基于转换和映射的语义单元自动获取①

方 森② 曹井香

(大连理工大学计算机科学与工程系 大连 116024)

摘 要 提出了一种基于转换和映射的英汉双语语料的语义单元自动获取方法,即构造一套规则系统对英语句子的链语法分析结果进行处理,以设定单词的语义层次并转换生成语义单元的英语表示(ESER);然后利用统计双语词对齐将之映射为语义单元的汉语表示(CSER),从而获得双语语义单元;最后通过语义单元表示实量竞争、合并以及召回等一系列策略对自动获取结果进行优化。该方法能够在缺少完全语义分析的情况下实现语义单元的自动获取。实验结果表明双语语义单元自动获取的 F 值达到了 74.06%,基于语义单元的机器翻译系统具有准确率高的特点。

关键词 语义单元,自动获取,链语法,词语对齐

0 引 言

大规模语料库的建立,为语言知识的自动获取研究提供了很好的基础。由双语文本组成的双语语料库包含了两种语言的对照翻译信息,因而具有更高的价值。语义单元[1-3]是一个表达完整意义的单元。语义单元包括带变量的语义单元和不带变量的语义单元,通过变量代换可以构成新的语义单元。语义单元在不同语言中的表示称为语义单元表示。基于语义单元的机器翻译可以看作语义单元在不同语言中的表示之间的转换,具有快速、准确的特点[2,3]。

人工获取语义单元的缺点是效率低,代价巨大,而且一致性难以保证。半自动获取[4]需要在已有的语义单元的基础上统计推断新的语义单元,并且需要语言专家的交互式参与。双语的结构对齐是自动获取翻译知识的一个常用手段[5-12]。然而,句法结构、平行语法结构和依存结构难以确定语义单元的边界和层次,不能直接用于语义单元的获取。自动获取语义单元的关键在于两点:一、单语或双语语义结构的自动获取;二、准确建立语义单元在两种语言中表示的对应关系。一个句子的完全语义结构分析须借助包含该句子句义所涉及的语义单元的语义单元库[2,3],这个库又需要通过自动获取建立。

链语法[13]是根据单词的连接要求来分析句子和判断句子合法性的文法,它将句子结构表示成满足一定条件的词汇之间的连接关系。这比短语和词性标记信息更丰富,便于语义信息的获取。本文以英汉双语为研究对象,进行了基于转换和映射的语义单元自动获取的研究,得出了提高自动获取准确率的方法,即先根据链语法和语义单元人工提取经验制定一套将英语句子转换成语义单元的英语表示的规则,再根据统计词语对齐结果将语义单元的英语表示映射到汉语表示,最后使用一套优化策略提高语义单元获取的准确率。

1 背景知识

1.1 语义单元理论

1.1.1 基本概念

定义 1(语义单元) 语义单元[2,3]是表达一个完整语义的单元。

例如,"茶杯"是一个实体,是语义单元。"$<N_{人}>$把$<N>$倒入$<N_{容器}>$中"表示一个动作,是语义单元。"$<N>$是$<N>$"表示一种关系,是语义单元。语义单元由实量和虚量组成。其中"把"是实量,"$<N_{人}>$"是虚量,"$N_{人}$"是表示人的语义单元类型。虚量可被满足类型要求的其它语义单元所替代。

① 863 计划(2006AA01Z140)资助项目。
② 男,1980 年生,博士生;研究方向:自然语言处理与机器翻译;联系人,E-mail: fangmiao@sohu.com
(收稿日期:2007-11-07)

定义 2(语义单元表示)　语义单元表示是语义单元在具体自然语言中的反映。

上面的语义单元在英语中的表示分别为“*cup*”,“< $N_{人}$ > *put* < *N* > *into* < $N_{容器}$ >”和“< *N* > be * < *N* >”。同样语义单元表示也由实量和虚量组成。

定义 3(基本语义单元与基本语义单元表示)　不能由其他语义单元代入产生的语义单元被称为基本语义单元;不能由其他语义单元表示代入产生的语义单元表示被称为基本语义单元表示。

定义 4(可弃语义单元与可弃语义单元表示)　可以由其他语义单元代入产生的语义单元被称为可弃语义单元;可以由其他语义单元表示代入产生的语义单元表示称为可弃语义单元表示。

