基于双语语料库的英文辅助写作等。因此长期以来，相似句子检索问题，一直为人们所热衷。

目前句子相似度计算一般分为三个等级[1]，分别为语法相似度、语义相似度和语用相似度。计算句子之间的语用相似度，一直是人们的目标，但是其计算具有相当的难度，效果还不尽如人意。而在一般的应用中，只计算句子的语义相似度就能够达到我们的需要。句子的语义相似，指的是两个句子之间结构类似并且词汇使用同义或者近义词代替。例如：“我喜欢吃苹果”与“我爱吃香蕉”就是一对语义相似的句子。

目前对句子语义相似度计算的研究方法主要有：基于相同词汇的方法[1]、使用语义词典的方法[2][3]、使用编辑距离[4]的方法，以及基于统计的方法[5]等。其中，基于相同词汇的方法有很明显的局限性，对于同义词之间的替换则无能为力。而使用语义词典的方法，可以很好的解决这一问题，但是单纯的使用语义词典的方法，并没有考虑到句子内部的结构和词语之间的相互作用关系，准确率不高。

编辑距离通常被用于句子的快速模糊匹配领域，但是其规定的编辑操作不够灵活，也没有考虑词语的同义替换。最后基于统计的方法，需要构造大量的训练语料，工作量是十分巨大的，而且还存在着数据稀疏的问题。

我们所提出的改进编辑距离的方法，吸取了基于语义词典的方法和编辑距离方法的优点，同时克服了它们的一些不足。与普通编辑距离算法不同，改进编辑距离方法同时使用了 HowNet[6]和《同义词词林》[7]两种语义资源，计算词汇之间的语义距离，同时赋予不同编辑操作不同的权重，在不用经过词义消歧和句法分析的情况下，兼顾了词汇的顺序和语义等信息，最终获得了 81.33%的查准率和 95.31%的查全率。本文的第 1 部分描述了英文辅助写作系统框架以及各个模块的算法。第 2 部分给出了测试结果。第 3 部分讨论该系统的优点和缺点。第 4 部分给出了最后的结论。

1. 相似句子检索

相似句子检索过程如图 1 所示。

①863 计划(2002AA147020-11)、国家自然科学基金(60203020)资助项目。
②男，1980 年生，博士生；研究方向为自然语言处理，信息检索；联系人。

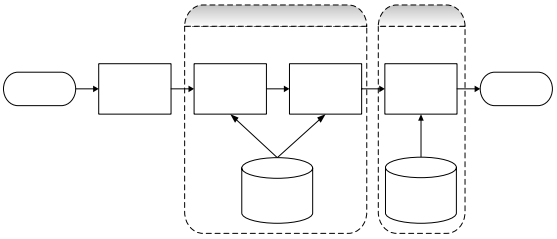


图 1. 相似句子检索流程图

核心的相似句子检索算法由候选句子粗匹配和句子相似度计算两部分组成。其中，候选句子粗匹配模块首先到句库中找到可能与输入为相似句的候选句子，目的是提高检索效率。句子相似度计算模块将每一个候选句子与输入句进行仔细的相似度计算，并按照相似度的大小排序，获得最终的输出结果。

1.1 候选句子粗匹配

候选句子粗匹配由词扩展和快速检索两个子部分构成。

首先为了获得较高的召回率，必须对分词以后的各个词汇进行适当的同义词扩展。在此，对同义词的定义即不能太宽泛，又不能太严格。如果太宽泛，将检索到许多无关的句子，降低了系统的准确率和效率；而如果太严格，又可能漏掉许多有用的句子，降低了系统的召回率。

在此使用 HowNet 语义辞典作为同义词扩展的资源。HowNet 中同义词的定义为具有相同的英语译文（*W_E*）和语义定义（*DEF*）的词汇。例如“我”和“俺”，简化词条如下：

NO.=085498 W_C=**我** *W_E=I* *DEF=firstPerson|我*	NO.=000701 W_C=**俺** *W_E=I* *DEF=firstPerson|我*

可见，“我”和“俺”具有相同的英语译文 (*W_E*) “I”和语义定义 (*DEF*) “firstPerson|我”，是一对同义词。在表 1 中，我们给出了使用 HowNet 进行词扩展的一些例子。从中可以看出，使用 HowNet 进行词的扩展，效果是比较理想的。

