翻译记忆中数据筛选方法的研究①

Research of Data Screening Methods in Translation Memory System

赵善祥　刘万军　(辽宁工程技术大学 电子与信息工程学院 辽宁 葫芦岛 125105)

摘　要： 翻译记忆是国内外专业领域计算机辅助翻译市场中使用最为广泛、有效的技术。记忆数据筛选（模糊匹配或语句相似度评价）决定了翻译记忆系统的有效性，这篇文章主要就其核心方法进行研究，介绍了使用编辑距离、向量余弦和 Jaccard 系数进行文本相似度评价的方法，使用基于《知网》和《同义词词林》的本体论、基于统计以及语法驱动的语义相似度评价方法，以及四种改进的评价方法，最后通过实验数据对其中四种典型方法进行比较分析，表明融合多目标评价可以提高数据筛选结果的实际有效性。

关键词： 计算机辅助翻译 翻译记忆 数据筛选 模糊匹配 语句相似度

本文在总结多种翻译记忆数据筛选方法的基础之上，对四种考察文本和语义相似度的典型方法做实验对比，阐述语义相似度评价对数据筛选结果的实际有效性的重要作用，多层次、多角度的考察影响结果有效性的因素可以增强数据筛选的准确性，因而在融合多目标考察的基础之上加深语义相似度评价方法体系的研究，将对计算机辅助翻译领域产生有益影响。

1　引言

翻译记忆(Translation Memory)[1]针对语言学的复杂性与多变性，充分利用数据库的强大功能，注重提高已有翻译资料的复用，从而无需重复翻译相同的句子。因此在资料重复度高达 20％～70％的专业翻译领域中，翻译记忆能大量消除译者的重复劳动，避免相同句子出现多种翻译结果，极大地提高了工作效率。

如何从庞大的翻译记忆数据库中快速有效地查找与原文最为相似的目标数据，以及如何在原文中正确识别专业术语并保持前后术语翻译的一致性是基于翻译记忆技术的计算机辅助翻译系统中两个主要问题，而前者的核心技术记忆数据筛选又叫做模糊匹配或语句相似度评价，决定了辅助翻译系统的准确性和有效性。目前被广泛接受的数据筛选目标主要包括文本相似度和语义相似度，前者仅考虑字面相似程度，通常以编辑距离[2]作为评价标准，也可采用计算向量余弦以及 Jaccard 系数[3]的方法，后者国内多以基于《知网》[4-6]和《同义词词林》[5]的本体论方法进行评价，国外有隐含语义索引(Latent Semantic Indexing) [7]等方法，此外还有语法驱动以及基于统计的方法等。

2　数据筛选的基本方法

2.1 文本相似度评价方法

2.1.1 编辑距离

编辑距离算法主要有俄国科学家 Vladimir Levenshtein 在 1965 年提出的 Levenshtein Edit Distance，它度量了一个字符串 $T=t_1t_2\cdots t_m$ 变换为另一个字符串 $T'=t_1't_2'\cdots t_n'$ 所花费的最小代价即为字符串 T 到 T' 的编辑距离 D(m,n)，而在实际应用中一般将句子中的单词或词组规划为相应的元素，求解句子 $S=w_1w_2\cdots w_m$ 到句子 $S'=w_1'w_2'\cdots w_n'$ 的编辑距离 D(m,n)。

