文章编号: 1003-0077(2007)03-0034-06

双语知识库中关联实例的多策略提取机制

张桂平[1,2],姚天顺[1],尹宝生[2],蔡东风[2],宋 彦[2]

(1. 东北大学 软件与理论研究所,辽宁 沈阳 110004;
2. 沈阳航空工业学院 人机智能研究中心,辽宁 沈阳 110034)

摘 要: 双语库是翻译记忆系统最重要的组成部分之一。从有限规模的双语库中提取更多的符合用户当前翻译需要的关联实例是翻译记忆技术研究的主要内容,本文首先对当前基于单一方法的实例检索算法存在的局限性进行了分析,并在对双语库进行知识化表示的基础上,提出了基于多策略的关联实例提取机制,即综合运用句子句法结构匹配、句子编辑距离计算、句子短语片段匹配、词汇语义泛化、基于扩展信息(如: 句子来源、所属专业、应用频度等信息)的优选等策略进行关联实例提取。试验结果表明,该方法有效提高了关联实例的召回数量和质量,明显改善了对用户的辅助效果。

关键词: 人工智能;机器翻译;双语知识库;关联实例;多策略提取机制;翻译记忆
中图分类号: TP391 **文献标识码**: A

The Association Example Multi-Strategy Extraction Mechanism Based on Bilingual Knowledge Corpus

ZHANG Gui-ping[1,2], YAO Tian-shun[1], YIN Bao-sheng[2], CAI Dong-feng, SONG Yan[2]

(1. Institute of Computer Software and Theory, Northeastern University, Shenyang, Liaoning 110004, China;
2. Human computer Intelligence Center, Shenyang Institute of Aeronautical Engineering, Shenyang, Liaoning 110034, China)

Abstract: Bilingual corpus is one of the most important parts in translation memory system. To extract more association examples which meet the present needs of users from limited scale of bilingual corpus is the main content of the research of translation memory technology. First of all, this paper analyzes the limits of the current example search method. Based on the knowledge representation of the bilingual corpus, this paper proposes multi-strategy based association example extraction mechanism, that is, to extract association example by using comprehensively the methods of tree matching, sentence edit-distance calculating, phrase chunk matching, lexicon semantic generalization, extended information based optimization (for instance, the information on sentence source, major belonged to, application frequency, etc.). Experimental results indicate that the method effectively improved the recall quantity and quality of association example and the assistant effect to users.

Key words: artificial intelligence; machine translation; bilingual knowledge corpus; association example; multi-strategy extraction mechanism; translation memory

1 前言

双语库规模的大小是影响翻译记忆效果的主要因素之一[1]。然而,在语料库规模相同的情况下,不同的实例提取算法为用户所提供的有帮助内容却不尽相同。一方面是受到单一检索方法本身的限制,另一方面就是双语库的信息描述内容和形式不尽

收稿日期: 2006-12-08 **定稿日期**: 2007-03-08
基金项目: 国家 863 计划资助项目(2006AA012148);国家航空基金资助项目(05J54011);辽宁省自然科学基金资助项目(20042004)

作者简介: 张桂平(1962—),女,教授,主要研究方向是自然语言处理、知识工程、知识管理。

完善。

目前对相关实例提取的研究方法主要有比较相同词汇的方法、使用编辑距离的方法、使用语义词典的方法、基于句法结构的方法[2]、以及基于统计的方法等[3]。其中每种方法所应用的计算信息是不同的,从效果上来看也存在各自的局限性。如表 1 所示。

通过表 1 可以看出,目前的相关实例提取算法主要用于对句子字符串信息、句子句法结构信息进行孤立应用,同时忽略了包括句子片断信息、句子来源、专业、频度等各类信息的综合运用。而从针对关联实例提取的应用需求来看,上述信息也十分重要。如何有效地组织上述各方面信息,并提出关联实例的多策略提取机制是本文探讨的主要内容。

