一种改进的句子相似度计算模型

杨思春

(安徽工业大学计算机学院 安徽 马鞍山 243002)

【摘要】在基于实例的机器翻译中，句子相似度计算是实例匹配的有效机制。该文对基于相同词的句子相似模型作进一步的改进，包括关键词抽取，以及在句子相似度的定义中引入同义词的情形。实验结果表明，改进方法比原方法具有较高的准确率。

关 键 词 自然语言处理; 基于实例的机器翻译; 句子相似度; 基于词
中图分类号 TP391 文献标识码 A

An Improved Model for Sentence Similarity Computing

YANG Si-chun

(School of Computer, Anhui University of Technology Maanshan Anhui 243002)

Abstract In example based machine translation, sentence similarity computing is an effective mechanism for example matching. Aiming at a sentence similarity model based on same words, an improved method is put forward, including the extraction of keywords, and the induction of synonyms in sentence similarity definition. Experiment result shows that the improved method has better accuracy than the former.

Key words natural language processing; example based machine translation; sentence similarity; based on word

基于实例的机器翻译(Example Based Machine Translation，EBMT)是以双语对齐的实例库为主要知识源，输入一个待翻译的源语言句子，从实例库中查找与输入句最相似的例句，再模仿例句的译文来实现输入句的翻译。在EBMT中，实例匹配是关键，直接关系到系统本身的翻译质量。实现实例匹配的有效机制是进行句子相似度计算[1-2]，目前主要有基于词[3-5]和基于句法语义分析[6-7]的两类方法。两类方法各有优缺点，基于词的方法简单、流行，但由于仅利用句子的表层信息，即组成句子的有关词汇的词法和语义信息，因此，在判断句子整体结构相似方面有欠缺；基于句法语义分析的方法折衷考虑句子的组成词汇语义信息与整体框架结构信息，但在折衷考虑的层次上较难把握。本文研究基于词的句子相似度计算问题，并在文献[5]的基础上提出了一种改进的方法。

1 基于词的句子相似度计算

基于词的方法是目前最简单、最流行的方法，依据词的形态变化、同义词、反义词以及更进一步的语义距离来判断孤立词之间的相似度，再通过这种词间相似度的不同组合来确定句子间的相似度。文献[3]利用同义词表计算两句词之间的语义距离，进而计算两句之间的相似度。文献[4]通过正反双向比较两句相同词的个数及其位置关系，得到一个转换表达式和子块库，再通过系统预定义的翻译模式和限制条件实现两句相似度的计算。文献[5]通过比较两句相同词的个数及其位置关系，得到两句的词形相似度和词序相似度，再通过词形相似度和词序相似度计算两句的相似度。

文献[4-5]采用的方法实质上是相同的，均通过比较相同词的个数及其位置关系来计算两句的相似度。但在相似度的定义中，仅考虑了形态上相同的词，而没有考虑同义词的情形。例如，对两个简单的句子“我/ 是/ 老师/ 。/”与“他/ 为/ 学生/。/”，按照文献[5]中的方法计算则相似度很低(只有0.1)，实际上这两句

收稿日期：2004 - 09 - 01
基金项目：安徽省教育厅自然科学基金资助项目(2004kj060)；安徽省高等学校青年教师科研计划资助项目(2004jq131)
作者简介：杨思春(1970 –)，男，硕士，副教授，主要从事自然语言处理与机器翻译方面的研究.

