面向双语句对检索的汉语句子相似度计算

车万翔　刘挺　秦兵 李生
哈尔滨工业大学计算机学院
(哈尔滨工业大学 321 信箱，哈尔滨 150001)
E-mail:{carl, tliu, qinb, ls}@ir.hit.edu.cn

摘 要：在基于大规模的双语句对语料库的英文辅助写作系统中，我们采用了一种改进编辑距离的句子相似度计算方法，即对以往的编辑距离算法进行适当的调整，考虑了更多的汉语结构信息，使之更加符合汉语的句子相似度计算。同时使用了 HowNet 和《同义词词林》两部语义辞典作为语义资源，计算词汇之间的相似度。改进编辑距离的算法与单纯基于语义辞典计算句子相似度的算法相比，具有便于扩展，准确率高等优点，在英文辅助写作领域取得了令人满意的效果。对其进行适当的改进后，可适于多数需要计算句子相似度的应用领域。

关键词：句子相似度 改进编辑距离

Chinese Sentences Similarity Computation Oriented the Searching in Bilingual Sentence Pairs

Che Wanxiang　Liu Ting　Qin Bing　Li Sheng
School of Computer Science and Technology of HIT
(Harbin Institute of Technology 321 Box, Harbin 150001)
E-mail:{carl, tliu, qinb, ls}@ir.hit.edu.cn

Abstract: An improved edit-distance approach was used to compute the sentence similarity in an assistant English writing system based on a very large bilingual sentence pairs resource. It means that the original edit-distance approach was improved and made it more fit for the computation of Chinese sentence semantic similarity. At the same time we use two thesauruses – HowNet and "Cilin" as the semantic resource to compute the semantic similarity between two words. The approach of improved edit-distance has more advantages than original edit-distance algorithm, such as easily extending, high precision and so on. It has gotten a satisfying result. After modified appropriately, it could be used in most fields using the computation of sentence semantic similarity.

Keywords: sentences similarity, improved edit-distance

1. 引言

句子相似度的计算，在自然语言处理领域具有非常广泛的应用，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等。因此长期以来，句子相似度的计算问题，一直为人们所热衷。

辅助写作系统，在现今机器翻译效果不令人满意的情况下，也更加引起了人们的重视。而我们所采用的基于大规模的双语句对语料库的辅助写作系统，容许用户输入汉语整句或者短语，系统快速的到双语句对库中检索与之相似的汉语句子，然后给出这些句子的英语翻译。其具有翻译准确，示例性强等优点。并且随着收集的双语语料库的增加，其覆盖面的扩大，辅助写作的效果也会越来越好。目前，我们收集的双语句对达到 25 万。

在该系统中，我们主要使用了汉语句子相似度计算技术，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。而在我们的英文辅助写作系统中，只计算句子的语义相似度就能够达到我们的需要。句子的语义相似度，指的是两个句子之间结构类似，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。只要将检索到的汉语句子对应的英文翻译中的个别单词进行替换，就能达到辅助写作的目的。

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（Nirenburg 1993）[2]、使用语义词典的方法（王洋 2002, Li Sujian 2002）[3][4]、使用编辑距离（E. S. Ristad 1998）[5]的方法，以及基于统计的方法（Chatterjee, 1999）[1]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，其无能为力，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，其工作量是十分巨大的，而且还存在着数据稀疏的问题。

我们所采用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。与普通编辑距离不同之处在于，改进编辑距离的方法，同时使用了 HowNet[6]和《同义词词林》[7]两种语义辞典，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前提下，兼顾了结构和词汇等信息，最终获得了较好的效果。本文的第二部分描述了英文辅助写作系统框架的描述。第三部分给出了测试结果。第四部分讨论系统的优点和缺点。第五部分给出了最后的结论。第六部分介绍了一些未来的工作。

