无双语词典的英汉词对齐

吕学强[1)]　吴宏林[2)]

[1)]（北京大学信息科学技术学院计算语言学研究所 北京　10871）

[2)]（东北大学信息科学与工程学院计算机软件与理论研究所　沈阳　110004）

摘　要　该文提出了一种基于语料库的无双语词典的英汉词对齐模型. 它把自然语言的句子形式化地表示为集合，通过集合的交运算和差运算实现单词对齐，同时还考虑了词序和重复词的影响. 该模型不仅能对齐高频单词，而且能对齐低频单词，对未登录词和汉语分词错误具有兼容能力. 该模型几乎不需要任何语言学知识和语言学资源，使语料库方法可独立应用. 实验表明，同质语料规模越大，词对齐的正确率和召回率越高.

关键词　自然语言处理，双语语料库，词对齐，最小求交，最小求差

中图分类号：　TP391

Aligning English-Chinese Words without Bilingual Dictionary

LŰ Xue-Qiang[1)]　　WU Hong-Lin[2)]

[1)](Institute of Computational Linguistics, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

[2)](Institute of Software and Theory, School of Information Science and Engineering, Northeastern University, Shenyang 110004)

Abstract　One of the bilingual corpus processing methods is the alignment of two languages on each linguistic level. Much research on word alignment between Indo-European languages has been done before, however, much less has been done on English-Chinese alignment. This paper proposes a corpus-based model for word alignment between English and Chinese. It formalizes natural languages into sets, the intersection and difference of which implement the word alignment, and, at the same time, the effect of word order and repetition is considered. The model includes a set of sub-models: minimum intersection model, minimum difference model, hybrid model, mono-directional model, bi-directional model, union model, and surrounding model. The English−>Chinese mono-directional model is used to generate 1-m parallels, and the English<−Chinese model is used to generate n-1 parallels. The union model and surrounding model are used to generate n-m parallels from the 1-m and n-1 parallels. The intersection of any two generated parallels in a sentence pair is empty, and the parallels themselves are minimum. This method can be used for alignment of both high-frequency words and low-frequency words, and is tolerant with Chinese word segmentation errors and unknown words. The typical characteristic of this model is that it needs a few linguistic knowledge and resource. Experimental results show that the larger of homogeneous corpus scale, the higher precision and recall can be got.

Keywords　natural language processing, bilingual corpora, word alignment, minimum intersection, minimum difference

1 引　言

随着语料库语言学的发展，双语语料库因含有双语信息而受到越来越多的重视. 生语料一般不能直接在自然语言处理中应用，需要进行加工，从中抽取有用信息. 对双语语料库的加工方式之一就是在各层次实现对齐. 双语语料库的对齐按粒度可分为篇章级[1]、段落级[2]、句子级[3-8]、短语级[9]和单词级[10-15].

本课题得到国家自然科学基金(60083006)、国家“九七三”重点基础研究发展规划项目(G19980305011)和国家“八六三”高技术研究发展计划项目(2001AA114019, 2001AA114210, 2002AA117010-08)资助. 吕学强，男，1970 年生，博士,主要研究方向为自然语言理解、机器翻译、语料库语言学. E-mail: studystrong@sohu.com. 吴宏林，男，1978 年生，硕士，主要研究方向为自然语言理解、机器翻译.

单词对齐(也称词对齐)就是在双语句对中把每个单词和它的译文建立对应关系.印欧语种间的单词对齐已得到较多的研究[10-12]，而英汉单词对齐的研究较少[13-15].单词对齐过程中通常用到同源词信息、双语词典、高频词的共现信息等.其中同源词信息不能用于英汉单词对齐中.由于自然语言翻译的灵活性和双语词典的有限性，词典译项对真实文本的覆盖率很低，仅用双语词典进行机械匹配来对齐英汉单词无法达到满意的效果.而且利用双语词典对齐英汉单词不能兼容汉语分词错误和未登录词，当汉语分词发生错误和出现未登录词时对齐结果必然是错的.对于含有大量双语句对的双语语料库可利用统计共现信息对齐单词，但共现方法通常只能对齐高频单词，而对低频单词无能为力.单词对齐中还面临一个问题是部分对应，即只把单词与其部分译文对应上，而没有与全部译文对应上.

本文提出的无双语词典的英汉词对齐模型吸收了翻译模板自动获取[16-19]的思想，是以语料库为基础的方法.它把自然语言的句子形式化地表示为集合，通过集合的交运算和差运算实现单词对齐，同时还考虑了词序和重复词的影响.该方法不仅能对齐高频单词，而且能对齐低频单词，对汉语分词错误和未登录词具有兼容能力，并且单词最终是与它的全部译文对应.

2 词对齐的形式化定义

2.1 自然语言句子的表示形式

定义 1. 设自然语言 L 的单词和符号(以下统称单词)的集合为 LW，则 L 中的一个句子 S 可表示为 LW 中元素的一个序列，即 $S=w_1w_2w_3\cdots w_n$.其中 w_1、w_2、w_3、…、$w_n\in LW$,并且是有序的.此表示形式称为句子 S 的原始表示形式,记为 $S^0=w_1w_2w_3\cdots w_n$.

定义 2. 设 S 的原始表示形式为 $S^0=w_1w_2w_3\cdots w_n$.若把单词 w_i 及其位置 p_i 表示为一个二元组$\langle w_i,p_i\rangle$，则 S 可表示为 $S^2=\{\langle w_1,p_1\rangle,\langle w_2,p_2\rangle,\langle w_3,p_3\rangle,\cdots,\langle w_n,p_n\rangle\}$.此表示形式称为 S 的二元集合表示形式.

为描述不同的语言，可增加一维表示语言种类的分量从而构成三元组.为论述方便，本文没有使用表示语言种类的分量.

二元集合表示形式与原始表示形式是等价的.即根据原始形式可求出二元集合形式；根据二元集合形式，也可求出原始形式.例如已知 S^0=the dog is running after the cat. 把各单词加上位置序号可求出二元集合表示形式为$S^2=\{\langle the,1\rangle,\langle dog,2\rangle,\langle is,3\rangle,\langle running,4\rangle,\langle after,5\rangle,\langle the,6\rangle,\langle cat,7\rangle\}$.若已知二元集合表示形式为 $S^2=\{\langle after,5\rangle,\langle cat,7\rangle,\langle dog,2\rangle,\langle is,3\rangle,\langle running,4\rangle,\langle the,1\rangle,\langle the,6\rangle\}$，把各元素按位置分量排序后得到的单词序列即为原始形式 S^0=the dog is running after the cat.