例如,句子“*Mr. Smith is an engineer*”和“史密斯先生是一位工程师”的句义是一个语义单元,可以表示为: $I_S(M_r(S_{mith}), A_n(E_{ngineer}))$,还可以记为:1(2(3),4(5))。其所有的基本语义单元及表示如表 1所示。

表 1　语义单元示例表

语义单元	参数数目类型	汉语表示	英语表示	类型
1($N_{人}$,$N_{人}$)	2, $N_{人}$,$N_{人}$	X_1 是 X_2	X_1 be * X_2	*J*
2($N_{姓}$)	1, $N_{姓}$	*X* 先生	*Mr. X*	$N_{人}$
3	0	史密斯	*Smith*	$N_{姓}$
4($N_{职称}$)	1, $N_{职称}$	一位 *X*	an *X*	$N_{人}$
5	0	工程师	*engineer*	$N_{职称}$

1.1.2　基于语义单元的机器翻译

自然语言的翻译可以通过两步来实现:第 1 步,通过在源语言上的语义分析将源语言的句子 *S* 变为句义表达式。第 2 步,把句义表达式在目的语言上的代入展开成为目标语言的句子 *T*。如输入英语句子“*Mr. Smith is an engineer*”,经语义分析产生句义表达式“$I_S(M_{r.}(S_{mith}), A_n(E_{ngineer}))$”,或记为“1(2(3),4(5))”;展开成汉语句子:1(2(3),4(5))⇒1(2(史密斯),4(工程师))⇒1(史密斯先生,一位工程师)⇒史密斯先生是一位工程师。

1.2　链语法

1.2.1　链语法及其特性

链语法[13]是一种分析英语句子的计算机可读文法,其表达能力等同于上下文无关文法,由卡耐基梅隆大学计算机学院 Sleator 等人开发。链语法具有如下特性:①平面性,画在句子上面的链不能相互交叉;②连接性,句子中所有单词必须相互直接或间接地连接在一起;③顺序性,当单词 $w_1, w_2, \cdots, w_m$ 均连接到 *w* 时, $w_1, w_2, \cdots, w_m$ 在句子中的顺序由 *w* 的连接因子的顺序决定;④排他性,一对单词之间不能同时有两个连接。

1.2.2　连接因子

连接因子是描述单词连接要求的字符串,一个连接因子由一个或者多个大写字母开头,紧跟着若干个(或零个)小写字母,最后是后缀“+”或者“-”。如 S_s+, $B_{sw}-$。“+”表示需要一个单词在右边与它相连;“-”表示需要一个单词在左边与它相连。

一般一个连接因子和另一个字符串相同但方向相反的连接因子相互连接。下标也可以控制连接因子是否可以相互连接。下标是紧跟在连接因子名称后面的小写字母。如“*Ss* +”能够和没有下标的连接因子“*S* -”连接,或者有下标的连接因子“*Ss* -”连接,但是它不能和“*Sp* -”连接。连接因子还可以有多个下标,如“*Spa* +”。“*”是一个通配符,可以和任何类型的下标匹配。如“*S* * *a* +”,能和“*S* -”,“*Ss* -”,“*Sp* -”,“*Ssa* -”等连接,但不能与“*Ssb* -”连接。

1.2.3　链语法分析结果

图 1 是一个英语句子的链语法分析结果,单词之间通过链连接起来。该图等价于如下的链序列:“1:2:*Ds* 2:3:*Ss* 3:4:*MVp* 4:7:*Jp* 5:7:*Dsu* 6:7:*AN*”。链序列由一系列链节点组成,每个节点分别表示链起点位置、终点位置和连接因子,它们之间由冒号隔开。如“1:2:*Ds*”表示一个从单词节点 1 到单词节点 2 的链“*Ds*”。

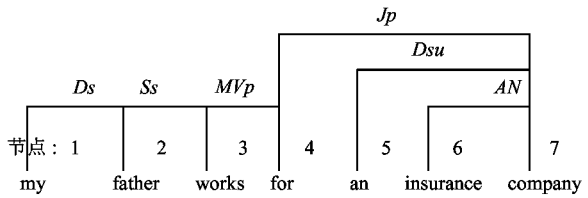

图 1　句子“*my father works for an insurance company*”的链

2　语义单元自动获取方法

2.1　句子链到语义单元表示的转换

在链语法中,每个链代表特定的意义且链的种类有限,这保证了英语句子链向语义单元表示转换切实可行。转换分两步:第一步根据链和转换规则为英语句子中每个节点(单词)设定语义层次;第二