编辑操作包含 3 种：插入某个字符；删除某个字符；将某个字符替换为另一个字符。

比如，字符串 Cat 和 Kate 之间的编辑距离为 2，计算过程如下：

(1) Cat->Kat(替换字符 C 为字符 K)；

① 收稿时间:2008-09-19

(2) Kat->Kate(插入字符 e)；

编辑距离为 0，表示两个字符串是完全相等的。距离越大，两个字符串的相似程度越小。

下面是计算编辑距离的动态规划算法[2]的核心部分：

$$D(m,n)=\begin{cases}0, m=n=0\\ D(0,m-1)+f(w_n'), m=0\text{且}n\neq 0\\ D(m-1,0)+f(w_m), m\neq 0\text{且 n}=0\\ Min\begin{bmatrix}D(m-1,n)+f(w_m),\\ D(m,n-1)+f(wn'),\\ M(m,n)\end{bmatrix}, \text{其 他}\end{cases}$$

$$M(m,n)=\begin{cases}Max[f(w_m),f(w_n')]\\ 0, w_m=w_n\end{cases}$$

其中 $f(w_n')$表示在 S 插入 w_n'所需的代价，$f(w_m)$表示在 S 中删除 w_m所需的代价，而 $Max[f(w_m), f(w_n')]$表示将 w_m替换为 w_n'所需的代价。

2.1.2 向量余弦和广义 Jaccard 系数

这两种方法有别于编辑距离的方法，它们都是从句子向量的角度去分析和评价句子间的相似程度。

将句子看作向量空间中的向量，句子中单词或词组看作向量在每个维度上的度量，这样两个句子的相似度可以通过两个句子所代表的向量在向量空间中的夹角大小来评价，两向量余弦的取值在 0 与 1 之间，适于作为评价度量，与此类似，广义 Jaccard 系数同样适用于对两向量夹角大小的评价。

以计算广义 Jaccard 系数的方法对句子 $S=w_1w_2\cdots w_m$和 $S'=w_1'w_2'\cdots w_n'$进行相似度评价，S 与 S’中总共所包含的单词或词组集合为 M，两个句子所代表的向量所在的向量空间的维度为 M 中元素的数量 c。句子 S 在向量空间中的向量 $X=(x_1, x_2\circ\circ\circ x_c)$，当 wi 属于 S’时，$x_i$为 1，当 w_i不属于 S’时，x_i为 0；句子 S 在向量空间中的向量 $Y=(y_1, y_2\circ\circ\circ y_c)$，当 w_i'属于 S 时，y_i为 1，当 w_i'不属于 S 时，y_i为 0，S 与 S’的 Jaccard 系数为：

$$EJ(X,Y)=\frac{X\times Y}{\|X\|\times\|X\|+\|Y\|\times\|Y\|-X\times Y}$$

但利用向量空间评价句子相似程度存在局限性，当句子维度过大时，计算的时间复杂度极大，因此有人提出通过关键词筛选进行降阶计算，总体来说此方法不具有可靠的普适性。

2.2 语义相似度评价方法

2.2.1 基于本体的方法

基于本体的方法是以语义义原为本体，将句子看作由若干义原有序组成的集合，通过对义原层面上的分析来对两个句子做相似度的评价。

该方法的基础一般建立在特定的概念领域之上，特定的概念领域内部一般是具有联系的语义义原组成的一个或多个网状或树状结构体，目前广泛应用的领域模型为《知网》和《同义词词林》，两词语在领域模型中的节点之间路径的最短距离是词语在语义相关程度上的度量，因此可以用来评价语义相似度，以此为基础可以进一步评价两个句子的相似程度。下面是一种基于《同义词词林》树状模型计算词语语义相似度算法[8]的核心部分：

$$Sim_{onto\log y}(w_1,w_2)=\sum_{i=1}^{n}\theta_i\mu_i(w_1,w_2)$$

其中 n 为 w_1、w_2在领域模型中最大深度，θ_i是第 i 阶权重，通常选用 1/n，$\mu_i(w_1,w_2)$为度量系数，当 w_1、w_2前 i 个父类相同时 $\mu_i(w_1,w_2)=1$，不同时 $\mu_i(w_1,w_2)=0$。该方法的问题在于词语间相似量化的前提是获取有效的网状或树状词语关系集合，目前《知网》研究取得一些实质进展，情感分析用词语集可以在《知网》网站获得。

