基于多尺度句子信息的语义距离计算

王忠林
(中国电子科技集团公司第12研究所大功率微波电真空器件技术国防科技重点实验室 北京 100015)

摘 要 句子语义距离计算是许多智能系统的一项基础技术。基于多尺度分析思想,提出一个多级语义距离计算方法。首先通过词汇级语义距离算法对句子对进行初步过滤,然后对于语义距离小于一定阈值的例子进行语法分析、语义分析;获得标准语义分析框架之后,再次对框架中的中心概念进行比较,最后对通过二级筛选的句子对使用基于动态权重的语义同构算法进行计算,得到最终的语义距离。最后通过实验验证,该方法总精度达到73.3%,对相关度比较高的情况,到达和基于语义级算法相近的91.4%。

关键词 语义距离,多尺度分析,词汇级,语义级
中图法分类号 TP391.1 **文献标识码** A

Semantic Calculation Framework Based on Multiscale Analysis

WANG Zhong-lin
(Vacuum Electronics National Lab, Beijing Vacuum Electronics Research Institute, Beijing 100015, China)

Abstract The calculation of the sentence semantic distance acts a base role in many intelligent systems. Based on multiscale analysis, a multi-level semantic distance calculation framework was proposed. All sentence pairs were filtered by word-level semantic distance algorithm first, and then syntax parsing and semantic parsing were executed for the sentence pairs which semantic distances were below the threshold. After getting the standard semantic frameworks, the core conceptions in the frameworks were compared then. The final semantic distances of the sentence pairs passing the second level filter were obtained using isomorphism-based semantic distance algorithm, which could dynamically adjust its weights. Experiments show that the total precision of the method reaches 73.3%. For the cases, it has higher relevance, reaches 91.4%, similar to the semantic-level algorithm.

Keywords Semantic distance, Multiscale analysis, Word-level, Semantic level

1 概述

在信息检索、自动翻译的评估、基于实例机器翻译、信息过滤、自动问答、自动文摘系统等许多智能系统中,句子语义距离(相似度)的计算都发挥着重要的作用。随着相关系统的发展,句子相似度的计算问题越来越受到重视。研究人员已经提出许多句子相似度计算方法:基于关键词的算法、基于语义词典的算法[1]、基于语义依存的算法[2,3]、基于编辑距离的算法[4-6]、基于语境框架的算法[7]、基于属性论的算法[8]以及基于统计的算法[9,10]等。归结起来可概括为两类:基于词特征、词义特征的算法和基于句法分析、语义分析的算法。基于语法分析、语义分析的算法需要进行语法分析、语义分析,而且计算语义距离时更复杂,所以时间效率普遍比较低。而基于词汇级的算法则在时间效率方面占优,在精确度方面比较低。

为了综合词汇级和语义级算法的优点,本文提出一种基于多尺度分析思想的语义距离计算方法。该方法利用多尺度分析思想,首先使用效率高的词汇级算法给出一个粗略的语义距离,然后进行初级过滤,并只对相关度比较高的句子对进行语法分析、语义分析和语义级的距离计算,避免了对所有情况都进行语法、语义分析及语义级距离计算的工作,从而在保证计算精度的同时,提高了语义距离计算的精度。

2 方法的基本思想与过程

由粗到细或由细到粗地在不同尺度(分辨率)上对事物进行分析称为多尺度分析[11]。该方法首先在图像处理中被提出,后来在生物、材料等许多领域得到应用。在句子分析的各个阶段,随着分析的深入,句子语义信息也越全面、越细致,在此基础上对句子语义距离的计算也越准确,这和多尺度分析的思想是一致的。基于多尺度语句信息的语义距离计算方法的基本过程如图1所示。在经过分词后,首先通过词汇级的语义距离计算方法对句子进行初步筛选。对于语义距离大于一定阈值的情况,直接输出语义距离。对于语义距离小于一定阈值的情况,继续进行语法分析、语义分析。获得语义框架之后,再次对语义框架中的中心概念进行比较。如果中心概念的范畴差别比较大,直接输出预定的语义距离。只有经过再次筛选的句子才进行基于句子语义的语义距离计算,从而得到最终的结果。

