服务于汉英机器翻译的双语对齐语料库和短语库建设*

常宝宝 詹卫东† 柏晓静 吴云芳† 张化瑞
北京大学计算语言学研究所，100871
†北京大学中文系，100871
{chbb,zwd,baixj,wuyf,hrzhang}@pku.edu.cn

摘要：机器翻译研究是一项十分具有挑战性的课题，机器翻译系统的翻译质量不但依赖于机器翻译方法和语言计算模型的创新性研究，也有赖于服务于机器翻译的语言资源的建设和积累，本文描述了服务于汉英机器翻译翻译的双语对齐语料库以及汉英双语短语信息数据库的描述内容以及在机器翻译中的部分应用情况。

关键词：双语对齐语料库 双语短语信息数据库 机器翻译

一、引言

从四十年代后期开始，机器翻译研究已经进行了五十多年，在这期间，机器翻译方法和系统都有了很大的进展。目前已有很多商品化的机器翻译系统在售。近年来，和 Internet 紧密结合联机翻译系统也有了很多发展，用户可以通过 Internet 访问和使用机器翻译系统，联机机器翻译系统在帮助用户阅读网上外文材料已经开始发挥作用。尽管如此，目前机器翻译系统不能令人满意的现状也不容否认。机器翻译问题仍然是一个十分具有挑战性的课题。

机器翻译系统表现不佳的原因是多方面的，其中一个很重要的原因在于资源缺乏，无论采用何种机器翻译方法，都需要大量大规模的知识资源。基于规则的机器翻译系统需要大量的规则知识、词典知识。基于统计的方法和基于实例的方法需要大规模的双语对齐语料。一个好的机器翻译系统所必备的资源往往需要经年累月的积累。

北京大学计算语言学研究所、中国科学院计算所以及清华大学自 2000 年以来在国家重点基础研究项目(973)支持下，一直在致力于开发一个“面向新闻领域的汉英机器翻译系统”。为了综合运用机器翻译研究近年来所取得的各项研究成果，该系统被设计成为一个微引擎流水线结构(Qun Liu et al. 2001)。在系统中，机器翻译的各个功能部件或同一功能的不同策略分别对应于系统中的一个个微引擎，在系统运行时，各个微引擎同时发挥作用，然后由系统进行评价综合，选择或组合出最优的处理结果。从微观角度看，目前该系统可以使采用不同方法的词法处理模块、句法处理模块等共处于一个系统之中，从宏观角度看，系统也允许把目前提出的不同的机器翻译方法以多引擎的方式组织起来。目前该系统中，不仅仅有一个基于规则的转换式机器翻译引擎，同时也有基于实例的机器翻译引擎和短语翻译引擎。这些微引擎要充分发挥作用，各项基础资源建设就显得至关重要。本文主要介绍围绕这一系统的研发我们在双语资源建设方面所进行的努力。内容主要集中在双语语料库以及双语短语信息数据库的建设。

二、机器翻译系统对双语对齐语料库和双语短语信息数据库的需求

构建汉英双语对齐语料库以及汉英双语短语信息数据库的主要目标是为汉英机器翻译系统提供资源和服务。在面向新闻领域的汉英机器翻译系统中，目前已经实现了一个基于实例的翻译引擎，基于实例的翻译引擎维护着一个汉英双语翻译实例库，在实例库中存储着句

* 本文工作得到国家重点基础研究项目(973)支持，项目号为 G1998030507-4。

子一级对齐的汉英翻译实例。翻译用户输入待翻译的汉语句子后，基于实例的翻译引擎利用用户输入的汉语句子到实例库中寻找类似的翻译实例，如果在实例库中存在类似的翻译实例，引擎则对实例中的译文部分进行修改作为待翻译句子的译文输出。对于基于实例的翻译引擎而言，实例库对真实文本的覆盖率将是影响实例翻译引擎翻译质量的一个很重要的因素，只有实例库达到一定的规模，基于实例的翻译引擎在翻译匹配时，才能保持足够的命中率，基于实例的翻译引擎才能发挥一定的作用。双语对齐语料库建设的直接应用目标正是为基于实例的翻译引擎提供翻译实例。其次，双语对齐语料库也为挖掘各种机器翻译知识提供了一个基础资源，例如可以基于双语对齐语料库挖掘词语和短语的对译知识，训练统计翻译模型等等。同时，双语对齐语料库除在机器翻译领域的应用价值外，在语言教学和研究领域、辞书自动编纂领域也有着重要的应用价值。