定义 3. 设 S 的原始表示形式为 $S^0=w_1w_2w_3\cdots w_n$.若不考虑单词的顺序，并把相同的单词看作一个单词，则句子可表示为单词的集合,S 可表示为 $S^1=\{m_1,m_2,m_3,\cdots,m_t\}$,其中 S^1 中出现的单词与 S^0 中出现的单词相同.此表示形式称为 S 的一元集合表示形式.

一元集合表示形式只能描述 S 的部分性质，有以下弱点：(1) 不能描述与词序相关的性质；(2) 不能完全描述重复单词的性质.对于非重复单词,S 的一元集合表示形式的一个元素与二元集合表示形式中的一个元素对应.对于重复单词，S 的一元集合表示形式的一个元素与二元集合表示形式中的多个元素对应.例如若 S^0=the dog is running after the cat,则 $S^1=\{$ after, cat, dog, is, running, the$\}$，$S^2=\{\langle the,1\rangle,\langle dog,2\rangle,\langle is,3\rangle,\ \langle running,4\rangle,\langle after,5\rangle,\langle the,6\rangle,\langle cat,7\rangle\}$.$S^1$ 中的 after、cat、dog、is、running 分别与 S^2 中的$\langle after,5\rangle$、$\langle cat,7\rangle$、$\langle dog,2\rangle$、$\langle is,3\rangle$、$\langle running,4\rangle$对应，而 S^1 中的 the 与 S^2 中的$\langle the,1\rangle$和$\langle the,6\rangle$对应.

在不会发生混淆的情况下，一个句子的三种表示形式 S^0、S^1、S^2 都用 S 表示.

定义 4. S 的一元集合表示形式和二元集合表示形式统称集合表示形式.

引入集合表示形式以后，可用集合理论来研究句子的性质，为自然语言的研究提供了一个强力工具.

2.2 词对齐的相关概念

设 BC（bilingual corpora）是已实现句子对齐的英汉双语语料库.本文把 BC 看作由英汉对应句子构成的集合.每个英汉对应句子(也称句对)S 表示为 S =ES<->CS,其中 ES 和 CS 是互为译文的英语句子和汉语句子.在 BC 中，S 以原始表示形式存放.根据原始表示形式可求出 S 的二元集合表示形式和一元集合表示形

式.

定义 5. 设 S=ES<->CS 是句对 S 的集合表示形式. 若 P=EP<->CP, 其中 EP⊆ES, CP⊆CS, 并且满足下列条件时，称 P 是 S 的一个对应(parallel), 记为 P|S.

(1) 非空性. EP∪CP≠Φ.

(2)互译性.EP 与 CP 在 S 中互为译文. 若 EP 的译文在 CS 中不存在，则 CP=Φ；若 CP 的译文在 ES 中不存在，则 EP=Φ.

若 EP =Φ或 CP =Φ, 则称 P 为空对应，否则称非空对应. 若 EP 中含 a 个元素，CP 中含 b 个元素，则称 P 的匹配模式为 a-b，记为 match(P)= |EP|-|CP|=a-b. 其中|EP|、|CP|表示 EP、CP 中元素的个数，“-”是连字符.

由计算机自动生成的对应不一定满足互译性，称为候选对应，在不会发生混淆的情况下也简称为对应.

定义 6. 设 P= EP<->CP 是 S 的对应，若 EP 是 CP 在 S 中的部分译文或 CP 是 EP 的在 S 中的部分译文，则称 P 为部分对应；若 EP 是 CP 在 S 中的全部译文且 CP 是 EP 的在 S 中的全部译文，则称 P 为完整对应.

若 EP 中的英语单词 e 的译文在 CS 中不存在，则规定{e}<->Φ是完整对应；若 CP 中的汉语单词 c 的译文在 ES 中不存在，则规定Φ<->{c}是完整对应.

定义 7. 设 P_1=EP_1<->CP_1、 P_2=EP_2<->CP_2是 S 的对应. 则 P_1与 P_2的交为 $P_1∩P_2$= $EP_1∩EP_2$<-> $CP_1∩CP_2$; P_1与 P_2的并为 $P_1∪P_2$= $EP_1∪EP_2$<-> $CP_1∪CP_2$，P_1与 P_2的差为 $P_1\P_2$= $EP_1\EP_2$<-> $CP_1\CP_2$.

定义 8. 设 P= EP<->CP 是 S 的对应，若 P 不能表示为两个或两个以上完整对应之并，则称 P 是 S 的最小完整对应.

定义 9. 当 S 的对应的集合 A={P_1, P_2, P_3, …, P_k}满足以下条件时，称 A 是 S 的对齐.

(1)完备性. $\bigcup_{i=1}^{k} CP_i = CS, \bigcup_{i=1}^{k} EP_i = ES$. 且对任何 1≤i≤k， 1≤j≤k，i≠k，有 $CP_i∩CP_j$=Φ, $EP_i∩EP_j$=Φ.

(2) 最小完整性. 对任何 1≤i≤k，P_i是 S 的最小完整对应.

词对齐的最终目的是根据句对 S 的原始形式找到 S 的二元集合形式的对齐结果. 如果先求出了 S 的一元集合形式的对齐则需进一步转化为二元集合形式. 一个一元集合形式的最小完整对应 P 中若只含非重复单词，因为非重复单词只与句对 S 的二元集合表示形式中的一个元素对应, 所以 P 转化为二元集合形式后仍是最小完整对应. 一个一元集合形式的最小完整对应 P 中若含重复单词，因为重复单词与句对 S 的二元集合表示形式中的多个元素对应, 所以 P 转化为二元集合形式后可能不再是最小完整对应，需进一步分解为最小完整对应. 本文提出的方法就是先求出 S 的一元集合形式的对齐，然后再根据一元集合形式的对齐求出二元集合形式的对齐.

3 模型

基于语料库的无双语词典的英汉词对齐模型包括最小求交模型、最小求差模型、混合模型、单向模型、双向模型、聚合模型、夹逼模型等多个子模型. 各子模型中除夹逼模型使用句子的二元集合表示形式外，其它子模型都使用句子的一元集合表示形式.