步根据节点的层次和链序列产生语义单元的英语表示。

2.1.1 转换规则

对一个可弃语义单元剖析,可以看出其构成层次关系。如句义“$I_S(M_r(S_{mith}), A_n(E_{ngineer}))$”有 3 个层次。相应地,语义单元表示实量也有层次。如“(*Mr.* (*Smith*)) *is* (*an* (*engineer*))”中“*is*”比“*Mr.*”高一个层次,而“Mr.”比“Smith”高一个层次。

语义单元表示的层次与链语法中的链有关。如“*S*”链连接主语名词和限定性动词,左边的词语一般作为以右边词语为实量的语义单元的参数。可以认为“*S*”链左边节点比右边节点的层次低。类似地,“*O*”链连接动词和直接或间接宾语,左边节点的层次比右边的高。“*K*”链连接动词和小品词,左右两个节点层次相同。

本文根据链表示的意义和人工语义单元提取经验,针对 129 种不同的链,总结了 141 条转换规则。该套规则能够覆盖链语法的所有链。经过对全部实验语料测试(81863 个英语句子),链语法正确分析处理的句子达 74.91%,其中完全正确分析的句子达 68.92%,部分正确分析的句子占 5.99%。从链语法完全正确分析的结果中随机抽取 235 句作为测试集,实验表明转换的正确率达 88.53%,F 值达 91.97%。规则如“*if link-name* = '*S*' *and priority* = 2 *then level-change* = 1”表示链名称等于“*S*”,右侧节点比左侧节点高 1 个层次,规则的优先级为 2。

2.1.2 转换算法

首先根据链和规则调整各词节点的语义层次和链节点;然后针对每个词节点,搜集与之有链连接的词节点,将与当前节点层次相同的节点置入实量集合(*rArr*)中,比当前节点层次低的节点置入虚量集合(*vArr*)中,对所有实量集合中节点重复上述操作,直至集合中的节点被处理完。转换算法如下:

算法 1 转换算法

输入:英语句子 *S*(长度为 *n*),链 *L*

输出:英语语义单元表示集{*ESER*}

① Set _ and _ Adjust(S, L);

//根据规则设定 *S* 中各节点层次(*level*)和调整链 *L*

② For i = 1 to n

If $node_i \notin fArr$ then //*fArr* 为处理过的节点集合

$node_i \rightarrow rArr$; //*rArr* 为 *ESER* 实量集合

Else

Continue;

For each *node* in *rArr*

For each $nodeT \leftrightarrow node \in L$

If level(*nodeT*) = level(*node*)&$nodeT \notin rArr$ then

// level(*node*)为节点 node 的层次

nodeT→rArr;

If level(*nodeT*) < level(*node*)&$nodeT \notin vArr$ then

$nodeT \rightarrow vArr$; //vArr 为 *ESER* 虚量集合

Sort *rArr* and *vArr* respectively;

$rArr \rightarrow fArr$;

Output an *ESER* according to *rArr* and *vArr*。

2.1.3 转换实例

图 1 的句子“*my father works for an insurance company*”经过链语法分析后产生链序列:“1:2:*Ds* 2:3:*Ss* 3:4:*MVp* 4:7:*Jp* 5:7:*Dsu* 6:7:*AN*”。根据转换规则,得到层次序列“1:1 2:0 3:1 4:1 7:0 5:1 6:0”,也可以表示成图 2。层次序列是一系列的节点,每个节点由单词节点序号和层次组成,用冒号隔开。第一节点的层次默认为 1,后面的节点根据相对层次关系得到其层次。然后利用转换算法可以得到 5 个英语基本语义单元表示:“*my*(X_1)”,“*father*”,“(X_1) *works for* (X_2)”,“*an* (X_1)”和“*insurance company*”。