到稿日期:2010-09-09 返修日期:2010-12-07
王忠林(1970-),男,博士,讲师,主要研究方向为自然语言处理、知识管理、计算机工程。

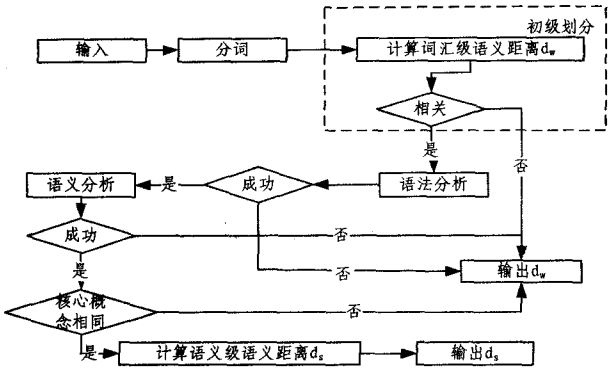

图 1　基本过程

3　算法的选择

初级划分分析句子对在比较粗的尺度上的语义距离，并决定是否进行进一步的分析。由于应用中比较关注语意距离比较小的句子对，可以在该阶段过滤语义距离较大的句子对，减少后续处理的对象数量。用于初级划分的算法首先要求时间效率比较高，而对精度要求稍低，因此选择词汇级算法进行初级过滤。基于词汇级的句子语义距离算法包括基于编辑距离的算法基于向量空间的算法、基于 N-gram 的算法以及基于最大词汇语义匹配算法[12]等。其中 N-gram 算法主要适用于英语等以字母拼写为基础的语言，不适用于汉语。基于向量空间的模型不能很好地处理包含同义词的情况。基于改进编辑距离的算法和基于最大词汇语义匹配的算法在准确率、时间效率上相差不大。基于编辑距离的算法对于存在结构迁移的情况会产生较大的误差（例如“被”字句、“把”字句等）。采用基于最大词汇语义匹配的算法可以很好地处理结构迁移的情况，但可能会导致较大语义距离的句子对被误放过的缺陷。例如对句子“函数 B 调用函数 A”和“函数 A 调用函数 B”，用基于编辑距离的算法作为初级划分算法，得到结论可能是不通过，这是合理的。但是对于“函数 B 调用函数 A”和“函数 A 被函数 B 调用”得到结论也不通过，这显然是错误的。而对基于最大词汇语义匹配的算法则会让两个句子对都通过，而进行深一步的分析，并最终计算出它们之间的精确的语义距离，就不会发生由于误判而导致语义距离比较小的情况被过滤掉的情况。由于基于最大词汇语义匹配的算法更安全，因此在初级划分阶段采用该算法计算句子的语义距离。

基于语义分析的方法在很大程度上依赖于句子语义的不同表示形式，并且要对句子进行深入的语法、语义分析。文献[13]中提出一种基于语义同构计算的语义距离计算方法，该方法通过浅层语义分析，将句子语义表示到一个标准语义框架中，然后通过计算两个句子语义框架的同构来计算两个句子的语义距离。该方法不需要进行精确的语法、语义分析，只需要根据词汇语义和形式语法结构将句子成分进行层次性聚类，然后在后续的过程中根据这种聚类的类型进行对应的比较，从而得到句子对的语义距离。该算法既考虑了句子的结构信息，又考虑了结构中不同节点对句子语义的权重是不同的，并且根据两个句子的实际结构可以动态变化，更加接近人的理解过程，因此可以更好地反映句子的语义距离，经实验验证其具有较高的精度。而基于多尺度句子语义信息的语义距离算法中的最终算法要求具有较高的精度和适应性，因此采用该算法。