3.1 最小求交模型

定义 10. 设 S_i、S_j∈BC. S_i= ES_i<->CS_i={e_{i_1}, e_{i_2}, e_{i_3}, …, $e_{i_{n_i}}$}<->{c_{i_1}, c_{i_2}, c_{i_3}, …, $c_{i_{m_i}}$}, S_j= ES_j<->CS_j={e_{j_1}, e_{j_2}, e_{j_3}, …, $e_{j_{n_j}}$}<->{c_{j_1}, c_{j_2}, c_{j_3}, …, $c_{j_{m_j}}$}. 若两个句对的交集为 $S_i∩S_j$=P=E<->C={e_1, e_2, e_3, …, e_n}<->{c_1, c_2, c_3, …, c_m}. 则称 S_j求交支持 E<->C|S_i. 它表示在 S_j支持下通过求交可决定 E 中英语单词 e_1、e_2、e_3、…、e_n与 C 中汉语单词 c_1、c_2、c_3、…、c_m对应. 若语料库 BC 中有 x 个句对求交支持 E<->C|S_i，则称 BC 中 E<->C|S_i的求交支持度为 x. 记为 Sup_{int}(E<->C|S_i)=x.

在 S_i= ES_i<->CS_i中, 对于任何 E(Φ⊂E⊆ES_i), E 中英文单词的对译词集合 C 可由下述公式决定：

$$E \leftrightarrow C|S_i = \begin{cases} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\text{Arg max}}\, \text{Sup}_{int}(E \leftrightarrow C|S_i), & \text{当} \max\limits_{\substack{C\subseteq CS_i \\ C\neq\varnothing}} \text{Sup}_{int}(E \leftrightarrow C|S_i)>0 \\ E \leftrightarrow \varnothing|S_i & ,\text{当} \max\limits_{\substack{C\subseteq CS_i \\ C\neq\varnothing}} \text{Sup}_{int}(E \leftrightarrow C|S_i)=0 \end{cases}$$

此模型称为“英->汉”单向求交 n-m 模型,即从 n 个英语单词出发找出它们的对译词.该模型实际上定义了 n-m 对应，即 n 个英语单词与 m 个汉语单词之间的对应(n≥1,m≥0).当 m、n 较大时通常不是最小对应，对词对齐来说意义不大.

类似可定义 “英<-汉”单向求交 n-m 模型，即从 m 个汉语单词出发找出它们的对译词.

定义 11. “英->汉”单向求交 n-m 模型中,当 n=|E|=1 时，称 S_j 最小求交支持 E<->C|S_i＝{e}<->{c_1, c_2, c_3, …, c_m}| S_i.它表示在 S_j 支持下通过最小求交可决定 S_i 中的英语单词 e 与汉语单词 c_1、c_2、c_3、…、c_m 对应.若语料库 BC 中有 x 个句对最小求交支持{e}<->C|S_i，则称 BC 中{e}<->C|S_i 的最小求交支持度为 x,记为 Sup_{minint}({e}<->C|S_i)=x.

在 S_i= ES_i<->CS_i 中， ES_i 中的一个英文单词 e 的对译词集合 C 可由下述公式决定：

$$\{e\} \leftrightarrow C|S_i = \begin{cases} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\text{Arg max}}\, \text{Sup}_{minint}(\{e\} \leftrightarrow C|S_i), & \text{当} \max\limits_{\substack{C\subseteq CS_i \\ C\neq\varnothing}} \text{Sup}_{minint}(\{e\} \leftrightarrow C|S_i)>0 \\ \{e\} \leftrightarrow \varnothing|S_i & ,\text{当} \max\limits_{\substack{C\subseteq CS_i \\ C\neq\varnothing}} \text{Sup}_{minint}(\{e\} \leftrightarrow C|S_i)=0 \end{cases}$$

此模型称为“英->汉”单向最小求交模型，也称最小求交 1-m 模型,可通过最小求交找出 S_i 中一个英语单词的对译词.类似可定义“英<-汉”单向最小求交模型,也称最小求交 n-1 模型， 可通过最小求交找出 S_i 中一个汉语单词的对译词.

例1. 设语料库由以下句对构成.

S_1=he left Beijing<->他 / 离开 / 了 / 北京

S_2=he likes playing football<->他 / 喜欢 / 踢 / 足球

S_3=he will come here<->他 / 将 / 来 / 这

S_4=he is eating lunch<->小王 / 正在 / 吃 / 午饭

S_5=she is eating lunch with her mother<->她 / 正在 / 和 / 母亲 / 一起 / 吃 / 午饭

S_6=I left Beijing<->我 / 离开 / 了 / 北京

S_7=Divoc will come here->迪 / 瓦 / 瓷 / 将 / 来 / 这

S_8= Divoc likes playing football->迪 / 瓦 / 瓷 / 喜欢 / 踢 / 足球

对于 S_1 中的单词“he”,满足最小求交条件的有 S_2、S_3、S_4. $S_1 \cap S_2$={he}->{他}, $S_1 \cap S_3$={he}->{他}, $S_1 \cap S_4$={he}->Φ.所以在该语料库中{he}->{他}|S_1 的最小求交支持度为 2,即 Sup_{minint}({he}->{他}|S_1)=2; {he}->Φ|S_1的最小求交支持度为 1，即 Sup_{minint}({he}->Φ|S_1)=1.按最小求交模型在 S_1 中选择“他”作为“he”的对译词.

3.2 最小求差模型

定义 12. 设 S_i、$S_j \in BC$. S_i= ES_i<->CS_i={e_{i_1}, e_{i_2}, e_{i_3}, …, $e_{i_{n_i}}$}<->{c_{i_1}, c_{i_2}, c_{i_3}, …, $c_{i_{m_i}}$}, S_j= ES_j<->CS_j={e_{j_1}, e_{j_2}, e_{j_3}, …, $e_{j_{n_j}}$}<->{c_{j_1}, c_{j_2}, c_{j_3}, …, $c_{j_{m_j}}$}.若两个句对的差集为 $S_i \backslash S_j$=E<->C={e_1, e_2, e_3, …, e_n}<->{c_1, c_2, c_3, …, c_m}.则称 S_j 求差支持 E<->C|S_i.它表示在 S_j 支持下通过求差可决定 E 中英语单词 e_1、e_2、e_3、…、e_n 与 C 中汉语单词 c_1、c_2、c_3、…、c_m 对应.若语料库 BC 中有 x 个句

对求差支持 E<->C|S_i，则称 BC 中 E<->C|S_i的求差支持度为 x. 记为 Sup_{dif}(E<->C|S_i)=x.