2.2.2 语法驱动

语法驱动的方法是一种深层结构分析法 ，它对句子进行深层句法与语义分析，同时将分析结果以格框架或依存树的形式表示，然后在此基础上进行相似度计算。这种方法计算相似度虽然比较准确，但是句子表示存在需要结构分析器，目前为止，还没有一个满意的结构分析器出现。因此，纯粹的基于语法驱动的计算方法难以得到实质性的进展。

2.2.3 基于统计的方法

在论断“词语的上下文可以为词语定义提供足够信息”的基础之上，利用词语的相关性来计算词语的相似度，将词汇的上下文信息的概率分布作为词汇语义相似度计算的参照，一般是通过语料库来进行统计的。基于特定语料库进行统计的方法比较客观，综合反映了词语在句法、语义和语用等方面的相似性和差异。但需要构造大量的训练语料，计算量大，计算方法复杂，此外还存在数据稀疏和数据噪声的干扰较大等问题。

3 改进的方法

3.1 改进的编辑距离方法

改进的编辑距离方法吸取了基于语义词典的方法和编辑距离方法的优点，在计算时以词汇为基本计算单位，以《知网》和《同义词词林 a》作为语义距离的计算资源，同时赋予不同编辑操作不同的权重，减小插入操作的代价，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。

一种编辑距离的改进算法[9]的核心是对操作的代价依据《知网》和《同义词词林》的语义联系做不同权重的调整，W 表示句子 S 中的单词，W_1 表示《知网》定义的 W 同义词，W_2 表示《同义词词林》定义的 W 近义词，有句子 S 变换为 S’的操作中，规定 W 到 W 的操作代价为 0，插入的代价为 0.1，W 到 W_1 的代价为 0.4，W 到 W_2 的代价为 $Dist(w,w_2)/10+0.5$，其他操作的代价为 1，这样句子 S 到 S’编辑距离计算公式如下：

$$D(i,j)=\begin{cases}0, i=j=0\\ D(0,j-1)+w(t_j), i=0且j\neq 0\\ D(i-1,0)+w(s_i), i\neq 0且j=0\\ Min\begin{bmatrix}D(i-1,j)+1,\\ D(i,j-1)+0.1,\\ D(i-1,j-1)+Dist(w(s_i),w(t_j))\end{bmatrix}, 其他\end{cases}$$

3.2 改进的向量空间模型

基本的基于向量空间模型的方法只考虑文本中出现的单词或词组，而没有考虑单词或词组出现的顺序关系，改进的向量空间模型[10]将句子用如下三个向量表示：

向量 V1 中的各维代表每个单词的 tf×idf 值，向量 V2、V3 中的各维分别代表每个单词的 bi-gram、tri-gram 是否出现在这个句子中(0 表示没有出现，1 表示出现)。两个句子间的相似度就根据这两个句子的三个向量间分别的余弦值决定。于是句子的相似度问题就转化为回归问题。很多回归模型都可以使用，如：线性回归，Logistic 回归等。

3.3 改进的骨架依存树方法

骨架依存分析方法的目标是仅分析出句子的整体句法结构。这里整体句法结构用该句的谓语中心词及其直接支配成分来表示。分析结果可看作一棵简化了的依存树(称为骨架依存树)，其简化之处在于：骨架依存树仅限两层，第一层为根结点，即句子的谓语中心词，第二层为叶结点，是句中谓语中心词的直接支配成分，包括主干成分以及谓语中心词左右的附加成分等(总称谓语中心词的直接支配语块)；仅标注了谓语中心词与其直接支配成分之间存在依存关系以及这些成分与谓语中心词的相对位置(左或右)，而没有标注具体的依存关系的名称[11]。