表 1　基于单一策略的实例提取算法的局限性分析

相关技术	用到的相关信息或知识	主要问题分析
完全字符串比较或编辑距离计算	原文字符串信息	在双语库规模较小的情况下,匹配结果很低。此外,对于同义词间的替换不能处理。
同义词扩展匹配	词汇信息和语义词典	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。
语法结构匹配	句子语法结构信息	对于语法结构简单的句子,匹配结果会很多,甚至无从选择。而对于语法结构复杂的句子,匹配结果又会很少。
基于统计的方法	统计信息	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。

本文主要从用户实际需求角度出发,首先对双语库进行了知识化描述和存储,形成了一个结构化的、具备更多关联信息的双语知识库,并利用该知识库提出了基于多策略的实例提取机制,即综合运用句法结构匹配、句子编辑距离计算、句子短语片段匹配、词汇语义泛化、基于扩展信息(如: 句子来源、所属专业、应用频度等信息)的优选等策略进行关联实例提取。试验结果表明,该方法有效提高了关联实例的召回数量和质量,明显改善了对用户的辅助程度。同时研制开发了一套较为完整的知识库批量建立、自动积累、检索应用和共享交换工具,为基于知识的翻译系统的研究打下了基础。

2　双语知识库的建立

双语库的直接应用目标是为用户提供可参考的翻译实例。因为其中蕴含着诸多翻译可用信息,如词语和短语的对译知识,某种句式的翻译知识,专业上的特殊用法知识等等。所以若想最大程度地从双语库中挖掘出更有效的翻译参考实例,需要对其进行知识化描述和计算。

我们首先分析一下人在进行关联实例提取时所用到的各方面主要知识。These include, but are not limited to[4]:

1) 更灵活的字符串比较知识(包括整句匹配和部分(Unit/ Chunk) 匹配); 2) 相关语法知识(如句子语法结构信息); 3) 概念类比知识(如相同概念的替换: 铅笔和钢笔,苹果和鸭梨等); 4) 专业领域知识; 5) 其他语境和语用知识等。

参考 TMX 标准[5]并结合实际需求,我们定义了双语库的知识化描述方法。包括源语言句子,目标语言句子(可为多个),源语言短语信息,源语言句法树信息,提交者、提交时间、专业信息、审校信息、应用频度等信息。该描述结构为后面基于多策略的实例提取奠定了基础,同时具有良好的扩充性和交换性。下面是对双语句对进行知识化描述的部分参考内容:

在〈header〉部分出现的属性(部分)

```
datatype = "PlainText"              //数据类型
segtype = "sentence"                //分段类型
srclang = "EN"                      //指定源文本的语言
creationdate = "20050411130812"     //创建日期
creationid = "ybs"                  //创建人
changedate = "20051012353010"       //修改日期
changeid = "ybs"                    //修改人
o-encoding = "iso-8859-1"           // Original encoding
… …
```

在〈body〉部分出现的属性(部分)

```
tuid = "281"                        //翻译记忆单位的序号
usagecount = "2"                    //使用次数
lastusagedate = "20060102134012"    //最近修改时间
```

creationdate = "20051015143002" //创建时间
creationid = "ybs" //创建人
changeid = "ybs" //修改人

srctext = "The police aren't allowed to accept rewards." //源语言
srctextex = "The police <be> allow to accept reward" //源语言初步泛化结果
srcchunks = "[NP The/DT police/NN] [VP are/VBP n't/RB allowed/VBN to/TO accept/VB] [NP rewards/NNS]"
srctree = = "(TOP (S (NP (DT The) (NN police)) (VP (VBP are) (RB n't) (VP (VBN allowed) (S (VP (TO to) (VP (VB accept) (NP (NNS rewards)))))))) (. .))"
dstlang = "CH" //目标语言
dst01 = "警察是不允许接受酬金的"
… …