在 S_i= ES_i<->CS_i中,对于任何 E($\Phi \subset E \subseteq ES_i$),E 中英文单词的对译词集合 C 可由下述公式决定：

$$E \leftrightarrow C|S_i = \begin{cases} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\text{Arg max}}\, Sup_{dif}(E \leftrightarrow C|S_i), & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup_{dif}(E \leftrightarrow C|S_i)>0 \\ E \leftrightarrow \varnothing|S_i, & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup_{dif}(E \leftrightarrow C|S_i)=0 \end{cases}$$

此模型称为“英->汉”单向求差 n-m 模型.类似可定义 “英<-汉”单向求差 n-m 模型，

定义 13. “英->汉”单向求差 n-m 模型中,当 n=|E|=1 时，称 S_j 最小求差支持 E<->C|S_i＝{e}<->{c1,c2,c3,…,cm}.若语料库 BC 中有 x 个句对最小求差支持{e}<->C|S_i，则称 BC 中{e}<->C|S_i的最小求差支持度为 x,记为 Sup_{mindif}({e}<->C|S_i)=x.

在 S_i= ES_i<->CS_i中， ES_i 中的一个英文单词 e 的对译词集合 C 可由下述公式决定：

$$\{e\} \leftrightarrow C|S_i = \begin{cases} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\text{Arg max}}\, Sup_{mindif}(\{e\} \leftrightarrow C|S_i), & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup_{mindif}(\{e\} \leftrightarrow C|S_i)>0 \\ \{e\} \leftrightarrow \varnothing|S_i, & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup_{mindif}(\{e\} \leftrightarrow C|S_i)=0 \end{cases}$$

此模型称为“英->汉”单向最小求差模型，也称最小求差 1-m 模型.类似可定义“英<-汉”单向最小求差模型,也称最小求差 n-1 模型.

例 1 中，对于 S_1 中的单词“he”,满足最小求差条件的只有 S_6. S_1\S_6={he}->{他}.所以在该语料库中{he}->{他}|S_1的最小求差支持度为 1,即 Sup_{mindif}({he}->{他}|S_1)=1.按最小求差模型在 S_1中选择“他”作为“he”的对译词.

3.3 混合模型

S_i中的一个英文单词 e 能通过最小求交模型确定其在 S_i中的对译词的必要条件是至少还有一个句子 S_j 中含有 e，但不含 S_i中除 e 外的其它单词. S_i中的一个英文单词 e 能通过最小求差模型确定其在 S_i中的对译词的必要条件是至少还有一个句子 S_j中含有 S_i中除 e 之外的所有单词.因为两个模型适用条件不同，所以可互相弥补不足.

定义 14. {e}<->C|S_i在语料库 BC 中的支持度是它在 BC 中的最小求交支持度与最小求差支持度之和,记为 Sup({e}<->C|S_i),即 Sup({e}<->C|S_i)= Sup_{minint}({e}<->C|S_i)+ Sup_{mindif}({e}<->C|S_i).

在 S_i= ES_i<->CS_i中， ES_i 中的一个英文单词 e 的对译词集合 C 可由下述公式决定：

$$\{e\} \leftrightarrow C|S_i = \begin{cases} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\text{Arg max}}\, Sup(\{e\} \leftrightarrow C|S_i), & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup(\{e\} \leftrightarrow C|S_i)>0 \\ \{e\} \leftrightarrow \varnothing|S_i, & \text{当} \underset{\substack{C\subseteq CS_i \\ C\neq\varnothing}}{\max} Sup(\{e\} \leftrightarrow C|S_i)=0 \end{cases}$$

此模型称为“英->汉”单向最小求差、最小求交混合模型,简称“英->汉”单向混合模型.类似可定义“英<-汉”单向最小求差、最小求交混合模型，简称“英<-汉”单向混合模型.

“英->汉”单向混合模型中，e 与支持度最大的非空集合 C 中的单词对应.只有非空对应的支持度全为 0 时才与Φ对应，即没有对应.该模型倾向于为 e 尽可能找到对应.

{e}<->Φ|S_i在实际语料中通常会有较大的支持度，但不能把它与非空对应同样对待.如例 1 中，仅 S_5 最小求差支持{he}<->{小王}|S_4，所以 Sup({he}<->{小王}|S_4)=1.而 S_1、S_2、S_3最小求交都支持{he}<->Φ|S_4,所以 Sup({he}<->Φ|S_4)=3.但是按照混合模型最终选择{he}<->{小王}|S_4，而不是{he}<->Φ|S_4.

该模型具有兼容未登录词和汉语分词错误的能力.如例 1 的 S_7中“Divoc”和“迪瓦瓷”都是未登录词，并且汉语句子中“迪瓦瓷”分词错误.但 S_8 最小求交支持 Divoc}->{迪,瓦,瓷}|S_7,S_3 最小求差支持

{Divoc}->{迪,瓦,瓷}|S_7,所以 Sup({Divoc}->{迪,瓦,瓷}|S_7)=2，按照混合模型最终选择{Divoc}->{迪,瓦,瓷}|S_7.

3.4 哑词表

为使更多的句对满足最小求交和最小求差的必要条件，我们把部分英语单词如“the”、“a”、“that”、“which”、“of”等规定为哑词，放入英语哑词表 EDumbTable,即在对齐过程中忽略它们的存在，留到夹逼模型中再考虑它们.汉语中的“的”、“地”、“得”、“了”、“吗”、“一”、“个”、“种”、“条”等放入汉语哑词表 CDumbTable 中，它们单独成词时也将被忽略.同时两种语言的标点符号及其它一些符号也归入相应哑词表中.

3.5 双向模型

“英->汉”单向混合模型和“英<-汉”单向混合模型, 可结合成双向模型，使句对中更多的单词得到对齐.设“英->汉”单向混合模型得到的对应集合为 E_CPSET、 “英<-汉”单向混合模型得到的对应集合为 C_EPSET，则双向模型的对应集合为 PSET=E_CPSET∪C_EPSET.