建设汉英双语短语信息数据库的初衷是改善基于规则的翻译系统的译文质量，根据观察在规则翻译引擎的翻译过程中，短语处理不当常常会造成翻译失败和翻译质量问题。首先待译中文句子中有时候会包含一些难以分析的短语，由于短语的分析失败，从而导致整个句子的分析失败，进而翻译失败。其次，有时候，即使分析成功，但句中的某些短语的翻译不能通过逐词对译的方式进行翻译，短语需要作为一个整体来进行翻译。有些短语即使可以通过逐词对译的方式进行翻译，但短语的各个组成成分均是多义词，很容易因为译词选择不恰当造成翻译错误。在机器翻译系统中增加有关短语结构及其译文的知识有利于减轻由于短语处理不当造成的翻译问题。

三、汉英对齐双语语料库

汉英句子对齐语料库目前描述了下面的内容：

(1) 文本属性信息

对于收录的任何一对双语文本，双语语料库都设置了一组属性来描述文本的基本信息，这些属性包括：

中文标题： 记录文本的中文标题。

英文标题： 记录文本的英文标题。

作者：记录文本的作者。

译者：记录文本的译者。

源语言：记录文本的创作语言。

文体：记录文本的文体，目前对文体的分类比较简单，所有文本被分作新闻、文学和应用文三类。

领域：记录文本的领域，目前按领域将所有的文本分为艺术、工商、政治、科技、体育、社会文化六类。

语体：记录文本的语体，所有文本按照语体分作书面语和口语。

创作时期：记录源文本的创作时期，中文文本分作古代、近代、现代、当代，英文文本分作 Old English、Middle English、Early Modern English、Present-day English。

(2) 文本结构信息

在语料库中，同时也标记了文本的篇章结构信息，这些标记包括：

中文标题：标记中文文本的标题。

英文标题：标记英文文本的标题。

子标题：标记各级子标题。

段落标记：标记文本中段落的开始和结束。

句子标记：标记文本中句子的开始和结束。

图表公式标记：标记文本中图表公式的开始和结束。

单语背景信息标记：单语背景信息指文本的编译者在文本中加入的关于文本背景的说明性信息，通常背景信息仅出现在一种语言的文本中，这样的文本内容用单语背景信息标记出来。

(3) 句子一级对齐标记

句子一级对齐标记用于标记汉英两种语言文本的句子之间的对译关系。

目前语料的所有标记均采用 XML 语言进行标记，两种语言的文本分别存放。任何一个文本均划分做两个部分，文本头部和文本体，文本属性信息记录在文本头部，文本体为标记了文本结构以及句子一级对齐关系的文本。所有标记及其含义见下表：

被标记内容		标记
文本		<TEXT>…</TEXT>
文本头	文本头	<TEXT_HEAD>…</TEXT_HEAD>
	中文标题	<CH_TITLE>…</CH_TITLE>
	英文标题	<EN_TITLE>…</EN_TITLE>
	作者名	<AUTHOR>…</AUTHOR>
	译者名	<TRANSLATOR>…</TRANSLATOR>
	文体	<STYLE>…</STYLE>
	领域	<FIELD>…</FIELD>
	语体	<MODE>…</MODE>
	创作时期	<PERIOD>…</PERIOD>
文本体	文本体	<TEXT_BODY>…</TEXT_BODY>
	段落	…
	句子	<s>…</s>
	句子级对齐单位	<a>…
	中文标题	<CH_TITLE>…</CH_TITLE>
	英文标题	<EN_TITLE>…</EN_TITLE>
	作者名	<AUTHOR>…</AUTHOR>
	译者名	<TRANSLATOR>…</TRANSLATOR>
	创作时间	<TIME>…</TIME>
	子标题	<SUBTITLE>…</SUBTITLE>
	图表公式和程序源码	<DIAGRAM>…</DIAGRAM>
	单语背景信息	<BACKGROUND>…</BACKGROUND>