3 基于双语知识库的关联实例多策略提取方法

基于双语知识库的关联实例多策略提取算法主要包括句子预处理、句子短语分析、句子语法树分析、编辑距离计算、基于短语片断的实例检索、基于语法树的粗匹配、基于语义词典的词汇泛化处理和多策略实例提取算法等部分组成。在英文句子短语分析和语法树生成方面,我们使用 OpenNLP. Tools 作为主要分析器[6];在句子模糊匹配和检索引擎方面,使用 CMU 的 Lemur Tookit 检索工具包作为底层支持[7];同时使用 Hownet 的英文词汇的概念定义作为词汇泛化的基础[8,12]。主要处理流程见图 1。下面对图 1 中主要处理环节给以说明。

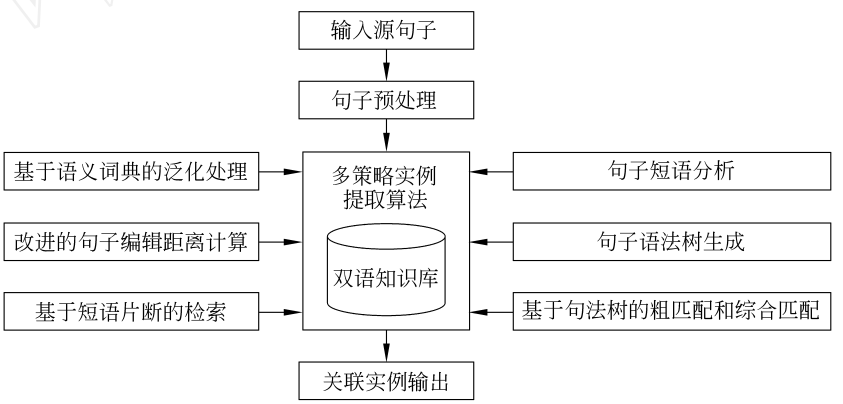

图 1 基于双语知识库的关联实例多策略提取流程示意图

3.1 句子短语分析和句法树的生成

我们使用 OpenNLP 对于用户输入的句子进行短语分析和句法分析。

例如: The police aren't allowed to accept rewards.

短语分析结果:

[NP The/DT police/NN] [VP are/VBP n't/RB allowed/VBN to/TO accept/VB] [NP rewards/NNS] ./.

句法分析结果如图 2、图 3 所示。

我们构建的句法分析树包含两种节点: 由句法范畴构成的非终结符节点(the nonterminal node that consists of the category)和由单词和它的词性构成的叶子节点(and the leaf node that consists of the word and its Part-of-Speech.)。这样的构建方法具有如下的特征: 由句法范畴(Category)构成的非终结符节点包含了句子的结构信息,和由单词和它的词性构成的叶子节点包含了词形、词性信息并能将其进行相似性计算。在图 2 中,节点 1、2、5、8、10、11、13、15 属于非终结符节点,节点 3、4、6、7、9、12、14、16 属于叶子节点。

为了便于存储和计算,树的实际存储方式定义如下: (TOP (S (NP (DT The) (NN police)) (VP (VBP are) (RB n't) (VP (VBN allowed) (S (VP (TO to) (VP (VB accept) (NP (NNS rewards)))))))) (. .))。

3.2 多策略实例提取算法

算法的总体设计思路是: 首先使用全文检索技术和编辑距离计算方法将当前输入语句在实例库中进行匹配和计算,如果存在字符串相似度在 95% (95%为用户设置值,主要是确保高相似度实例的快速检出)的以上的句子则返回该结果,否则直接利用