3.6 聚合模型

“英->汉”单向混合模型可以求出 1-m 对应，即一个英文单词对应 m(m≥0)个汉语单词.“英<-汉”单向混合模型可以求出 n-1 对应，即 n(n≥0)个英语单词对应 1 个汉语单词.也就是双向模型可求出 1-m 和 n-1 对应.但不同对应之间交集不一定为空,即可能不满足对齐的完备性.同时对应可能是部分对应，而不是完整对应,即可能不满足对齐的最小完整性.聚合是把交集非空的对应并为一个更大的对应，使各对应的交集为空，且更可能是完整对应.例如，对应{e_1}->{c_1, c_3}|S 和对应{e_4}->{c_3, c_4}交集非空，则聚合为新的对应{e_1, e_4}->{c_1, c_3, c_4}|S.经过聚合以后，实际上产生了 n-m 对应(n≥0,m≥0,m、n 不同时为 0).{e_1, e_4}->{c_1, c_3, c_4}|S 是一个 2-3 对应.聚合模型就是重复这一过程直至任何两个对应的交集为空.

3.7 夹逼模型

定义 15. 对于句对 S 中的单词 w，若存在 S 的非空对应 P=EP<->CP,使 w∈EP 或 w∈CP,则称 w 有对应，否则称无对应.对于句对 S 中的英语单词 e 和汉语单词 c，若存在 S 的对应 P=EP<->CP,使 e∈EP 且 c∈CP,则称 e 与 c 对应，否则称 e 与 c 不对应.

前述模型中得到的对应是一元集合形式的对应 ，没有考虑单词在句子中的顺序，相同的单词被看作一个单词，并且没有对齐哑词.夹逼模型主要考虑单词的顺序，首先把一元集合形式的对应转化为二元集合形式的对应.相同的单词因位置不同而成为不同的元素.哑词与普通单词同样对待.夹逼模型主要解决两件事情:(1)用于解决无对应问题,包括哑词；(2) 用于解决重复单词问题.

设句对 S 的原始形式为 S^0，一元集合形式为 S^1, 二元集合形式为 S^2.则 S^1中的每个元素与 S^2中的一个或多个元素对应.

定义 16. 设 P 是 S^1的对应，把 P 中元素用它对应的 S^2中的元素代替就得到 S^2中的对应.这一过程称为扩展.

例如,S^0=the dog is running after the cat<->那/条/狗/正在/追/那/只/猫，假设有 S^1中的对应{the}<->Φ|S^1 和{dog}<->{狗}|S^1,则扩展后分别成为 S^2 中的对应{<the,1>,<the,6>}<->Φ|S^2 和{<dog,2>}<->{<狗,3>}|S^2.

定义 17. 对于句对 S=ES<->CS={<e_1, ep_1>,<e_2, ep_2>,<e_3, ep_3>,…,<e_n, ep_n>}<->{<c_1, cp_1>,<c_2, cp_2>,<c_3, cp_3>, …,<c_m, cp_m>}，设<e_0, ep_0>和<e_{n+1}, ep_{n+1}>分别代表 ES 的左边界和右边界， <c_0, cp_0>和<c_{m+1}, cp_{m+1}>分别代表 CS 的左边界和右边界，则 S 可表示为 S=ES<->CS={<e_0, ep_0>,<e_1, ep_1>,<e_2, ep_2>,<e_3, ep_3>,…,<e_n, ep_n>,<e_{n+1}, ep_{n+1}>}<->{<c_0, cp_0>,<c_1, cp_1>,<c_2, cp_2>,<c_3, cp_3>,…,<c_m, cp_m>,<c_{m+1}, cp_{m+1}>}.这一表示形式称为 S 的增广二元集合表示形式.

规定 S 的增广二元集合表示形式中<e_0, ep_0>与<c_0, cp_0>对应，<e_{n+1}, ep_{n+1}>与<c_{m+1}, cp_{m+1}>对应.若<e_n, ep_n>与<c_m, cp_m>都无对应，则规定<e_n, ep_n>与<c_m, cp_m>对应，它们通常是英语句子和汉语句子的最后一个标点.

定义 18. 设 S=ES<->CS={<e_0, ep_0>,<e_1, ep_1>,<e_2, ep_2>,<e_3, ep_3>,…,<e_n, ep_n>,<e_{n+1}, ep_{n+1}>}<->{<c_0, cp_0>,<c_1, cp_1>,<c_2, cp_2>,<c_3, cp_3>,…,<c_m, cp_m>,<c_{m+1}, cp_{m+1}>}是 S 的增广二元集合表示形式，则 ES 的子集

ESEG={<e_i, ep_i>, <e_{i+1}, ep_i+1>, <e_{i+2}, ep_i+2>, …, <e_{i+p}, ep_i+p>}(0≤i≤i+p≤n+1)称为 S 的英语片段,它相当于原始形式的单词序列 $e_ie_{i+1}e_{i+2}$, …, e_{i+p}. ESEG 中若 p≥2 且只<e_i, ep_i>和<e_{i+p}, ep_i+p>有对应，其它元素无对应，则称为 S 的英语夹逼片段.

类似可定义汉语片段和汉语夹逼片段.

定义 19. 若 S 的英语夹逼片段 ESEG={<e_i, ep_i>, <e_{i+1}, ep_i+1>, <e_{i+2}, ep_i+2>, …, <e_{i+p}, ep_i+p>}和汉语夹逼片段 CSEG={<c_j, cp_j>, <c_{j+1}, cp_j+1>, <c_{j+2}, cp_j+2>, …, <c_{j+q}, cp_j+q>}中，<e_i, ep_i>与<c_j, cp_j>对应且<e_{i+p}, ep_i+p>与<c_{j+q}, cp_j+q>对应，或<e_i, ep_i>与<c_{j+q}, cp_j+q>对应且<e_{i+p}, ep_i+p>与<c_j, cp_j>对应,则称 ESEG 和 CSEG 满足夹逼条件.并规定 ESEG 和 CSEG 可生成新对应{<e_{i+1}, ep_i+1>, <e_{i+2}, ep_i+2>, …, <e_{i+p-1}, ep_i+p−1>}<->{<c_{j+1}, cp_j+1>, <c_{j+2}, cp_j+2>, …, <c_{j+q-1}, cp_j+q−1>}|S.

通过无对应单词的夹逼处理可使一部分无对应单词找到对应，再通过聚合使新生成的对应交集为空.

夹逼处理还可用于解决重复单词的对应问题.

定义 20. 设二元集合表示形式的对应 P= E<->C 中含有重复单词，即 E 中两个以上元素的第一分量相同或 C 中两个以上元素的第一分量相同.若 P 为非空对应， E 不是英语片段且 C 不是汉语片段,则称 P 满足分裂条件.