4　初级划分的影响分析

4.1　初级划分对时间的影响

在初级划分阶段，通过率越高，则通过第一层筛选，进入后续过程的句子也越多，则其计算所花费的时间也越多。设总的测试用例数量为 N 对句子，每个句子分词的平均时间为 T_s，语法分析的平均时间为 T_y，语义分析的平均时间为 T_e，词汇级语义距离计算每对句子的平均时间为 T_l，基于语义同构的距离计算每对句子的平均时间为 T_i，第一层筛选的通过率为 P。因为第二层中心概念层的筛选所耗费时间比较少，可以忽略不计，并假设第二层筛选通过率为 50%，则总的耗费时间为

$$\begin{aligned} T &= 2NT_s + NT_l + 2NP(T_y + T_e) + 2NP \times 50\% \times T_i \\ &= (2NT_s + NT_l) + [2N(T_y + T_e) + NT_i]P \\ &= a + bP \end{aligned}$$

式中，$a = 2NT_s + NT_l$，$b = 2N(T_y + T_e) + NT_i$，T 与 P 成线性关系。通过分析可以看出，初级划分通过比率与总的运行时间之间基本具有线性关系。对于基于语义距离的大部分应用来说，所计算的句子对之间不相关的比率非常高，这些句子对间的语义距离都是比较大的，会在初级划分中被过滤掉，即通常 P 比较小，因此通过初级划分对于提高整个语义距离计算的效率具有十分显著的影响。

4.2　阈值对初级过滤精度的影响

无论对于基于词汇级的语义距离计算方法还是基于语法语义分析的语义距离计算方法，当两个句子之间的结构、语义差异比较大时，算法的计算精度都会不同程度地有所下降，即语义距离计算精度随着语义距离的增大而降低。而初级划分的语义距离阈值 D_s 直接影响着进入后续分析的对象的数量，从而影响着初级划分的过滤精度。当 D_s 选择比较小的值时，通过过滤的句子对是语义距离比较小的例子，相应的计算精度比较高，出现错误划分的机会比较低，因此，过滤的精度比较高；当阈值 D_s 选择比较大的值时，通过过滤的句子对中包含大语义距离的情况增多，相应的计算精度会下降，出现错误划分的几率增加，初级划分的过滤精度就会降低。为了研究阈值 D_s 对初级过滤精度的影响，通过改变阈值 D_s 得到如图 2 所示的统计结果。实验表明，随着划分阈值 D_s 的增加，初级划分的精度会随之下降。

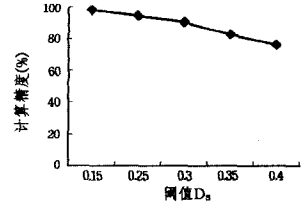

图 2　阈值 D_s 对第一层过滤精度的影响的图示结果

5　算法的有效性验证

5.1　实验概述

为了对句子语义距离算法和程序进行系统的分析、比较，笔者将语义距离空间[−1，1]划分为不同的等价类：相同(same)、相近(close)、相似(similar)、远离(far)、不相关(unrelate)、相反(inverse)、负相似(anti_similar)、负远离(anti_far)。不同的等价类对应各自的语义距离范围。测试用例共计

4717 对，分选自电子、计算机网络、计算机操作系统、机械等领域中比较有代表性的句子。将测试用例进行等价类划分，然后根据各个等价类的准确率对算法进行评测。各个等价类的测试用例数量如表 1 所列。

表 1　各个等价类用例分布

Same	close	similar	far	unrelate
591	367	543	966	1330
inverse	anti_similar	anti_far	Total	
342	340	238	4717	

为了对语义距离算法进行系统的分析，测试指标主要包括准确率和时间效率。准确率主要评价算法计算的语义距离与标准语义距离的符合程度。方案对准确率的测试主要包括两个方面：不同等价类区间内的准确率和总的准确率。不同等价类区间内的准确率使用同一等价类的测试集进行测试，并分别统计算法在该等价类测试集上的准确率。