2. 系统描述

英文辅助写作系统流程如图 1 所示。以下我们将详细描述系统各个模块。

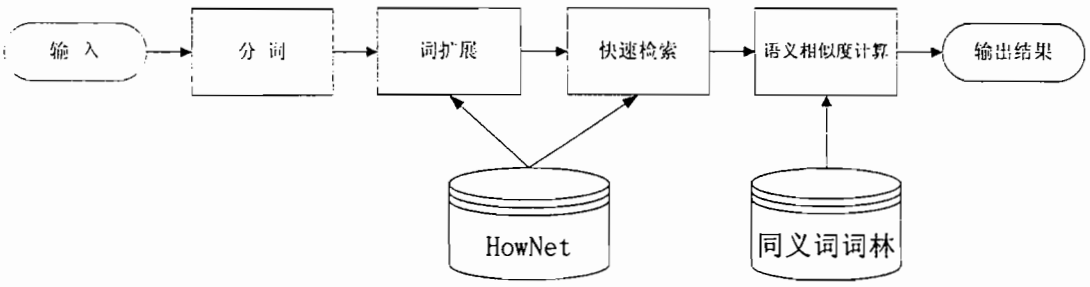

图 1. 英文辅助写作系统流程图

2.1 分词

英文辅助写作系统首先将用户输入的汉语整句或者短语分成单独的词汇。在此，我们采用正向最大匹配的分词算法[8]，虽然其准确率在各种分词算法中是最低的，但是其实现简单，分词效率高，同时便于添加新的词汇。

2.2 词扩展

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 语义词典作为词扩展的资源。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”，其简化词条如下：

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

2.3 快速检索

为提高系统的效率，首先对整个语料库进行初步的筛选，确定数量不多的，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，其就越有可能与用户的需求相匹配，即权重越大。我们采用倒排文档索引[9]的方法进行检索。

在此，并不需要对用户需求中的词进行词义消歧，而直接对所有扩展后的词进行检索，其依据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。显然，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”的句子被排在较靠前的位置上，更容易成为候选句。按照句子权重由大倒小的顺序，我们选择前 100 个作为候选句。

2.4 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，首先介绍什么是两个句子之间的编辑距离以及如何进行计算。

编辑距离指的从一个句子变为另一个句子所需要的最小的编辑操作的个数。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。

(a) (b)

图 2. (a)“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；(b)“爱吃苹果”与“喜欢吃香蕉”之间的改进编辑距离为 1.1，其中“爱” → “喜欢”代价为 0.5，“苹果” → “香蕉”代价为 0.6

从上面的计算过程可以看出，单纯使用编辑距离的方法，计算出的语义距离和实际情况是有很大的出处的。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，其代价不应该很大。最后，如果在被检索句子或短语中间加入为数不多的词，其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。

基于以上的观点，我们提出了编辑距离的改进算法，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。

表 2 中定义了以上几种编辑操作改变编辑距离的次序。我们假设“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. 词“A”与“B”进行各种编辑操作后改变编辑距离的顺序，(其中，“*”代表 1 至 4 个词)

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’

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，将词按照语义类分成树状的结构，越靠近根节点，语义的概念越抽象。具体的汉语词，只分布在叶子结点上。于是，每个汉语词都按照其语义，赋予了一个或多个 4 位的语义代码。例如："苹果" Bh07，"香蕉" Bh07，"西红柿" Bh06，……。计算词与词之间的语义距离，只是简单的代数操作。

按照表 2 规定的各种编辑操作后语义距离的次序，就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。

表 3. 改进编辑距离各种编辑操作的代价，其中"→"代表替换操作；"d"为"A"和"A""在《同义词词林》中最小语义距离

编辑操作	操作代价
A → A	0
插入	0.1
A → A'	0.4
A → A"	d/10 + 0.5
替换	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，最后改进编辑距离结果为 1.1，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。表 4 给出了一个计算两个句子之间改进编辑距离的动态规划过程的例子。

表 4. "顺利到达职场生涯的顶峰"与"他当选为总统是他职业生涯的顶峰"两个句子之间改进编辑距离的动态规划过程

		他	当选	为	总统	是他	职业	生涯	的	顶峰
	0	0	0	0	0	0	0	0	0	0
顺利	1	1	0.7	0.8	0.9	1	1	1	1	1
到达	2	2	1.7	1.6	1.7	1.8	1.9	2	2	2
职	3	2.7	2.7	2.6	2.5	2.4	2.3	2.4	2.5	2.6
场	4	3.7	3.7	3.6	3.5	3.4	3.3	3.2	3.3	3.4
生涯	5	4.7	4.7	4.6	4.5	4.4	4.3	3.3	3.4	3.5
的	6	5.7	5.7	5.4	5.5	5.4	5.3	4.3	3.3	3.4
顶峰	7	6.7	6.7	6.4	6.4	6.4	6.3	5.3	4.3	3.3

由表 4 可见，这两个句子之间的改进编辑距离为 3.3。

最后，经快速检索步骤检索出来的每个句子都与用户的需求计算出改进编辑距离，最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。