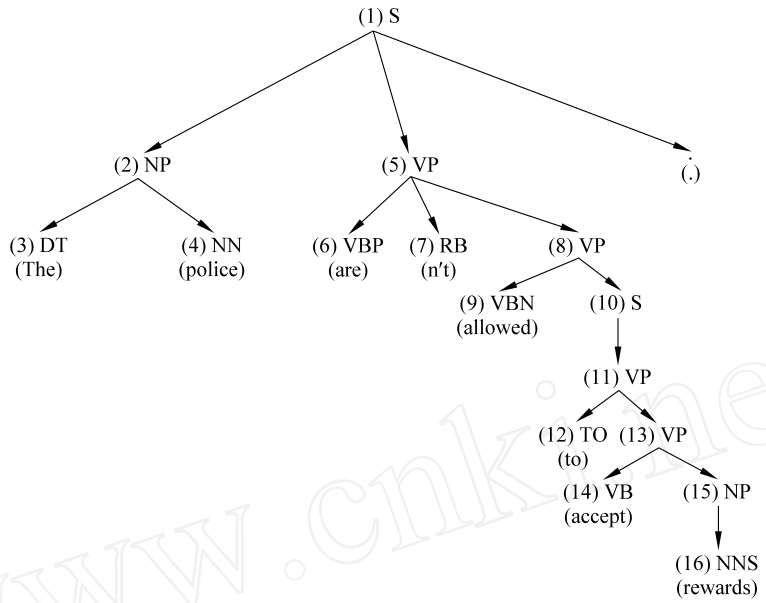

图 2　句法分析树 1

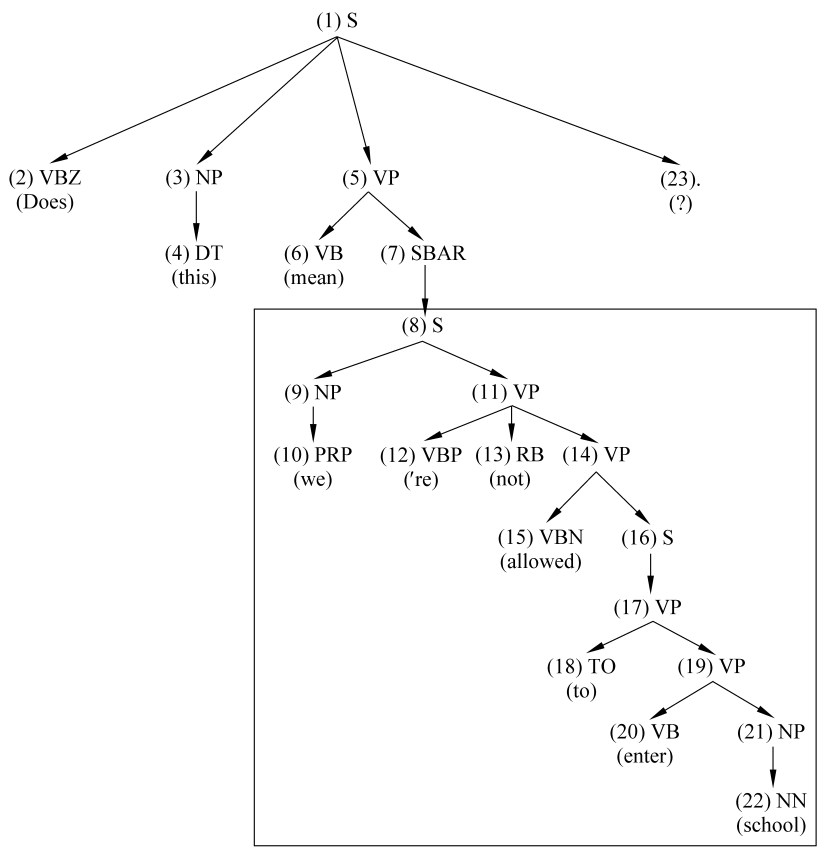

图 3　句法分析树 2

输入句的句法分析树信息在实例树库中进行粗匹配,然后在粗匹配结果集上利用句子的词形、词性、和语义信息结合句子树结构信息进行综合评价,最终确定输入句和实例句间的综合匹配度。如果不存在满足指定匹配度阈值的结果时,则返回基于短语片段检索的关联实例结果。算法的具体描述如下:

1) 改进的编辑距离计算

由于句子间编辑距离计算算法和全文检索算法已经比较成熟[3,9],所以本文不再赘述。需要指出的是,本文采用的是一种改进的编辑距离计算算法,

一方面以单词为最小编辑单位,另一方面是利用预处理模块对句子进行了初步泛化处理。如 is,was 等统一转换成〈be〉,数字信息换成〈num〉,books 还原成 book 等。

2) 基于句法树的关联实例粗匹配

基于用户输入句的句法分析树在双语知识库中已经建立的实例树(Translation Example of Tree)库(Database)进行模糊匹配,匹配度大于指定阈值(根据系统性能确定尽量低的值)的实例形成粗匹配结果集。下面举例说明输入句为"The police aren't allowed to accept rewards."时返回的一个检索结果。

实例 1: The government servants are not allowed to accept rewards.