是比较相似的。究其原因，主要是没有考虑两句中的同义词“是”和“为”。另外，也没有考虑任何句法结构信息。因此，在算法实现上虽然较为简单，但准确率却不高。基于词的方法依据句子的表层信息，通过对这些表层信息的加工也可以获得一些有用的句法结构信息，如抽取一些能够近似表达部分句法结构信息的关键性的词(以下简称关键词)。在此基础上进行句子相似度计算，就会具有较高的准确率。本文针对以上两点作了进一步的改进。

2 一种改进的方法

本文对文献[5]中的句子相似模型作了进一步改进，包括关键词抽取，以及在句子相似度的定义中考虑同义词的情形。令inp为待翻译的输入句，exa为对应的m个例句中的一个，先分别抽取inp和exa中所有的名词、代词、动词或形容词，并组成相应的关键词序列，再求出inp和exa中关键词序列的相似度，最后选取大于规定阈值的最大相似度例句作为输入句的最相似例句。

2.1 关键词抽取

由语言学知识可知，任何句子都是由关键成分(主、谓、宾等)和修饰成分(定、状、补等)构成的。关键成分对句子起主要作用，修饰成分对句子起次要作用。进行句子相似度计算时，只要考虑句中的关键成分。基于词的方法不考虑句法结构分析，因此，不能确定句子的内部成分，包括关键成分和修饰成分。在通常情况下，一个句子中作主语和宾语的多为名词或代词，作谓语的多为动词或形容词。因此，可以将一个句子中的所有名词、代词、动词和形容词作为关键词，并在计算句子相似度时只考虑这些关键词。例如，句子“我/ 当然/ 愿意/ 了解/ 她们/ 的/ 要求/ 。/”的关键词序列为“我/ 愿意/ 了解/ 她们/ 要求/ 。/”。对于特定句中的某个名词、代词、动词或形容词，不一定就是该句中的主语、宾语或谓语成分，但相对于句中所有的词构成的词序列而言，关键词序列却具有一定的句法结构信息表达能力，至少可以了解句子中的哪些词在组成句子框架结构方面是比较重要的。在此基础上进行相似度计算，比一般基于词的方法准确一些。

对句子进行关键词抽取的算法如下：

算法 1　关键词抽取算法

令S为句子，w为S中任一词，S'为S中关键词序列。

(1) for S 中任一词w do；

if　w为名词、代词、动词或形容词，then 抽取w；

读入下一词；

end for。

(2) 由S中抽取的所有关键词组成关键词序列S'。

2.2 有关定义和计算

按上述方法对句子进行关键词抽取以后，可以依据文献[5]中的句子相似模型实现任意两个句子之间的相似度计算。本文在文献[5]定义的句子相似度基础上进一步考虑了同义词的情形，有关定义和计算如下：

定义 1　词形相似度

反映两个句子形态上的相似程度，以两个句子中所含相同词或同义词的个数来衡量。设S_1、S_2为两个句子，则S_1、S_2的词形相似度为：

$$\mathrm{Sim}_{\mathrm{word}}(S_1, S_2)=2*(\mathrm{SameWord}(S_1, S_2)/(\mathrm{Len}(S_1)+\mathrm{Len}(S_2)))$$

式中　$\mathrm{SameWord}(S_1, S_2)$为$S_1$、$S_2$中所含相同词或同义词的个数；$\mathrm{Len}(S)$为句子$S$中所含词的个数。

定义 2　词序相似度

反映两个句子中所含相同词或同义词在位置关系上的相似程度，以两个句子中所含相同词或同义词的相邻顺序逆向的个数来衡量。设S_1、S_2为两个句子，$\mathrm{OnceWord}(S_1, S_2)$为$S_1$、$S_2$中所含仅一次的相同词或同义词的集合，$\boldsymbol{P}_{\mathrm{first}}(S_1, S_2)$为$\mathrm{OnceWord}(S_1, S_2)$中的词在$S_1$中的位置序号构成的向量，$\boldsymbol{P}_{\mathrm{second}}(S_1, S_2)$为$\boldsymbol{P}_{\mathrm{first}}(S_1, S_2)$中的分量按对应词在$S_2$中的次序排序生成的向量，$\mathrm{RevOrd}(S_1, S_2)$为$\boldsymbol{P}_{\mathrm{second}}(S_1, S_2)$各相邻分量的逆序数，则$S_1$、$S_2$的词序相似度为：

$$\mathrm{Sim}_{\mathrm{ord}}(S_1, S_2)=\begin{cases}1-(\mathrm{RevOrd}(S_1, S_2)/(|\mathrm{OnceWord}(S_1, S_2)|-1)) & |\mathrm{OnceWord}(S_1, S_2)|>1\\ 1 & |\mathrm{OnceWord}(S_1, S_2)|=1\\ 0 & |\mathrm{OnceWord}(S_1, S_2)|=0\end{cases}$$