等价类准确率采用 $P_i=\frac{C_{pi}}{C_i}$ 计算。式中，C_{pi} 表示计算结果正确的隶属于对应等价类例子的数量；C_i 表示对应等价类的测试例子总量。对所有不同等价类集测试数据进行汇总，获得总的准确率。等价类总准确率采用 $P=\frac{C_p}{C}$ 计算。式中，C_p 表示计算结果正确的隶属于对应等价类的例子数量的总和；C 表示所有测试例子总量。时间效率主要通过计算所有测试用例所用的总时间进行评测。

5.2　实验结果

评测对本文提出的基于多尺度分析方法（MSA）、基于语义同构的算法（IBSD）和基于词汇级的算法（MLSM）进行了比较。在 MSA 的初级划分中 $D_S=0.3$，语法分析采用文献[14]介绍的方法，语义分析通过映射规则将语法树中的成分映射到语义框架的相应部分。各个算法对应于不同的测试用例等价类的准确率如图 3 所示，总的准确率如图 4 所示。

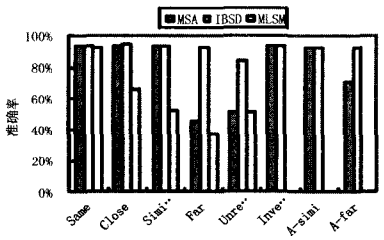

图 3　各个算法对不同等价类测试用例的准确率的图示

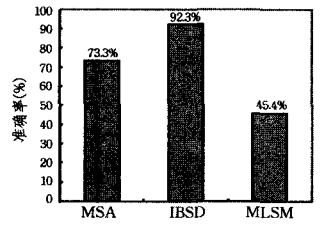

图 4　各种算法总的准确率

图 3 和图 4 显示，IBSD 算法的准确率最高，MSA 算法次之，MLSM 最差。这是由于 IBSD 对所有的测试用例都进行语法、语义分析，建立对应该句子的语义框架，并利用该框架计算两个语义框架之间的最大同构子框架和最大包容子框架，进而计算两个句子的语义距离，具有较高的准确率。而 MLSM 算法只是利用了句子词汇级的语义信息，而忽略了句子的语法语义信息，并且对于句子中的词汇平等对待，而忽略了不同词汇在表达句子语义上的差异，因此其准确率最低。MSA 对于语义距离较大的情况，其语义距离为初级划分时利用基于词汇级的语义距离计算方法得到，其计算精度较低；而对于语义距离较小的情况，采用 IBSD 的方法计算，具有较高的计算精度。综合起来，其计算精度介于 IBSD 算法和 MLSM 算法之间。但是对于相关度比较高的等价类（除远离、不相关、负远离以外的等价类），MSA 算法与 IBSD 算法具有相近的准确率：MSA 的精度为 91.4%，IBSD 算法的精度为 92.2%。

各个算法总的时间及各个部分（分词、词汇级语义距离计算、语法分析、语义分析、语义级语义距离计算，MLSM 只包含分词和词汇级语义距离计算）所用时间如图 5 所示（评测在主频 2.8G、内存 1G、操作系统为 Red Hat9.0 的主机进行，各个部分由下向上排列）。从图 5 可以看出，IBSD 算法的计算时间最长，这是因为其对所有的测试用例都相同对待，都要进行分词、语法分析、语义分析、语义距离计算等阶段，而语法分析、语义分析和基于语义同构的语义距离计算都比较复杂，要占用较多的时间，因此其效率最低；MSA 对所有的测试用例都经过分词和初级划分，然后对于通过初级划分的测试用例才进行语法分析、语义分析以及语义距离计算，节省了许多对于语义距离较大情况的不必要的分析和计算，提高了效率。MLSM 方法仅仅进行分词和基于词汇级的语义距离计算，其计算相对简单，效率最高。评测说明，基于多尺度句子信息的语义距离计算方法综合了词汇级算法时间效率高和语义级算法计算精度高的优点，综合效果比较理想。