改进的骨架依存树方法是在骨架依存分析方法基础之上利用语义词典进行语义相似度计算。这种方法考虑了句子的结构与语义信息，但是存在依存分析准确率不高的问题。

3.4 语法分块

该方法是将句子按语法分块，依照实验数据赋予不同权重，再计算句子间加权相似值。基本思想是在存储时，我们根据句子的五种基本句型，按照块的形式来划分。每个句子用四个块表示，即主语块(NB)、谓语块(VB)、宾语块(OB)和补充块(CB)[12]。算法如下：

$$Sim(S,S')=\lambda_1 SimNB(S,S')+\lambda_2 SimVB(S,S')+\lambda_3 SimOB(S,S')+\lambda_4 SimCB(S,S')$$

其中 λ_1，λ_2，λ_3，λ_4 为各块权重，$\lambda_1+\lambda_2+\lambda_3+\lambda_4=1$，且 λ_1，λ_2，λ_3，$\lambda_4>=0$。

次方法实际上是语法驱动基础上的整体降阶算法的实现，虽然算法公式貌似具有较好的可操作性，但由于语法驱动的适应性非常局限因此存在以下严重的问题：自然语言语法分析非常困难，目前尚没有令人满意的分析方法，导致整个方法没有良好的基础支撑；由于在不同语境环境中各块对于语义表达的重要程度的权重会有所不同，导致实际权重值无法从实验数据中准确萃取；对于复合句式以及省略表达不能进行有效的相似比较。

3.5 多目标筛选

以翻译记忆为核心的辅助翻译系统中，为使从记

忆数据中筛选得到的数据对翻译者足够有用，往往在筛选的过程中需要同时考察多个目标的合理性，于是有人提出基于多目标的融合算法，考察的目标一般细化了文本与语义相似度，从词形，词序，语法，语义，语用等角度综合考察目标语句的相似程度，从而用以度量数据的有效性。

3.5.1 综合词形、词序和词义相似度

该方法是通过计算句子间词形、词序和词义相似度并依据各自权重加权得到综合相似度，着重考察目标数据在词形、词序和词义角度与原文句子的相似程度，从而强化目标数据的对使用者的有效性，一种算法[13]如下：

$$Sim(S,S')=\lambda_1 SimP(S,S')+\lambda_2 SimO(S,S')+\lambda_3 SimM(S,S')$$

其中 λ_1，λ_2，λ_3 为各块权重，$\lambda_1+\lambda_2+\lambda_3=1$，且 λ_1，λ_2，$\lambda_3>0$，经实验数据得到的经验值一般为 0.3、0.1、0.6，SimP(S,S')、SimO(S,S')、SimM(S,S') 分别表示词形相似度、词序相似度、词义相似度。

$$SimP(S,S')=\frac{2\times[\mu_1\times ws(S,S')+\mu_2\times wl(S,S')]}{L(S)+L(S')}$$

其中 ws(S,S')和 wl(S,S')分别表示句子 S 与 S' 中相同单词个数和相似单词个数，μ_1、μ_2 为它们所代表的权重，人工设定均为 0.5，L(X)表示句子 X 中包含单词的个数。

$$SimO(S,S')=\begin{cases}1-reO(S,S')/[|G(S,S')|-1], |G(S,S')|>1\\ 1, |G(S,S')|=1\\ 0, |G(S,S')|=0\end{cases}$$

其中 G(S,S')表示两句子中仅出现一次的单词或扩展词、同义词的集合，reO(S,S')表示以上集合中的单词在 S 中位置构成的向量的分量按对应单词在 S' 中顺序排列的向量相邻分量的逆序数，以此消除词组在句子中的长距离移动的造成的值偏移。

$$SimM(S,S')=\left(\sum_{i=1}^{m} S_i\frac{1}{n}+\sum_{j=1}^{n} S_i'\frac{1}{m}\right)/2$$

$$S_i=Max[SimM(S_i,S_j')], j=1,2,\cdots,n$$

$$S_i'=Max[SimM(S_i,S_j')], i=1,2,\cdots,m$$

这里语义相似计算采用基于《知网》本体论方法。

3.5.2 其他

类似还有综合考察词法、语法和语义相似度方法和综合考察语法、语义和语用相似度的方法[14]，此类方法提高了筛选数据的有效性，但其实质仍然是以考察文本和语义两个目标为基础，只要在对此目标尤其是语义目标的评价算法有更深刻的突破，数据筛选的有效性也会随之提高，使用者将因此受益。