定义 3 句子相似度

反映两个句子之间的相似程度。通常为一个0～1之间的数值，0表示不相似，1表示完全相似，数值越大表示两句越相似。

令S_1、S_2为两个句子，则句子相似度为：

$$\mathrm{Sim}(S_1, S_2)=\lambda_1*\mathrm{Sim}_{\mathrm{word}}(S_1, S_2)+\lambda_2*\mathrm{Sim}_{\mathrm{ord}}(S_1, S_2)$$

式中 $\mathrm{Sim}_{\mathrm{word}}(S_1, S_2)$为$S_1$、$S_2$的词形相似度；$\mathrm{Sim}_{\mathrm{ord}}(S_1, S_2)$为词序相似度；$\lambda_1$、$\lambda_2$为常数，且满足$\lambda_1+\lambda_2=1$。本文取$\lambda_1=0.9$，$\lambda_2=0.1$。

2.3 算法描述

算法 2 一种改进的算法

令输入句为inp，例句为exa(个数为m个)，输入句inp中关键词序列为inp′，例句exa中关键词序列为exa′，

(1) 抽取输入句inp中的关键词，得到inp中的关键词序列inp′；抽取每个例句exa中的关键词，得到exa中的关键词序列exa′；

(2) 求出inp′、exa′的词形相似度和词序相似度；

(3) 求出inp′、exa′的句子相似度；

(4) 选择大于规定阈值的最大相似度例句作为输入句的最相似实例。

与原算法相比，该算法中的关键词抽取部分涉及分词与词性标注(原算法仅涉及分词)，在计算词形相似度时还需要借助一部同义词词典。该算法具有以下特点：(1) 简单，所利用的信息仍为句子的表层信息。(2) 保留了原算法的优点，可以保证句子中的分句或短语整体移动后仍与原来的句子相似。(3) 比原算法准确一些，所抽取的关键词可以近似地表达部分句法结构信息。

2.4 举例

下面给出说明算法的实现和处理流程的例子。

inp：我/ 当然/ 愿意/ 了解/ 她们/ 的/ 要求/ 。/

exa[1]：我/ 认为/ 我/ 当然/ 愿意/ 了解/ 她们/ 的/ 要求/ 。/

exa[2]：当然/ 我/ 想/ 知道/ 你/ 的/ 意见/ 。/

exa[3]：我/ 很/ 想/ 知道/ 他/ 的/ 决定/ 是/ 什么/ 。/

第1步 分别抽取inp与exa[i](i=1,2,3)中的关键词，组成相应的关键词序列。

inp′：我/ 愿意/ 了解/ 她们/ 要求/ 。/

exa[1]′：我/ 认为/ 我/ 愿意/ 了解/ 她们/ 要求/ 。/

exa[2]′：我/ 想/ 知道/ 你/ 意见/ 。/

exa[3]′：我/ 想/ 知道/ 他/ 决定/ 是/ 什么/ 。/

第2步 分别求出inp′与exa[i]′(i=1,2,3)的相似度。

inp′与exa[1]′的相似度为：0.9×2×[6÷(6+8)]+0.1×1=0.874；

inp′与exa[2]′的相似度为：0.9×2×[3÷(6+6)]+0.1×1=0.550 (考虑“了解”与“知道”是同义词)；

inp′与exa[3]′的相似度为：0.9×2×[3÷(6+8)]+0.1×1=0.485 (考虑“了解”与“知道”是同义词)；

因此，inp与exa[1]相似。

3 实验结果

以人工分词的50个汉语句子作为测试集，平均句长为11.2，并按相似程度分为16类，每类有3～4个彼此相似的句子。对测试集中每个句子，分别以文献[5]中的方法和本文的方法计算其与其他所有句子的相似度，仅当相似度最大的句子与人工评判的最相似句子一致时，才认为该句的相似度计算结果正确。