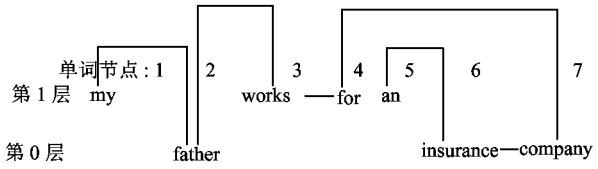

图 2 句子“*my father works for an insurance company*”经转换后的层次图

2.2 映射算法

实量映射比较简单,只需根据词对齐结果将英语语义单元表示的实量节点所对应的汉语的节点找到并加入到对应的汉语实量集合中。虚量映射较复杂,要求英汉表示的虚量之间一一对应。对于一个汉语节点同时对应英语表示的实量和虚量节点的情况,借助双向翻译概率来确定当前汉语节点对应实量还是虚量节点。对于一个英语虚量节点对应多个汉语节点的情况,选取一个最大翻译概率的汉语节点。映射算法如下:

算法 2 映射算法

输入:汉语句子 *SC*,英语句子 *SE*,词对齐 *WA*,语义单元英语表示 *ESER*(实量集 *erArr*,虚量集 *evArr*)

输出:语义单元汉语表示 *CSER*(实量集 *crArr*,虚量

集 *cvArr*)

① For each *nodeC*

If ∃ *nodeE*(*nodeE* ∈ *erArr*&*nodeE*↔*nodeC* ∈ *WA*) then

nodeC→crArr;

② For each nodeE′∈evArr

For each *nodeC*↔*nodeE′* ∈ *WA*

nodeC → *tempArr*;//*tempArr* 为临时节点集合

For each *nodeC* ∈ *tempArr*

If *nodeC* ∈ *crArr* then

Choose *nodeE* from *erArr*:

$\underset{nodeE}{argmax}\, P(nodeC \mid nodeE) + P(nodeE \mid nodeC)$

If $P(nodeC \mid nodeE) + P(nodeE \mid nodeC) \geqslant P(nodeC \mid nodeE') + P(nodeE' \mid nodeC')$ then

Remove *nodeC* from *tempArr*;

Remove *nodeE′*↔*nodeC* from *WA*;

Else

Remove *nodeC* from *crArr*;

Remove *nodeE*↔*nodeC* from *WA*;

If more than one node in *tempArr* then

Choose *nodeC* → *cvArr*:

$\underset{nodeC}{argmax} P(nodeC \mid nodeE) + P(nodeE \mid nodeC)$;

Else if only one node in *tempArr* then

nodeC → *cvArr*;

③ Sort *crArr* and *cvArr* respectively;

④ Output *CSER* according to *crArr* and *cvArr*。

2.3　优化策略

在转换和映射后,语义单元的汉英表示均已获得,但有一些错误。如“*He reversed his car into the garage*. ↔他 把 车子 倒 开进 车库 。”的词语对齐序列为“1:1 2:2 2:3 2:4 2:5 3:5 7:6 8:7”,链序列为“1:2:*Ss* 2:5:*MVp* 2:4:*Os* 3:4:*Ds* 5:7:*Js* 6:7:*Ds*”。通过转换和映射,得到英汉语义单元表示① *He* ↔他,②(*He*)*reversed*(*car*) *into*(*garage*)↔(他) 把 车子 倒 开进 (车库),③ *his car* ↔开进和④ *the garage* ↔车库。其中②和③是错误的,原因在于词语对齐序列中有两个错误“2:3 3:5”,同时遗漏了“4:3 5:5”。

本文采用如下策略对自动获取结果进行优化,消除词语对齐错误和英语表示构造错误的影响,提高准确率:

(1) 对齐补充策略

依据双语词典,将句对中双语互译概率大于一定阈值的英汉词语的节点加入词语对齐中。目的是依据双语词典补充遗漏的双语对齐。

(2) 实量竞争策略

如果两个或多个语义单元英语表示(*ESER*)的实量集合(*erArr*)同时对应一个汉语单词 *nodeC*,则保留一个具有最大概率的 $ESER_{\max}$,以式

$$ESER_{\max} = \underset{ESER_i}{\operatorname{argmax}} P(ESER_i \mid nodeC) + P(nodeC \mid ESER_i) + P_{Cross}(nodeC, ESER_i) \quad (1)$$