实例 1 句法树: (TOP (S (NP (DT The) (NN government) (NNS servants)) (VP (VBP are) (RB not) (VP (VBN allowed) (S (VP (TO to) (VP (VB accept) (NP (NNS rewards)))))))) (. .)))

实例 1 参考译文: 公务员不得接受酬谢。

3) 基于多信息的综合匹配

将输入句子的分析树与粗匹配结果集中的实例分析树进行子树匹配(Sub-Tree)[10],对于两棵树中非终结符节点以及该非终结符节点在树中位置一一匹配(Corresponding)上的子树,就将其定义为结构对应子树,例如图 3 中矩形范围内的子树和图 2 中的句法树为结构对应子树。图 3 的节点 8、9、11、14、16、17、19、21 分别对应于图 2 中的句法树的 1、2、5、8、10、11、13、15。对于两树间包含多个结构对应子树的情况,我们仅取对应项最多的那个子树进行计算。对于每个对应的非终结符节点,我们计算它的节点相似度,在计算节点相似度时,我们根据每对匹配上的非终结符节点在句子中语法的重要性赋予不同权值(如句法范畴为 PP 的打分较高而 NP 的打分较低),并对每个非终结符节点所属的叶子节点的单词的词形、词性相似度和语义相似度进行计算(对于语义相似度,我们使用了知网提供的词之间的语义相似度计算方法),综合句子的词形、词性、语法和语义信息,综合得出每一个节点的匹配度。

4) 实例结果的优选及输出

将各个节点匹配度进行累加,得到句子间的综合匹配度。最后基于综合匹配度完成关联实例的排序并输出结果。对于匹配度相同的实例,我们再通过实例句对的扩展知识(如专业信息、引用频度,更新时间等)进行实例优选。

具体试验结果请参见本文第 4 节。

3.3 系统完整应用模式

基于双语知识库和多策略提取算法,我们完成了如图 4 所示系统应用模式的设计[11]。

从图 4 我们可以看出,系统以双语知识库为核心,以关联实例多策略提取模块为主要应用接口,同时建立了完整的知识积累和导入/导出机制。

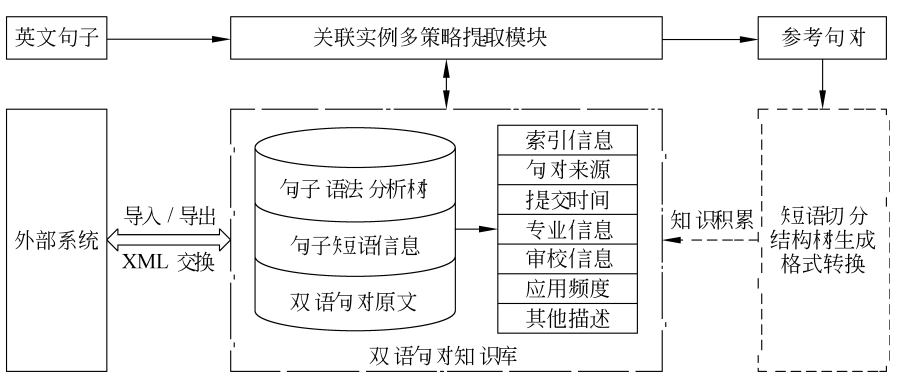

图 4 基于双语知识库的系统应用模式

4 系统测试结果

基于本文思想,我们建立了一套完整的测试系统。在系统的知识库准备方面,我们建立了一个通用领域的英汉双语知识库,知识库共包含 201 221 个英汉实例及每实例的知识化描述信息。在系统测试集方面,我们从通用领域的英文文章中随机抽取了 200 个句子作为我们的测试集。为了便于比较,我们分别使用基于改进编辑距离的实例检索算法和多策略实例提取方法进行测试,结果如表 2。