实验结果如表1所示。由此可见，改进方法的准确率明显高于原方法，这主要得益于改进方法是基于关键词抽取来进行相似度计算，近似地考虑了部分句法结构信息。通过对相似度计算结果不正确的另外15个句子的分析，发现错误的原因主要在于这些句子的长度较长、结构较为复杂，所抽取的关键词在近似表达句法结构信息方面能力减弱，进而在句子相似度计算方面，准确率也随之降低。同时，通过关键词抽取的方法计算句子相似度，仅在一定程度上改善了基于词的方法的准确率，要实现准确率的全面提高，必须借助较完全的句法语义分析，例如骨架依存分析[6]和语义依存分析[7]。

表**1**　实验结果

计算方法	测试句子数	结果正确的句子数	准确率/ (%)
原方法	50	31	62
改进方法	50	35	70

4　结　论

通过关键词抽取可以明显地提高基于词的句子相似度计算方法的准确率。自动分词和词性标注的质量直接影响该方法的准确率；关键词的抽取质量直接影响该方法的准确率。

本文研究工作得到了南京大学计算机系机器翻译研究室陈家骏教授的帮助，在此表示感谢。

参　考　文　献

[1] Satoshi S, Francis B, Yamato T. A hybrid rule and example based method for machine translation[C]//Proceedings of the 4th Natural Language Processing Pacific Rim Symposium, Puket, 1997.
[2] Malavazos C, Piperidis S. Application of analogical modeling to example based machine translation[C]//Proceedings of the 18th International Conference of Computational Linguistics, Saarbrucken, 2000.
[3] 张　民, 李　生, 赵铁军, 等. 一种汉语句子间相似度的度量算法及其实现[C]//计算语言学进展与应用, 北京, 1995.
[4] 王长胜, 刘　群. 基于实例的汉英机器翻译系统研究与实现[J]. 计算机工程与应用, 2002, 38(8): 126-127.
[5] 吕学强, 任飞亮, 黄志丹, 等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报(自然科学版), 2003, 24(6): 531-534.
[6] 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), 北京, 1998.
[7] 李 彬, 刘　挺, 秦　兵, 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, 20(12): 15-17.

编　辑　黄　莘

(上接第955页)

3　结 束 语

在Internet远程教学中，数据挖掘技术可用来从大量反馈中提取有用知识、发现教与学中的潜在问题、实现智能答疑、辅助考试分析、辅助课程设计者决策等。在实践中，如何搜集有效的隐式反馈信息、如何去除由于缓存和网络迟延而对系统日志数据带来的噪声等问题，都是值得进一步研究的问题。但可以肯定的是，基于数据挖掘的Internet远程教学研究将推动远程教学的进一步完善和发展。

参　考　文　献

[1] Timothy K S. Distance education technologies: Current trends and software systems[C]//Proceedings of the First International Symposium on Cyber Worlds(CW'2002), Tokyo, 2002.
[2] 邹建梅, 刘成新. 网络课程的交互设计与控制策略[J]. 中国电化教育, 2003, 202(11): 61-65.
[3] 袁　渊, 大　军. 基于Web的远程电子实验系统的设计与实现[J]. 实验科学与技术, 2005, 3(增刊): 105-106.
[4] 王　瑶, 孙景东. 基于流媒体VOD的远程教学系统[J]. 实验科学与技术, 2006, 4(4): 41-43.
[5] 朱　明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社, 2002.
[6] 曲霖洁, 刘培玉. 基于Agent的网上教学系统的研究[J]. 电化教育研究, 2002, 105(1): 38-40.

编　辑　熊思亮