图 6 是实验得到的初级划分通过率与总的运行时间之间的关系图示。可以看出，第一层筛选通过比率与总的运行时间之间基本具有线性关系，与理论分析一致。对于基于语义距离的大部分应用来说，所计算的句子对之间不相关的比率非常高，这些句子对间的语义距离都是比较大的，因此都会在初级划分中被过滤掉，从而更能充分体现基于多尺度句子信息的语义距离计算方法的优势。

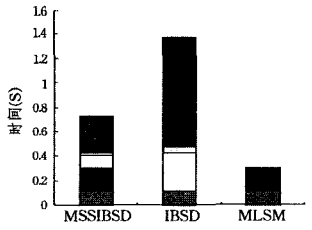

图 5　计算所有的测试用例各种算法所用的时间

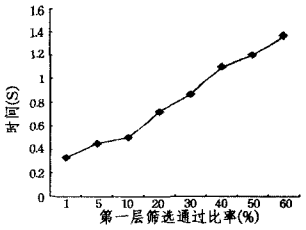

图 6　第一层筛选通过率与总时间的关系

结束语　基于多尺度句子信息的语义距离计算方法利用语义距离计算过程中不同阶段所产生的不同句子信息来由粗到精地计算句子语义距离。此方法结合了基于词汇级语义距离算法和基于语法语义分析的语义距离算法，在时间效率、计算精度方面有了一些新的特点。在计算精度方面，对于语义距离比较大的等价类，由于其句子语义距离主要通过基于词汇级算法来得到的，因此其计算精度比较低；而对于语义距离比较小的等价类，即句子相似度比较高的句子，其语义距离主要是通过基于语法语义分析的算法计算得到的，相应的计算精度比较高。在各种基于语义距离的应用中，系统往往更关

（下转第 274 页）

向与切向畸变。

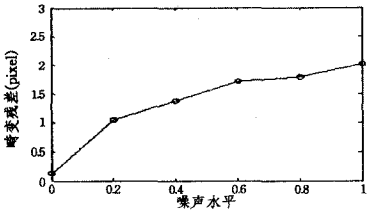

图 4　不同噪声水平下畸变残差分布

4.2　实际图像实验

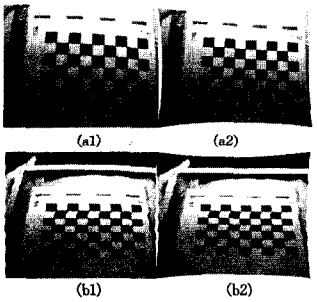

图 5　真实图像实验

表 2　图 5 中的图像计算结果

图	f_x	f_y	x_c	y_c	k_1	k_2	p_1	p_2
a	1954	1947	381.7	287.9	1.269×10^{-6}	8.722×10^{-12}	4.315×10^{-8}	8.149×10^{-8}
b	1949	1941	382.3	288.5	1.401×10^{-6}	8.150×10^{-12}	4.029×10^{-8}	8.299×10^{-8}

实际拍摄一棋盘格图像，图像大小为 764×576，畸变校正后的图像如图 5 所示，可以看出镜头畸变得到了很好的校正。表 2 为校正后的结果，可以看出标定及校正结果是合理、准确的。

结束语　基于射影几何原理，针对直线的变换特点，提出一种镜头校正并标定摄像机内参数的方法。利用直线的投影不变性计算镜头畸变参数，再通过直线的正交性标定摄像机的内参数。从实验结果可以看出，本文方法精度较高，计算量小，适用于直线特征明显的场合。实际测量中，选取多幅图像的多条直线联合校正，比单幅图像精度更高。

参 考 文 献

[1] Tsai R Y. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off the shelf TV cameras and lenses[J]. IEEE Journal of Robotics and Automation, 1987,3(4):323-344