4 实验及结果分析

从以上翻译记忆数据筛选的方法中，我们选择其中具有较高可操作性的四种方法进行试验，以考察哪种方法可以最大程度的评价数据的有效性，这四种方法分别是改进的编辑距离(ED)、改进的向量空间(EV)、词形词序综合考察(PO)和词形词序词义综合考察(POM)的方法。

我们从翻译记忆数据库中随机选取 9900 个噪音句集，另外人工加入 100 个手工获取的参照原始句集进行较大词语、语法和语义改动的句子作为标准句集。我们对 100 个原始句集经过四种方法筛选得到的结果人工按照有效，一般，无效进行划分，最终考察筛选的数据的有效性的计算公式如下：

$$E(f)=\frac{3\cdot S(f)+1\cdot C(f)+(-2)\cdot N(f)}{300}$$

其中 S(f)、C(f)、N(f)分别表示使用方法 f 筛选得到数据的有效、一般和无效的数量，实验结果见表 1。

表 1 实验结果

方法	有效	一般	无效	有效性
ED	67	33	0	.780
EV	45	52	3	.603
PO	39	55	6	.533
POM	83	17	0	.887

由于选取的样本句集和待考察原始句集以及最终的结果评定存在很多干扰因子导致实验结果数据可能

存在一定的偏差，但是从上表中的数据对比，我们可以清楚的了解，采用文本和语义双重相似评定所筛选得到的结果数据的实际有效性要远高于另外两种仅考察文本相似的方法，语义相似度评价在筛选数据实际有效性的提高中起到重要的作用，而语义相似度可深入量化的研究还有很大的发展空间，随着《知网》等语义分析模型的深入研究，翻译记忆筛选得到的数据将对使用者提供更高质量的参考结果。

参考文献

1 袁亦宁.国外计算机翻译的发展和近况.上海科技翻译, 2002,(2):58－59.

2 Baldwin T. A look under the Hood and Road Test. Proceedings of the 15th International Japanese/English Translation Conference (IJET-15).Yokohama, Japan, 2002.

3 姚清耘.基于向量空间模型的中文文本聚类方法的研究.上海:上海交通大学, 2005.

4 董振东.知网.http://www.keenage.com.

5 梅立军,周强,臧路等.知网与同义词词林的信息融合研究.中文信息学报, 2005,19(1):63－70.

6 唐歆瑜,乐文忠,李志成,等.基于知网语义相似度计算的特征降维方法研究.科学技术与工程, 2006,6(21):3442－3446.

7 Chris H, Ding Q. A Similarity-based Probability Model for Latent Se-mantic Indexing. Proc Of 22 ACM SIGIR Conference, 1999:59－65.

8 徐小娟,徐国梁,黄新.基于本体的英汉翻译记忆系统的研究.计算机科学与工程, 2008,8(10):2708－2710.

9 林贤明,李堂秋,陈毅东.句子相似度的动态规划求解及改进.计算机工程与应用, 2004,35(21):64－66.

10 张奇,黄萱菁,吴立德.一种新的句子相似度度量及其在文本自动摘要中的应用.中文信息学报, 2005,19(2):93－99.

11 穗志方,俞士汶.基于骨架依存树的语句相似度计算模型.中文信息处理国际会议(ICCIPY98).1998.

12 周文,徐国梁.翻译记忆中语句相似度计算方法的研究.计算机应用, 2007,27(5):1210－1213.

13 基于多层次融合的语句相似度计算模型.延边大学学报（自然科学版）,2007,33(3):191－194.

14 金博,史彦军,滕弘飞.基于语义理解的文本相似度