表 2 中提出的"有效关联实例"是通过人工方式判定的,主要判定原则是考查返回实例与输入句的相似性和对原句翻译的可参考性。需要说明的是,

表 2　测试结果

方　法	改进编辑距离计算的方法	多策略关联实例提取(不包含 Chunk 匹配)	多策略关联实例提取(包含 Chunk 匹配)
测试句子数	200 句	200 句	200 句
能够提取出"有效关联实例"的句子数	106 句	142 句	187 句
不能提取出有效关联实例的句子数	94 句	58 句	13 句
平均每句返回的"有效关联实例"数	0.75	0.97	1.84

我们对每个测试句返回的实例结果集只判定其前 5 句的有效性。同时为了方便比较,对于表 1 中第 3 列基于多策略实例提取方法的结果集中,单一短语匹配结果不计为有效实例,但实践表明,这方面的信息对用户的辅助翻译十分有用。这在一定程度上为用户提供了可多的可参考知识。

例如,输入句为"He bought a brand-new pair of expensive shoes."

则某一实例 The boy saw a pony with a brand-new saddle over its back. (实例中文: 男孩看到背上有崭新鞍子的小马) 不认为是有效实例——基于短语检索提取。

而实例 I bought a pair of cheap shoes which fell apart after two weeks. (实例中文: 我买了双便宜的鞋子,两个星期后就全破了。) 认为是有效实例——基于多策略提取。

通过对结果的分析,我们发现多策略关联提取算法能够充分运用句子字符串信息、结构信息和词汇语义信息等,获取令人满意的结果。此外,根据具体应用需求 (如: 辅助翻译、信息检索或辅助写作等) 可以动态调整各类信息的计算权值,从而获得更实用的结果。

5　结束语

与基于双语语料库的单一关联实例检索技术相比,基于双语知识库的多策略关联实例提取机制改进了检索效果,提取出的相关实例的数量和质量均有明显提高。我们已经将基于知识驱动的多策略关联实例检索技术应用到了英汉翻译工作室产品中,并在航空文献辅助翻译中得到了应用,取得了比在通用领域更好的匹配和辅助效果。

我们下一步的研究内容一方面是建立并完善面向专业领域的双语知识库结构,同时探讨该多策略提取机制在其他相关领域的应用,如英文辅助写作系统,自动问答系统中问题库检索以及问题与答案匹配等。

参考文献:

[1] 常宝宝, 詹卫东, 柏晓静. 服务于汉英机器翻译的双语对齐语料库和短语库建设[A]. 第二届中日自然语言处理专家研讨会论文集(北京)[C]. 2002.10, 147-154.

[2] Eiji Aramaki and Sadao Kurohashi. Example-Based Machine Translation Using Structural Translation Examples[A]. International Workshop on Spoken Language Translation (IWSLT)[C]. 2004. 91-94.

[3] 车万翔, 刘挺, 秦兵, 李生. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯,2004, 07

[4] An SDL White Paper. Knowledge-based Translation http://www.sdl.com [OL].

[5] TMX Specification http://www.lisa.org/standards/tmx/tmx.html [OL].

[6] OpenNLP.tools http://opennlp.sourceforge.net [OL].

[7] The Lemur Project http://www.lemurproject.org [OL].

[8] Dong Zhendong, Dong Qiang. HowNet and the Computation of Meaning[M]. World Scientific Publishing Co. Pte. Ltd. 2006.

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

[10] LIU Zhanyi, WANG Haifeng, WU Hua. Example-based Machine Translation Based on TSC and Statistical Generation[A]. MTSummit X, Phuket, Thailand[C]. 11-17 September 2005, 25-32.

[11] Macklovitch, Elliott and Graham Russell. What's Been Forgotten in Translation Memory [A]. In: Proceedings of AMTA 2000 [C]. Cuernavaca, Mexico.

[12] 董振东, 董强. 知网 http://www.keenae.com [OL].