下面是对齐双语语料库的一个样例，首先是中文文本，然后是对应的英文文本。中文文本和英文文本分别存放为两个文件，文本均以<TEXT>开始，</TEXT>结束。文本头部记录了有关文本的基本信息，均以<TEXT_HEAD>开始，并以</TEXT_HEAD>。文本体是包含各种标记的文本正文，其中段落以 标记开始， 标记结尾，段落的编号用 标记的属性 id 来表示。句子以<s>标记开始，</s>标记结束，段内句子编号用<s>标记的 id 属性来记录。句子对齐标记以<a>开始，结束。中文文本、英文文本中具有相同 id 值的两个句子级对齐单位为一个对齐句对。在句子一级对齐的语料库中，中文句子和英文句子的对应关系多数为一一对应关系，即一个汉语句子对应一个英文句子，但也存在多一、一多和多多对应关系，这种对应模式可以很容易地从对齐语料库中得到。任何一个对齐单位标记<a>除带有一个 id 属性外，还有一个 no 属性，中文文本中的<a>的 no 属性的值记录的是对齐单位中包含的中文句子的个数，英文文本中<a>的 no 属性的值记录的是对齐单位中包含的英文句子的个数，这样，要想得到一个对齐句对的构成模式，只须得到两个 no 属性的值就可以了。

<TEXT>
<TEXT_HEAD>
<CH_TITLE>动员新兵及新兵政治工作/CH_TITLE>
<AUTHOR>邓小平</AUTHOR>
……
</TEXT_HEAD>
<TEXT_BODY>
<s id="1"><CH_TITLE>动员新兵及新兵政治工作
</CH_TITLE></s>
<s id=1><Time>（一九三八年一月十二日）</Time></s>
<s id=1><Subtitle>一</Subtitle></s>
<s id=1>当前的战局，是处于暂时的局部的失利的境况，决不是抗日自卫战争的最后失败。</s><s id="2">战争的最后胜败，要在持久抗战中去解决。
</s>
……</TEXT_BODY>
</TEXT>
<TEXT>
<TEXT_HEAD>
<EN_TITLE>MOBILIZE NEW RECR…</EN_TITLE>
<AUTHOR>邓小平</AUTHOR>
……
</TEXT_HEAD>
<TEXT_BODY>
<s id="1"><EN_TITLE>MOBILIZE NEW RECRUITS AND CONDUCT POLITICAL WORK AMONG THEM</EN_TITLE></s>
 <s id="1"><Time>January 12, 1938</Time></s>
<s id=1><Subtitle>I</Subtitle></s>
<s id=1>Currently we are suffering a temporary and partial setback in our defensive war against Japan, but this is not final defeat.</s><s id="2">The final outcome of the war will be determined by a protracted war of resistance.</s>
……
</TEXT_BODY>
</TEXT>

四、双语短语信息数据库

双语短语信息数据库中主要收录汉语短语以及汉语短语的英语译文，双语短语信息数据库的建设不同于双语词典的建设。同短语不同，一般认为词的数量是有限的，而短语却是无限的。在建设双语短语信息数据库时，主要参考了下面的原则：（1）短语的收录应该考虑到短语的出现频率，收录的短语应该是高频短语，频率较高的短语再次出现的概率一般也比较大，收入这样的短语对机器翻译系统的改进效果应该比低频率短语明显。（2）短语的收录也应该有一定的针对性，即针对那些机器翻译系统按照常规方法不宜处理的短语，如上文所述的不能通过逐词翻译得到译文的短语、结构自动分析难度较大的短语以及短语组成成分歧义比较严重的短语。（3）短语应尽可能来自于真实文本。值的说明的是目前已经建成的短语库