若 P 为空对应，则不需处理；若 E 为英语片段或 C 为汉语片段,则 P 也不需处理.重复单词的夹逼处理就是对满足分裂条件的对应 P= E<->C 进行处理.处理方法是把 E 中的单词和 C 中的单词都看作无对应，用解决无对应问题的夹逼模型确定 P 中的各单词的对译词.新得到的各对应需要与 P 求交，再通过聚合使其满足交集为空.

4 算法

4.1 词对齐总体算法

算法 1. 词对齐总体算法.

输入：英汉句对文件.

输出：所有句对及其二元集合形式的对齐.

1. 预处理
2. 建立单词倒排索引表
3. for each 句对 S in 预处理后句对文件
4. “英->汉”单向混合模型词对齐
5. “英<-汉”单向混合模型词对齐
6. 聚合处理
7. 夹逼处理
8. 输出 S 及其二元集合形式的对齐

4.2 预处理

预处理主要完成以下功能：(1)英语单词词形还原；(2)在英语单词和标点之间加空格；(3)对汉语句子进行分词；(4)把经过以上处理的句对存入一个文本文件，称为预处理后句对文件(即前文所述的双语语料库 BC)，每个句对占一行，英语句子和汉语句子间用跳格符分开.

4.3 单词倒排索引表

设句对 S 在预处理后句对文件中的相对于文件头的偏移量是 I(当预处理后句对文件较小时，可把所有句对读入内存数组中，则偏移量 I 是存放 S 的数组元素下标).这样若已知 I 就可直接从预处理后句对文件中读出 S.以后用 I(S)表示句对 S 在预处理后句对文件中的偏移量，而用 S(I)表示偏移量是 I 的句对,同时用 ES(I)和 CS(I)表示句对 S(I)中的英语句子和汉语句子.

单词倒排索引表 InvTab 的每个记录为二元组<w, $ISET_w$>,其中 $ISET_w$是单词 w出现的句对的偏移量的集合.如<many,{345,567,678}>,表示“many”在 S(345)、S(567)和 S(678)中出现.单词倒排索引表按单词以散列方式组织，能够实现快速建立和查找.单词倒排索引表分为英语单词倒排索引表 EInvTab 和汉语单词

倒排索引表 CInvTab,用于存放语料库中除哑词外其它单词的倒排索引.

4.4 单向词对齐

单向词对齐算法包括“英->汉”单向和“英<-汉”单向最小求交、最小求差混合模型实现算法.因为这两个算法基本一致，所以本文仅讨论“英->汉” 单向的混合模型实现算法.

混合模型算法中将用到候选对应集合(candidate parallel set,CPS),CPS 中的记录是二元组＜parallel,support>,分别表示单词对应关系及其支持度.

算法 2. “英->汉”单向混合模型实现算法.

输入：句对 S、预处理后句对文件、英语单词倒排索引表、哑词表.

输出：S 的“英->汉”单向对应集合 E_CPSET.

1. 求出 S 的一元集合形式 S=ES<->CS
2. for each e_k in ES\EDumbTab
3. find ISET_{e_k} in EInvTab
4. for each e_k in ES\EDumbTab
5. 把 CPS 置空
6. “英->汉”单向最小求交词对齐
7. “英->汉”单向最小求差词对齐
8. if CPS 内有非空对应
9. then 输出支持度最大的非空对应到 E_CPSET 中
10. else 输出空对应到 E_CPSET 中

“英->汉”单向最小求交词对齐可细化为:

1. for each 句对偏移量值 j in $\mathrm{ISET}_{e_k} \setminus \bigcup_{\substack{e_p \in ES\setminus EDumbTab \\ p\neq k}} \mathrm{ISET}_{e_p}$
2. find S(j) in 预处理后句对文件
3. 求出 S(j)的一元集合形式 S(j)=ES(j)<->CS(j)
4. 生成对应 P={e_k}->CS∩CS(j)\ CdumbTab
5. if P 在 CPS 中
6. then P 的支持度加 1
7. else 把 P 加入 CPS 中且支持度置为 1

其中句对偏移量集合 $\mathrm{ISET}_{e_k} \setminus \bigcup_{\substack{e_p \in ES\setminus EDumbTab \\ p\neq k}} \mathrm{ISET}_{e_p}$ 是最小求交算法实现的关键. ISET_{e_k} 中存放的是含 e_k 的句对的偏移量集合. $\bigcup_{\substack{e_p \in ES\setminus EDumbTab \\ p\neq k}} \mathrm{ISET}_{e_p}$ 存放的是至少含有 ES 中除 e_k 和哑词外的一个单词的句对的偏移量集合.所以 $\mathrm{ISET}_{e_k} \setminus \bigcup_{\substack{e_p \in ES\setminus EDumbTab \\ p\neq k}} \mathrm{ISET}_{e_p}$ 是含 e_k 而不含 ES 中除 e_k 和哑词外其它单词的句对的偏移量集合.因此通过 $\mathrm{ISET}_{e_k} \setminus \bigcup_{\substack{e_p \in ES\setminus EDumbTab \\ p\neq k}} \mathrm{ISET}_{e_p}$ 中的元素（即句对的偏移量）找到的句对在不考虑哑词的条件

下与 S 满足最小求交条件.

“英->汉”单向最小求差词对齐的实现与“英->汉”单向最小求交词对齐的实现类似. 聚合处理与夹逼处理实现比较简单，因篇幅所限，这里不再详述.

5 实验和相关工作对比讨论

5.1 实验语料

英汉词对齐实验是以原始表示形式存放的双语句对库为输入的.实验中所有语料共含 244599 英汉句对，成份如表 1 所示.

表 1　双语句对库

类别	规模（句对）	英语词形数	英语词数	汉语词形数	汉语词数	英语句子平均词数	汉语句子平均词数
法律	28325	22327	808307	11275	592509	28.5	20.9
汽车	47564	13287	662242	12728	696135	13.9	14.6
词典	63249	28572	617229	31267	643696	9.8	10.2
日常	62378	16728	620527	27584	544889	9.9	8.7
口语	30267	8735	250622	13907	219657	8.3	7.3
经贸	12816	6273	144824	9721	124359	11.3	9.7

实验主要检测词对齐正确率和召回率两个指标：

$$\text{正确率} = \frac{\text{系统输出的正确的最小完整对应总数}}{\text{系统输出的最小完整对应总数}} \times 100\%$$

$$\text{召回率} = \frac{\text{系统输出的正确的最小完整对应总数}}{\text{语料库中实有的最小完整对应总数}} \times 100\%$$

因为无法做到对所有结果进行核查，实际是从每组中各抽出 100 个句对作为测试语料，手工检查对齐的正确率和召回率，用以估算整组的对齐效果.表 2 中说明了每组测试语料中所含的最小完整对应的形数和总数.