表 1. HowNet 进行词扩展示例

原词	扩展词
俺	本人 鄙人 我 吾
俺们	我方 我们 吾 吾侪 咱 咱们
按	按捺 把持 扼制 管束 收 束 压 压制 抑制 制 制约 羁
按键	按钮 电键

其次，为提高系统的效率，首先对整个语料库进行初步的筛选，确定数量不多但有可能与用户的需求相似的候选句，然后对这些候选句进行精确的语义相似度计算，得出最终的结果。

选择候选句的依据是，如果一个句子中与用户的需求相同或同义的词越多，越有可能与用户的需求相匹配，即权重越大。我们采用信息检索中常用的倒排文档索引[8]的方法进行检索。

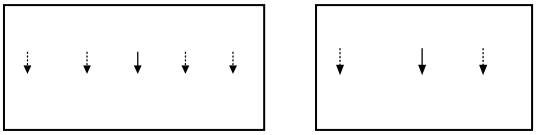
在此，并不需要对用户输入句子中的词进行词义消歧，而直接对所有扩展后的词进行检索。这是由于输入的词并非孤立，当与其余的词共同检索的时候，能达到消歧的目的。以“打”为例，当输入“打毛衣”时，“打”被扩展为“打击”，“编织”等。一个句子中同时含有“编织”和“毛衣”的可能性很大，而“打击”和“毛衣”几乎不可能同时出现在一个句子中。于是，含有“编织”和“毛衣”的句子更容易成为候选句。按照句子权重由大到小的顺序，我们选择前 100 个句子作为候选句。

1.2 语义相似度计算

本文采用改进编辑距离的算法计算句子之间的语义相似度。在介绍改进编辑距离的算法之前，首先介绍两个句子之间的编辑距离以及如何进行计算。

编辑距离指从一个以字为单位的句子变为另一个以字为单位的句子所需要最小的编辑操作个数。编辑操作共有“插入”、“删除”和“替换”三种。图 2(a)显示了“爱吃苹果”与“喜欢吃香蕉”之间的编辑距离为 4，如四条虚线所显示。

从该计算过程可以看出，单纯使用以字为单位编辑距离的方法，计算出的语义距离和实际情况是有很大的出入的。首先，编辑距离算法以字为基本计算单位，而在汉语中，单个的字往往是不具备意义的。例如上面的“苹”、“果”等字，并不能反映其所合成词的意义。其次，词语之间的替换操作的代价并非都是相同的。例如，“爱”被“喜欢”替换，代价不应该很大。最后，如果在被检索句子或短语中间加入为数不多的词，语义也不会有太大改变。例如“爱吃苹果”与“爱吃甜苹果”就非常相似。



(a)编辑距离的　　(b)改进编辑距离

图 2. 编辑距离与改进编辑距离的比较

基于以上的观点，我们提出了改进编辑距离算法，即以词汇为基本的计算单位，同时以 HowNet 和《同义词词林》作为语义距离的计算资源，并减小插入操作的代价。

HowNet 定义的同义词如前所述。下面介绍如何使用《同义词词林》进行语义距离计算。

在《同义词词林》中，将词的词义分为大、中、小类描述了一个由上到下，由宽泛概念到具体词义的语义分类体系，并将所收的词按词义分门别类组织在其中。于是，每个汉语词都按照其语义，赋予了一个或多个 3 位的语义代码，与此分类体系相对应的是一个词义的编码体系，描述如下：

<词义编码>::=<大类><中类><小类>

<大类>::=<大写英文字母>

<中类>::=<小写英文字母>

<小类>::=<数字><数字>

例如：“苹果”Bh07，“香蕉”Bh07，“西红柿”Bh06，……。则 A, B 两词之间的语义距离为：

$$Dist(A,B)=\min_{a\in\mathbf{A},b\in\mathbf{B}} dist(a,b)$$

其中，**A**，**B** 分别为 *A*，*B* 两词具有语义的集合，语义 *a*，*b* 之间的距离为：

$$dist(a,b)=2*(4-n)$$

其中 *n* 为它们之间的语义代码从第 *n* 类开始不同，全部相同语义距离为 0。可见 *Dist*(苹果，香蕉)=0，*Dist*(苹果，西红柿)=2。这种代数操作要较之使用 HowNet 计算语义相似度方便、快捷的多。

在此，也没有进行词义消歧工作，而使用两个词之间的最短语义距离作为它们之间的语义距离。这是因为我们在候选句子粗匹配时，检索到的句子中的词与用户输入句中的词具有相似的语义，起到了词义消歧的目的。

表 2 规定了词“A”与“B”进行各种编辑操作后编辑距离由小到大改变的顺序，其中，“*”代表 1 至 4 个词，“A”与“B”为用户输入的两个连续的词，X'为 HowNet 定义的 X 的同义词，X"为《同义词词林》定义的 X 的近义词。