[2] Zhang Z Y. A Flexible New Technique for Camera Calibration [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22 (11):1330-1334

[3] Weng J Y,Cohen P,Herniou M. Camera calibration with distortion models and accuracy evaluation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(10):965-980

[4] 马颂德，张正友. 计算机视觉——计算理论与算法基础[M]. 北京：科学出版社，1998:60

[5] 张靖，朱大勇，等. 摄像机镜头畸变的一种非量测校正方法[J]. 光学学报，2008,28(8):1552-1557

[6] 吴朝福. 计算机视觉中的计算方法[M]. 北京：科学出版社，2008:79

[7] 傅丹，周剑，等. 基于直线的几何不变性标定摄像机参数[J]. 中国图形图像学报，2009,14(6):1058-1063

(上接第 241 页)

注那些句子语义距离比较小的情况，而这些语义距离是通过语法语义分析的算法计算得到，其较高的计算精度保证了系统的精确率方面的性能。总体而言，基于 MSA 的方法的计算精度要比单纯基于词汇级的计算方法要高。但是由于有很大一部分的语义距离是基于词汇级的计算方法的，因此其总的计算精度要比完全使用基于同构的语义距离计算方法的计算精度要低。在时间效率方面，基于词汇级的语义距离算法在进行了词法分析后即计算语义距离，计算过程比较简单，效率比较高。IBSD 方法要经过词法分析、语法分析、语义分析、语义距离计算等多个过程，计算最复杂，因此其实现效率最低，MSA 方法会对所有的句子进行词法分析，并利用词汇级的算法计算语义距离，然后只对部分句子进行语法分析、语义分析、计算 IBSD 语义距离，其实现比单纯基于词汇级的算法要复杂，效率也低；但比单纯基于语义同构的方法要简单，效率也高。实验证明，本方法具有较高的计算精度和较快的时间效率，将有助于提高相关智能系统的性能，进一步的工作将沿着该方向进行。

参 考 文 献

[1] 秦兵，刘挺，王洋，等. 基于常问问题集的中文问答系统的研究[J]. 哈尔滨工业大学学报，2003,35(10):1179-1182

[2] 穗志方，俞士汶. 基于骨架依存树的语句相似度计算模型[C]// 中文信息处理国际会议(ICCIP'98). 1998

[3] 李彬，等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究，2003(12)

[4] 车万翔，等. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯，2004(7)

[5] Ribadas F J,Ferro M V,Ferro J V. Semantic Similarity Between Sentences Through Approximate Tree Matching[C]// IbPRIA. 2005:638-646

[6] Ilenko B,Cohenw M R,et al. Adaptive Name Matching in Information Integration [J]. IEEE Intelligent Systems,2003,18(5): 16-23

[7] 晋耀红，等. 基于语境框架的文本相似度计算[J]. 算机工程与应用，2004(16)

[8] 潘谦红，史忠植，等. 基于属性论的文本相似度计算[J]. 计算机学报，1999,22(6)

[9] Chatterjee N. A statistical Approach for Similarity Measurement Between Sentences for EBMT [C]// Proceedings of Symposium on Translation Support Systems(STRANS). 2001

[10] Poleicova G. Semantic Similarity in Content-based Filtering[C]// Proceedings of the 6th East European Conference on Advances in Databases and Information Systems. London, UK: Springer-Verlag,2002:80-85

[11] Engquist W E B. Multiscale modeling and computation[J]. Notice Amer. Math. Soc. ,2003,50:1062-1070

[12] 余刚，裴仰军，朱征宇. 基于词汇语义计算的文本相似度研究[J]. 计算机工程与设计，2006,27(2):241-244

[13] 王忠林，尹宝林. 基于同构的语义距离算法[J]. 山东师范大学学报：自然科学版，2008,23(3):50-53

[14] 王忠林，赵启阳，尹宝林. 基于确定信息的直接语法分析[J]. 中北大学学报：自然科学版，2008,29(2):131-135