表 2 测试语料

组别	类别	规模（句对）	最小完整对应形数	最小完整对应总数
TestSet1	法律	100	1831	2656
TestSet 2	汽车	100	967	1638
TestSet 3	词典	100	792	1115
TestSet 4	日常	100	463	872
TestSet 5	口语	100	398	710
TestSet 6	经贸	100	813	1451

实验进行的是开放测试,用于建立索引的训练语料不含以上测试语料。

5.2 实验结果

共进行了三个实验.

第一个实验使用全部训练语料，共 243999 句对.目的是研究在较大规模的异质语料支持下不同组别语

料的对齐效果.实验结果如表 3 所示.

第二个实验只使用与各测试集同类别的语料作为训练语料.目的是研究较小规模的同质语料支持下各组的对齐效果.实验结果如表 4 所示.

第三个实验是针对汽车语料进行的.目的是研究同质训练语料规模不同时对对齐结果的影响.实验结果如表 5 所示.

表 3　使用全部训练语料时词对齐正确率和召回率

组别	训练语料	正确率(%)	召回率(%)
TestSet 1	全部	78.3	80.4
TestSet 2	全部	90.7	91.2
TestSet 3	全部	80.6	82.1
TestSet 4	全部	84.3	85.8
TestSet 5	全部	58.9	62.3
TestSet6	全部	63.5	67.2

表 4　只使用同质训练语料时的词对齐正确率和召回率

组别	训练语料（句对）	正确率(%)	召回率(%)
TestSet 1	同质 28225	72.1	74.5
TestSet 2	同质 47464	86.6	88.3
TestSet 3	同质 63149	76.3	78.2
TestSet 4	同质 62278	80.2	83.3
TestSet 5	同质 30167	47.3	52.5
TestSet6	同质 12716	28.7	37.4

表 5 同质训练语料规模对词对齐的影响

组别	训练语料（句对）	正确率(%)	召回率(%)
TestSet 2	同质 10000	47.2	56.7
TestSet 2	同质 20000	63.5	69.3
TestSet 2	同质 30000	73.6	77.8
TestSet 2	同质 40000	82.2	84.3
TestSet 2	同质 47464	86.6	88.3

以上实验结果说明：

(1)语料本身的性质对词对齐结果影响极大.在相同的较大规模异质语料的支持下，翻译最规范的“汽车维修手册”对齐效果最好，翻译最灵活的“口语”语料对齐效果最差.“口语”语料有许多类似“This is Mayamm on the phone.<->我是玛亚姆.”这样的句对，给对齐造成了一定困难.

(2) 训练语料的规模对词对齐结果影响极大.同质训练语料的规模越大，效果越好.对于“汽车维修手册”，当本组的训练语料从 1 万句对增到约 5 万句对时，正确率由 47.2%增到 86.6%，召回率由 56.7%增加到 88.3%.

(3)同质训练语料规模不够大时，异质训练语料的支持也起一定作用.“经贸” 训练语料只有 12716 句对，仅用同质训练语料时正确率和召回率只有 28.7%和 37.4%，用全部 243999 句对训练语料时，正确率和召回率分别达到 63.5％和 67.2%.

5.3 相关工作对比讨论

因为词对齐对于双语知识获取、双语词典编纂、机器翻译等具有重要意义，近年来得到广泛的研究.但英汉词对齐的研究却相对滞后.当前英汉词对齐中主要使用双语词典、同义词词典等资源，同时辅以统计共现信息.很少有不用双语词典的.

刘小虎是国内较早进行英汉词对齐研究的[14].他主要采用了双语词典完全匹配策略、同义词词典完全匹配策略和共现互信息、t-score.刘小虎对 6 万英汉句对中的实词进行了对齐研究，但没有说明语料所属的领域.得到了 79.5%的对齐正确率.

吕雅娟的英汉词对齐研究中考虑了双语词典完全匹配、双语词典模糊匹配、语义相似匹配、统计词性匹配、共现统计补充词表以及位置因素[15].她的语义相似匹配也可理解为结合同义词词典的完全匹配，但她用语义分类码的层级体制更细地刻划了同义词的“同义”程度.她对初中、高中和大学英语课本中的 3 万句对进行了对齐研究，包含空对齐的词对齐正确率为 80.87，召回率为 78.75%.

在以上两人的研究中，统计方法只是整体方法的次要部分，是对双语词典方法和同义词词典方法的辅助和补充.相比他们的研究，本文的方法具有以下特点：

(1)语料库方法可独立应用.基于语料库的无双语词典的词对齐方法的提出对于大规模语料库的词对齐研究具有重大意义，除用于英语词形还原的英语词表和用于汉语分词的汉语词表外它不需要任何语言学知识和语言学资源，对机器可读知识匮乏的特殊语种、特定领域也适用.它不需要双语词典，反过来可用来编纂双语词典.实验表明，当同质训练语料规模较大时本文方法能取得比较好的词对齐结果.当同质训练语料规模不大时，异质训练语料的支持也起一定作用.

(2)仅就语料库方法而言，本文的最小求交、最小求差方法克服了共现方法理论上的不足.共现方法都要设置频次阈值，这使得源文与译文的低频共现不能被处理，这也是共现方法不能独立应用的原因.最小求交、最小求差方法不仅能处理高频共现，而且理论上在大规模语料库支持下能处理低频共现.最小求交方法有效的必要条件是源文与译文共现两次，最小求差方法有效的必要条件是源文与译文共现一次.在大规模语料库之下最小求交、最小求差的其它条件也可被满足，克服了共现方法不能处理低频共现这一模型本身的不足.