表 2.编辑距离改变的次序

级别	模式
1	AB
2	A*B
3	AB’; A’B
4	A*B’; A’*B
5	AB’’; A’’B
6	A; B
7	A’; B’

据此，我们可以定义改进编辑距离计算语义相似度方法中各种编辑操作的代价，如表 3 所示，其中“→”代表替换操作。

表 3. 改进编辑距离编辑操作代价定义

编辑操作	操作代价
A → A	0
插入	0.1
A → A’	0.4
A → A’’	*Dist*(A, A’’)/10 + 0.5
其它	1

根据以上对编辑距离的重新定义，“爱吃苹果”与“喜欢吃香蕉”之间的改进编辑距离计算如图 2(b)所示，其中“爱”→“喜

欢”代价为 0.5，“苹果”→ “香蕉”代价为 0.7。最后改进编辑距离结果为 1.2，要较之普通编辑距离计算的距离 4，更符合实际情况。

与计算普通的编辑距离相同，也使用动态规划算法计算改进编辑距离。图 3 为计算两个句子 S1 与 S2 改进编辑距离的算法。

输入：经过分词的两个句子 S1 和 S2
For i, M[i,0]=i
For j, M[0,j]=j
For i, j ≥1,
M[i,j] = min {M[i-1,j]+1, //删除操作
M[i,j-1]+0.1, //插入操作
M[i-1, j-1]+Dist($S1_i$, $S2_j$)}
输出：S1 和 S2 的改进编辑距离

图 3. 改进编辑距离算法

最后，经快速检索步骤检索出来的每个句子都与用户输入的句子计算出改进编辑距离，然后按照由小到大的顺序进行排列，取出前几个作为最终的结果。本文选择前 10 个结果。

2. 测试结果

在现今机器翻译效果不令人满意的情况下，辅助写作系统逐渐引起了人们的重视。我们使用基于相似度的句子检索技术，以大规模的双语句对语料库为资源建立的英文辅助写作系统，容许用户输入中文整句或者短语，系统在双语句对库中快速的检索与之相似的中文句子，对应给出这些句子的英语翻译。具有翻译准确，示例性强等优点。并且随着收集的双语语料库的增加，覆盖面的扩大，辅助写作的效果也会越来越好。目前，我们收集的双语句对已达到 25 万对。

英文辅助写作系统首先将用户输入的中文整句或者短语分成单独的词汇。在此，我们采用最大概率分词算法[9]。由于输入中一般不会有未登录词，因此采用最大概率分词算法即可达到 95%以上的准确率，足以满足系统的要求。

我们采用国际上通用的 MMR(Mean Reciprocal Rank)评分机制对算法进行评价，即第一个结果匹配，分数为 1，第二个匹配分数为 0.5，第 n 个匹配分数为 1/n，如果没有匹配的句子分数为 0。最终的分数为所有得分之和。

我们随机测试了 150 个句子或短语，使用全部 25 万句对双语语料库，最终的平均分数为：1.254。其中 122 句输入能够找到相匹配的结果。在没有给出相应匹配结果的 28 个测试句子中，有 22 个是在语料库中就没有相匹配的中文句子，6 个是系统没有给出相匹配的正确句子。在表 4 中，列出了一些系统输入和输出的最终结果。

表 4. 系统输入输出结果示例

用户需求	系统输出
属于那种人	他是那种随波逐流的人。 He is the kind of person who swim with the tide. 同那种人打交道得随机应变。 You need to keep your wits about you when you're dealing with a man like that. ……
挤进了大礼堂	大家拥进了房间。 Everyone piled into the room. 人们一起挤进了一个小房间。 People were packed together in a small room. ……

同时，我们又将基于改进编辑距离的中文相似句子检索方法与单纯基于语义辞典的方法进行比较。比较结果如表 5 所示。其中，查准率 *P* 和查全率 *R* 的定义如下所示：

$$P = \frac{\text{含有相匹配的句子数}}{\text{测试句子数}}$$

$$\mathrm{R} = \frac{\text{含有相匹配的句子数}}{\text{语料库中有相匹配的句子数}}$$

可以看到，改进编辑距离的方法比基于语义词典的方法在辅助写作系统中，取得了更好的效果。基于语义词典的方法除了进行语义距离的计算过程中使用单纯的语义词

典进行语义距离的计算外，其余步骤与改进编辑距离的方法完全相同。

另外，我们建立的英文辅助写作在线演示系统[③]，目前已经收集了超过 5000 个真实有效的翻译查询。经过分析可知，其中主要的查询为短语和短句，抽样检查系统的查准率在 80%以上。