只是部分的遵循了上述原则，例如到目前为止，完全来自于真实文本的短语库只有 8000 多个，这 8000 多个短语来自于上述双语对齐语料库。

对于双语短语信息数据库而言，要想成为汉英机器翻译系统中一个有益的资源，仅仅记录汉语短语以及英语译文本身是不够的，还必须描述有关双语短语的其它一系列信息，这些信息对基于规则的机器翻译系统而言，通常是十分重要的。也只有在双语短语库中描述了这些信息，在机器翻译系统中也才能得到短语的内部组成结构和向外组合能力，短语库中的双语短语资源才能无缝的进入到翻译程序的翻译过程中。

目前对于双语短语库中的每对短语，描述了下列信息：

（1） 汉语短语
记录汉语短语字串。

（2） 汉语短语的切词信息
记录汉语短语词语切分结果。

（3） 汉语短语的词性信息
记录汉语短语各组成单词的词性信息，双语短语信息数据库的汉语部分词性标注采用了北京大学计算语言学研究所词性标注集，选择了其中了 26 个词性标记。有关该标注集的情况参见俞士汶(1999)。

（4） 汉语短语的短语类信息
短语类描述了短语向外扩展的能力，目前短语库设置的短语类有名词短语(np)、动词短语(vp)、数量短语(mp)、处所短语(sp)、时间短语(tp)、形容词性短语(ap)、副词性短语(dp)以及小句(dj)共八个类别。

（5） 汉语短语的内部结构关系
内部结构关系描述的是汉语短语最高层直接组成成分之间的句法关系，目前短语库设置了九种短语内部结构关系，分别是：定中结构(DZ)、述宾结构(SB)、述补结构(SBU)、联合结构(LH)、主谓结构(ZW)、连谓结构(LW)、状中结构(ZZ)、的字结构(DE)和介宾结构(JB)。

（6） 汉语短语的中心词信息
描述了短语中在句法层面起中心作用的词，对于语言学上一般认可的向心结构，中心词的规定和语言学的规定的相同，对于所谓离心结构，则采用规定的方式设定中心词，短语库中中心词前面加标记“!”。

（7） 英语短语
记录英语短语的词串。

（8） 英语短语的词性信息
记录英语短语的各组成单词的词性信息，词性标记采用了宾州词性标记集,有关该标记集的情况可参见宾州(WEB)。

（9） 英语短语的短语类信息
记录英语短语的短语类信息，目前对于英语短语，共设置了名词性短语(NP)、动词性短语(VP)、形容词性短语(AdjP)、副词性短语(AdvP)、介词短语(PP)和小句(CS)六类。