(3)在大规模语料库下，最小求交、最小求差方法的潜力大于双语词典方法与共现方法的结合.由于自然语言的复杂性和翻译的灵活性，双语词典的译文对大规模语料库的覆盖很低，对于未登录词更是无能为力.而共现方法不能处理低频共现.这样大规模语料中大量存在的源文与译文低频共现、源文是未登录词或译文不被源文的词典译项覆盖现象，即使双语词典方法与共现方法结合，也是处理上的一个盲点.而在大规模语料库支持最小求交、最小求差方法有潜力触及这个盲点.

(4)最小求交、最小求差方法对汉语分词错误有一定兼容能力，一些切碎的词可被正确对齐.

(5)本文词对齐方法考虑了部分对应问题.通过聚合处理把部分对应聚为完整对应.

(6)本文词对齐方法考虑了一个句对中一词多现和一词的多个译项同现的问题.通过分裂处理和夹逼处理进行解决.

6 结论

通过以集合形式描述句子，可用集合理论来研究自然语言.本文提出的基于语料库的无双语词典的词对齐模型就是建立在句对的集合形式基础上.通过句对间最小求交、最小求差运算发现双语单词间的 1-m 和 n-1 对应，再通过聚合处理发现 n-m 对应.哑词表的使用排除了一些虚词的干扰，夹逼处理考虑了词序和重复词对于对齐的影响.单词倒排索引表的使用和集合理论的应用为模型的实现提供了高效的算法.本文方法的最大特点是几乎不需要任何语言学知识和语言学资源.实验表明，当同质语料规模较大时本文方法能取得比较好的词对齐结果.当同质语料规模较小时，异质语料的支持也起一定作用.

参 考 文 献

1 Xu Dong-Hua. Aligning and matching of English-Chinese bilingual texts of CNS news. Department of Information System and Computer Science, National Univerisity of Singapore,Technical Report: cmp-lg/9608017, 1996

2 Wang Bin,Liu Qun,Zhang Xiang. Automatic Chinese-English paragraph segmentation and alignment. Joural of Software,2000,11(11):1547-1553(in Chinese)
(王斌,刘群,张祥. 汉英双语库自动分段对齐研究.软件学报，2000,11(11):1547-1553)
3 Brown P. F., Lai J. C., Mercer R. L. *et al*. Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991, 169-176
4 Gale W. A., Church K. W.. A program for aligning sentences in bilingual corpora. Computational Linguistics, 1993,19(1): 75-102
5 Kay M., Roscheisen M.. Text-translation alignment.Computational linguistics, 1993, 19(1): 121-142
6 Chen S. F.. Aligning sentences in bilingual corpora using lexical information. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, OH, 1993, 9-16
7 Wu De-Kai. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In: Proceedings of the 32th Annual Conference of the Association for Computational Linguistics, Las Cruces, NM, 1994,80-87
8 Qian Li-ping,Zhao Tie-jun,Yang Mu-yun *et al*. Translation-based automatic alignment of Eglish and Chinese parallel corpora. Mini-micro System,2001,22(1):123-125(in Chinese)
(钱丽萍,赵铁军,杨沫昀等. 基于译文的英汉双语句子自动对齐. 小型微型计算机系统,2001,22(1):123-125)
9 Imamura K.. A hierarchical phrase alignment from English and Japanese bilingual text. In: Procceedings of the 2nd International Conference on Intelligent Text Processing and Computational Linguistics, Mexico, 2001, 206-207
10 Ker S. J.,Chang J. S.. A class-based approach to word alignment. Computational linguistics, 1997,23(2): 313-344
11 Borin L.. You'll take the high road and I'll take the low road: using a third language to improve bilingual word alignment. In: Procceedings of the 18th International Conference of Computational Linguistics, Saarbrucken, Germany,2000, 97-103
12 Dagan I., Church K. W., Gale W. A.. Robust bilingual word alignment for machine aided translation. In: Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Columbus,OH, 1993,1-8
13 Sun Le,Jin You-Bing, Du Lin *et al*. Word alignment of English-Chinese bilingual corpus based on chunks. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong, 2000,110-116
14 Liu Xiao-hu,Wu Wei,Li Sheng *et al*. Aligning algorithm for a corpus at word level based on dictionary and statistics. Journal of informatics,1997,16(1):20-26(in Chinese)
(刘小虎,吴葳,李生等. 基于词典和统计的语料库词汇级对齐算法. 情报学报,1997,16(1):20-26)
15 Lü Ya-juan, Zhao Tie-jun,Li Sheng *et al*. Word alignment based on statistic and lexicon. In: Hunang Chang-jing,Zhang Pu. Natural Language Understanding and Machine Translation. Beijing:Tsinghua University Press,2001,108-115(in Chinese)
(吕雅娟，赵铁军，李生，等.统计和词典方法相结合的双语语料库词对齐. 见：黄昌宁,张普. 自然语言理解与机器翻译. 北京：清华大学出版社,2001, 108-115)
16 Cicekli I.，Güvenir H. A.. Learning translation templates from bilingual translation examples. Applied Intelligence,2001,15(1): 57-76
17 Güvenir H. A., Cicekli I.. Learning translation templates from examples. Information systems, 1998, 23(6): 353-363
18 Oz Z.,Cicekli I.. Ordering translation templates by assigning confidence factors. In: Proceedings of the 3rd Conference of Association for Machine Translation in the Americas, Langhorne, PA, 1998,51-61
19 McTait K.,Trujillo A.. A language-neutral sparse-data algorithm for extracting translation patterns.In: Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation, Chester,England,1999,98-108

Background

The National Natural Science Foundation project "Chunk Based Machine Translation System" aims to apply bilingual chunk-base and partial parsing technology in example based machine translation. The 973 program "Evaluation of Machine Translation" aims to research the evaluation methods of machine translation and develop the automatically evaluating tools. The 863 program "Constructing Linguistic Knowledge-base and researching its application", "Chinese-English and Chinese-Japanese corpora" and "The Construction and Utilization of A Comprehensive Language Knowledge-base" aim to construct machine readable linguistic resource for natural language processing.

The research group has presented several machine translation theories, such as Lexical Semantic Drive Theory, Extend Chunk Theory, etc. They also developed Chinese-English-Korean multilingual machine translation system. With the cooperation of Institute of Computational Linguistics, Peking University and Natural Language Processing Laboratory, Northeastern University, the first word alignment Chinese-English corpus has been constructed.

The word alignment technology presented in this paper has been applied in: (1) Generating translation template; (2) Generating target language; (3) Constructing bilingual chunk-base; (4) Constructing word alignment bilingual corpus; (5) Compiling bilingual dictionary.