表 6. 改进编辑距离的方法与基于语义词典的方法比较结果

方法	改进编辑距离	基于语义词典的方法
测试句子数	150	150
含有正确匹配句子数	122	116
系统没有给出正确匹配的句子数	6	12
语料库中没有相匹配的句子数	22	22
查准率 *P*	81.33%	77.33%
查全率 *R*	95.31%	90.63%
MMR	1.254	1.209

3. 讨论

以上测试结果说明，在英文辅助写作系统中，使用改进的编辑距离进行中文相似句子检索取得了较为理想的效果，查准率达到了 81.33%。

对匹配不正确的例子进行分析可知，该方法目前的问题是如果用户需求的句子较长，就很难找到与之完全匹配的句子，例如输入“顺利到达职场生涯的顶峰”与输出“他当选为总统是他职业生涯的顶峰”，虽然匹配了后半部分，但是前面的“顺利到达”并没有体现出来，这就需要对句子进行恰当的分割，然后分别查询。

通过使用改进编辑距离的方法与基于语义的方法相比较可知，改进编辑距离的方法考虑了较多的词汇顺序以及距离的信息，例如对于输入句子

S_{input}=“匆匆忙忙交给她”

改进编辑距离方法首选结果为：

S_{edit}=“她急忙把孩子交给她妹妹照管”

而基于语义方法首选结果为：

S_{sem}=“她就*匆匆忙忙*挑了一件店里最贵重的衣服，把它*交给*售货员，此人为*她*尽快包好”

虽然 S_{sem} 与 S_{input} 中所有的词都匹配，可是并非最好的结果，反而是 S_{edit} 这个与原句词匹配不多的句子更加符合需求。可见，改进编辑距离的方法比基于语义的方法更能反映句子中词汇顺序和距离的信息，最终取得了更好的效果。

4. 结论

基于改进编辑距离的中文相似句子检索方法在英文辅助写作系统中获得了较好的结果。随着双语语料库的增加，覆盖面的增大，系统的效果也将有一定的提高。同时，又具有易于扩展的优点，我们可以方便的将该方法应用到其它的领域中，如：基于实例的机器翻译中的原语言搜索，自动问答中的常问问题库检索以及问题与答案匹配等等。

为了进一步提高基于改进编辑距离的中文相似句子检索算法的性能，我们以后需要将句子分成较小的独立子结构分别查询，因为一个较长的句子往往不容易匹配，这需要使用较复杂的句法分析技术。

参考文献

[1] Nirenburg S. Two Approaches of Matching in Example-Based Machine Translation, Proc. TMI-93, Kyoto, Japan, 1993

[2] 秦兵, 刘挺, 王洋等. 基于常问问题集的中文问答系统研究. 哈尔滨工业大学学报, 2003, 35(10): 1179

[3] Li S, Zhang J, et al. Semantic Computation in Chinese Question-Answering System. 2002, Journal of Computer Science and Technology, 17(6): 933

[4] Ristad E S and Yianilos P N Learning

[③] http://ir.hit.edu.cn/

string-edit distance. 1998, IEEE PAMI, 20(5): 522
[5] Chatterjee N, A Statistical Approach for Similarity Measurement Between Sentences for EBMT, 1999
[6] 董振东，董强．知网．http://www.keenage.com
[7] 梅家驹，竺一鸣，高蕴琦，殷鸿翔编，《同义词词林》第二版，上海：上海辞书出版社，1996 年
[8] Frakes W and Baeza-Yates R. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, N J, 1992.
[9] 刘挺，吴岩，王开铸．最大概率分词问题及其解法．哈尔滨工业大学学报，1998, 30(6): 37

Similar Chinese Sentence Retrieval based on Improved Edit-Distance

CHE Wan-xiang, LIU Ting, QIN Bing, LI Sheng
Information Retrieval Lab, School of Computer Science and Technology of HIT Harbin 150001

Abstract

The approach of similar Chinese sentence retrieval has been used widely in the field of Chinese information processing, such as Example Based Machine Translation (EBMT) and so on. The approach of similar Chinese sentence retrieval based on improved edit-distance not only uses the technology of information retrieval to improve the efficiency of retrieval, but also adds the semantic information of words into the normal edit-distance approach. The new approach is more consistent with the computation of Chinese sentence similarity. The approach of improved edit-distance has more advantages than original edit-distance algorithm, such as easily extending, high precision and so on. The new approach was used in the English writing assistant system based on a large bilingual sentences pairs and achieved 81.33% precision and 95.31% recall.

Keywords: improved edit-distance; similar sentence retrieval; English writing assistant