（10） 英语短语的中心词信息
记录了英语短语的中心词，同样在短语的中心词前加标记“!”。

（11） 若干属性信息，除上述信息外，双语短语信息数据库还描述了短语是否构成专名以及习用语等信息。

目前短语库采用 Accsess 数据库方式管理，下图是双语短语信息数据库的一个片断。

ZH_PHRASE	ZH_PHR	ZH_GI	EN_PHRASE	EN_PH
地方/n !法规/n	np	DZ	local/JJ !legislation/NN	NP
地方/n 各级/r !财政/n	np	DZ	local/JJ !budgets/NNS at/IN various/JJ levels/NNS	NP
地方/n 各级/r !人大/n	np	DZ	the/DT local/JJ people/NNS ' s/POS !congresses/NNS	NP
地方/n 各级/r 人大/n 代表/n 中/f 的/u 女性/r	np	DZ	!ratio/NN of/IN women/NNS deputies/NNS in/IN loca	NP
地方/n 各级/r 人民/n !法院/n	np	DZ	people/NNS ' s/POS !courts/NNS at/IN various/JJ le	NP
地方/n 各级/r 人民/n !政府/n	np	DZ	local/JJ people/NNS ' s/POS !governments/NNS at/IN	NP
地方/n 各级/r 政府/n 的/u 配套/a !资金/n	np	DZ	supporting/VBG !money/NN provided/VBN by/IN the/D	NP
地方/n !官员/n	np	DZ	regional/JJ government/NN !officials/NNS	NP
地方/n 国家/n 权力/n !机关/n	np	DZ	local/JJ organ/NN of/IN state/NN !power/NN	NP
地方/n !经济/n	np	DZ	local/JJ !economy/NN	NP
地方/n 劳动/v 争议/v 仲裁/v !委员会/n	np	DZ	local/JJ labor/NN dispute/NN arbitration/NN !comm	NP
地方/n !民族主义/n	np	DZ	local/JJ ethnic/JJ !chauvinism/NN	NP
地方/n 配套/a !资金/n	np	DZ	supporting/VBG !capital/NN to/TO be/VB contribute	NP
地方/n 人民/n 代表/n !大会/n	np	DZ	local/JJ people/NNS ' s/POS !congress/NN	NP
地方/n 社会/n 经济/n 文化/n 事业/n 的/u !发	np	DZ	the/DT !development/NN of/IN local/JJ social/JJ ,	NP
地方/n !台/n	np	DZ	local/JJ broadcasting/NN !stations/NNS	NP
地方/n !行政/n	np	DZ	local/JJ !administration/NN	NP

五、应用

（1）在基于实例的或基于存储的翻译引擎中的使用

正如前文所述，在基于实例或基于存储的翻译引擎中，需要大量的翻译实例，这些翻译实例至少要以句子一级的对齐方式存储在基于实例的引擎的翻译实例库或基于存储的翻译引擎的翻译记忆库中。在汉英双语对齐语料库中，已经标记了句子一级的对齐信息，从中可以很容易地抽取出汉英对照的双语句子对儿，这些句子对儿可以直接存放在基于实例的翻译引擎的实例库或基于存储的翻译引擎的翻译记忆库中使用。目前北京大学计算语言学研究所已经积累的句子一级对齐的双语语料已经达到 65000 余对[1]，目前双语对齐语料库的建设仍在进行当中，近期汉英双语对齐句子对儿的规模将会超过 100000 个，这个资源对面向新闻领域的汉英机器翻译系统中基于实例的翻译引擎将是一个有力的支持。

双语句子对齐句子对儿对进一步的基于实例翻译研究也提供了一个基础资源，为了提高实例翻译的命中率，实例库中的翻译实例还需要进一步的对齐处理，例如进行短语和小句一级的对齐，如果进行了短语和小句一级的对齐工作，在进行基于实例的翻译时，匹配的单位就可以下降到短语或小句层次，这会进一步提高实例库中实例的利用率，翻译引擎不仅可以以句子为单位匹配翻译实例库，也可以以短语为单位匹配实例库。我们已在短语对齐方面作了一些探索性工作，对部分双语句子对儿进行了基本名词短语和最常名词短语的对齐工作。

（2）双语相关集列系统

相关集列系统(concordancer)系统是伴随着语料库语言学的发展而出现的一种语料库应用工具，这个工具集成提供一组面向语料库的检索手段和搭配分析手段，国际上目前已经有许多成熟的单语相关集列系统存在，这些系统在单语语言研究和单语辞书编纂方面发挥了非常重要的作用。

双语对齐语料库包含两种语言的文本，同时还包含两种语言在句子一级的对译关系，双语相关集列系统的建立将会显示出双语对齐语料在基于实例翻译之外的其它应用价值。首先面向双语的相关集列系统包含单语相关集列系统的所有功能，除此以外，双语相关集列系统将会有效地发挥出双语对齐语料库在机器辅助翻译、双语词典编纂、双语对比研究的应用价值。

同全自动机器翻译研究的目标不同，机器辅助翻译的目标是为翻译工作者提供一组翻译辅助工具，利用这组翻译工具，翻译工作者的翻译效率将会得到明显提高，双语相关集列系统也是这样一个工具，它延伸了双语词典在翻译过程中的作用。翻译工作者在进行翻译时，常常会为某个词或某个表达难以翻译而大费斟酌，传统的解决办法是查询双语词典，然而在双语词典中，通常只记录了词汇一级的对译知识，特定表达的特定翻译往往不能通过查询词