3. 测试结果

我们定义的打分方法为，第一个匹配，分数为 1，第二个匹配分数为 0.9，以此类推，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。则最高的得分为 5.5。分数越高结果越好。因为辅助写作系统输出的示例越多，可供参考的译文就会越多。

我们随机测试了 50 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：2.178。其中 40 句输入能够找到相匹配的结果。即正确率为 80%。在没有给出正确匹配结果的 10 个测试句子中，有 7 个是在语料库中就没有相匹配的中文句子， 3 个是系统没有给出相匹配的正确句子。

为了测试召回率，随机的在双语语料库中选出 50 个句对，将这 50 个句对中的汉语句子，经过适当的改造，形成新的输入，例如："使用效能和设计优美两方面的协调一致"被改造为"使用和设计协调一致"等。最终共 46 个句子检索得到了原来的句子。即召回率为：92%

基于语义词典的方法除了进行语义距离的计算过程中仅用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相同。最终的平均分数为：2.078。正确率为 70%。

4. 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计算句子语义的相似度取得了较为理想的效果，正确率达到了 80%。

其还具有易于扩展（指的是我们可以方便的规定各种操作的代价，如同义词、近义词之间的替换插入等），考虑了句子的结构信息等优点。

对匹配不正确的例子进行分析可知，该方法目前的问题是没有涉及到语用的层面，如"多少钱？"与"怎么卖？"这两个本来是同义的句子却没有匹配成功。另外，假如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。最后，分词效果有时也影响了结果，例如"这里人人都想学科学。"被分为"这里 人人 都 想 *学科 学* 。"于是，就与"交叉学科"相匹配了。

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结构信息，例如对于输入句子"匆匆忙忙交给她"，改进编辑距离方法首选结果为"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"她就*匆匆忙忙*挑了一件店里最贵重的衣服，把它*交给*一个售货员，此人为*她*尽快包好"。虽然句子"她就*匆匆忙忙*挑了一件店里最贵重的衣服，把它*交给*一个售货员，此人为*她*尽快包好"中与原来的句子中所有的词都匹配，可是并非最好的结果，反倒是"她急忙把孩子交给她妹妹照管"这句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的结构信息，最终取得了更好的效果。

5. 结论

使用改进编辑距离计算句子语义相似度的方法在英文辅助写作系统中获得了较好的结果，并且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，其又具有其易于扩展的优点，我们可以方便的将改进编辑距离的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，自动问答中的常用问题库检索以及问题与答案的匹配等等。

6. 未来工作

为了进一步加强英文辅助写作系统的实用性，我们以后需要进一步改进使用改进编辑距离计算句子相似度的方法。

首先，使用更强大的分词模块，以提高分词的准确率。

同时加入词频信息，系统应该更倾向于匹配低频词，因为高频词往往不是人们所关注的，并且其存在影响了句子改进编辑距离的计算。例如“的”，“了”等等，当然，这些也可以通过词性信息加以解决。另外，标点信息也很重要，因为往往在用户需求中加入“，”、“。”等标点，其语义将有很大的变化。以上参数的改变，在改进编辑距离算法的整体框架之下，是非常容易实现的。

我们也应该将句子分成较小的独立子结构，然后分别查询，因为一个较长的句子往往不容易匹配，这需要使用一定的句法分析技术。

参 考 文 献

[1] N Chatterjee. A Statistical Approach for Similarity Measurement Between Sentences for EBMT, 1999

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

[3] 王洋. 秦兵. 郑实福. 句子相似度计算在 FAQ 中的应用. 第一届学生计算语言学研讨会论文集. 第一届学生计算语言学研讨会. 中国 北京 北京大学. 2002: 175~181

[4] Li Sujian. Zhang Jian. Huang Xiong and Bai Shuo. Semantic Computation in Chinese Question-Answering System. 2002. Journal of Computer Science and Technology

[5] E. S. Ristad and P. N. Yianilos. Learning string-edit distance. 1998, IEEE PAMI, 20(5):522--532

[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》，上海：上海辞书出版社，1996 年第二版

[8] 朱晓旭. 汉字输入教学系统中词组切分方法的设计. http://www.lesson.com.cn/co/proseminar/thesis05.asp

[9] William B. Frakes and Ricardo Baeza-Yates. editors. Information Retrieval: Data Structures and Algorithms. Prentice Hall. Englewood Cliffs. N J. 1992.