[1]北京大学长期从事机器翻译评价系统的研究，并设计实现了一个基于孤立测试点的自动机器翻译评价系统，为了评价译文质量，评价系统采用了句子一级对齐的汉英双语测试集，这部分语料目前也已经被继承进双语对齐语料库中。

典的方式得到，即使对于词汇层次，词典中对译知识往往是脱离语言环境的，利用双语相关集列系统，翻译工作者不但可以查询词汇的正确翻译，也可以查询某种表达的翻译，对每一个词或特定语言表达，不但可以找出其正确翻译，还可以在系统提供的上下文中更好的把握翻译的准确性。

（3）短语翻译引擎

汉英双语短语信息数据库目前规模已经达到 50000 短语对，并且已经以一个短语翻译引擎的形式在面向新闻领域的汉英机器翻译系统中发挥出初步作用。在面向新闻领域的汉英机器翻译系统中，多个翻译引擎同时发挥作用，线图为组合各种引擎翻译结果提供了一个公共的数据结构。对每一个待翻译的汉语句子，短语翻译引擎首先查询双语短语信息数据库，找出在句子中包含的在短语库中已经描述过的短语，并从短语库中提取这些短语的结构信息以及译文的结构信息作为部分处理结果加入线图，这样做的好处是，对于短语库中已经描述过的短语，无需再进行常规的分析处理，避免了由于这些短语分析失败而造成翻译错误，同时也无需对这些短语的翻译再进行常规的转换处理，短语库中记录的英文对译短语及其结构信息将直接作为该短语的翻译结果加入到译文结构中。

六、结束语

本文介绍了北大计算语言学研究所为提高汉英机器翻译系统翻译质量而进行的两项资源建设，分别为汉英对齐双语语料库和汉英短语信息数据库，介绍了这两项资源的描述内容、编码以及部分应用情况。目前这两项资源的建设还在进一步建设之中，我们期望这些资源建设将会为机器翻译系统研究提供积极的支持作用。

在双语对齐语料库以及汉英短语信息数据库的建设过程中，吴拥华、叶嘉明、王文涛等同学也先后提供了软件技术支持，刘群、王厚峰等老师也参加了双语短语信息数据库的一期加工工作，特此致谢。

参考文献

柏晓静，常宝宝，詹卫东 (2002)，构建大规模的汉英双语平行语料库，2002 全国机器翻译讨论会，已录用

Qun Liu, Baobao Chang, Weidong Zhan, Qiang Zhou (2001), A News-oriented Chinese-English Machine Translation System, International Conference on Chinese Computing (ICCC2001), Singapore, 2001

吴云芳，常宝宝，詹卫东(2002)，汉英双语短语信息数据库的构建，第一届计算语言学研讨会论文集，2002 年 8 月，北京

俞士汶，朱学锋，王惠，张芸芸(1998)，现代汉语语法信息词典详解，清华大学出版社，1998

俞士汶(1999)，现代汉语语料库加工——词语切分与词性标注规范与手册. (内部资料)，1999

宾州(WEB)，宾州树库词性标柱集及手册，见 http://www.cis.upenn.edu/~treebank

Building aligned parallel corpus and bilingual phrase information bank for Chinese-English Machine Translation

CHANG Baobao ZHAN Weidong† BAI Xiaojing WU Yunfang† ZHANG Huarui

Institute of Computational Linguistics, Peking University,

†Department of Chinese language and literature， Peking University,
Beijing, 100871
{chbb,zwd,bxj,wuyf,hrzhang}@pku.edu.cn

Abstract: Machine Translation has been proved to be a very challenging task. The quality of Machine Translation depends not only the advances in methodology of Machine Translation and computing modeling of language, but also a very large scale language knowledge bank. This paper gives a detailed description of building of an aligned parallel corpus and an bilingual phrase information bank and their application in the research of Machine Translation.

Keywords: Aligned Parallel Corpus, Bilingual Phrase Information Bank, Machine Translation