表示。删除其余的 $ESER_i$ ($i \neq \max$)到 *nodeC* 的映射,并且去掉词语 *nodeE*($\in ESER_i$)与 *nodeC* 的对齐。(1)式中,$P(ESER_i \mid nodeC)$ 和 $P(nodeC \mid ESER_i)$ 是双向翻译概率,$P_{Cross}(nodeC, ESER_i)$ 是交叉概率。计算公式如下:

$$P(ESER_i \mid nodeC) = \max_{nodeE \in R\{ESER_i\}} P(nodeE \mid nodeC)$$

$$P(nodeC \mid ESER_i) = \max_{nodeE \in R\{ESER_i\}} P(nodeC \mid nodeE)$$

$$P_{Cross}(nodeC, ESER_i) = \max_{nodeE \in R\{ESER_i\}} P_{Cross}(nodeC, nodeE)$$

$$P_{Cross}(nodeC, nodeE) = 1 - \frac{cross\_nodeCE}{total\_cross\_link}$$

“R{ $ESER_i$ }”表示由 $ESER_i$ 的实量所组成的集合。*cross_nodeCE* 是词语对齐 *nodeC*↔*nodeE* 所引起的交叉数目,*total_cross_link* 是当前对齐序列中所有的交叉数。引入交叉概率的目的是解决下面一类问题:*He sees that he has made a mistake*. ↔他 明白 他 犯 了 一个 错误 。在词语对齐中可能出现 1↔1,1↔3 和 4↔1,4↔3 对应,因为互译概率是相等的,不能区分。引入交叉概率后得到对应 1↔1 和 4↔3。

(3) 实量召回策略

实量召回有两种情况:①在语义单元表示的映射操作完成之后,如果汉语的实量表示不连续,并且中间间隔词语均未对应到其他 *ESER*,则将其加入到汉语的实量集合中;②实量未对应的其余情况,依据概率召回未对齐的词语,为每一个未对齐的词语 *nodeC*,选择一个 $ESER_{\max} = \underset{ESER_i}{\operatorname{argmax}} P(ESER_i \mid nodeC) + P(nodeC \mid ESER_i)$ 加上 $ESER_i$ 到 *nodeC* 的对应,同时在词语对齐中加上 $ESER_i$ 中的词语与 *nodeC* 的对齐。

(4) 虚量召回策略

词语对齐错误会造成英语表示中虚量的对应缺失。设英语语义单元表示 $ESER_i = X_1, E_1, X_2,$

…, E_m, X_n(E_i 表示实量, X_i 表示虚量)。在映射完成之后,汉语表示中 X_i 缺失。在英语语义单元表示中 $ESER_j$ 包含 X_i 对应的单词 E_j, 则在 $ESER_j$ 中其余实量表示 E_i 的对应中寻找一个 $nodeC = \underset{nodeC}{\operatorname{argmax}} P(ESER_i \mid nodeC) + P(nodeC \mid ESER_i)$ 作为该参数 X_i 的对应。

(5) 合并策略

不可避免地,仍然有些英语表示会对应到空。为此,合并不同的语义单元表示层次。设英语语义单元表示 $ESER_i = X_1, E_1, X_2, \cdots, E_m, X_n$。在映射完成之后,汉语表示中 X_i 缺失。在英语语义单元表示中 $ESER_j$ 包含 X_i 对应的单词 E_j。$ESER_j$ 的汉语对应为空,则将 $ESER_j$ 代入 $ESER_i$ 中,替换掉变量 X_i。

3 实验及结果分析

3.1 实验设计

为了检验语义单元自动获取的有效性和基于语义单元机器翻译系统的性能,本文进行了两项实验:①语义单元自动获取实验;②基于语义单元的机器翻译测试。

3.1.1 实验基础

(1) 实验语料:抽取厦门大学英汉双语平行语料库[14-19]中英语句子长度在 5 ~ 30 个单词的双语句对,共 81863 对,用于词语对齐和双语词典抽取。

(2) 预处理:

①汉语词法分析:采用中科院计算所词法分析系统 ICTCLAS[14]对汉语句子切词;

②英语词法分析:采用开源工具 tokenizeE.perl.tmpl[15]对英语句子断词;

③链语法分析:采用开源工具 linkparser4.0[16]对英语句子进行语法分析;

④双语词语对齐:采用开源工具 GIZA ++[17]进行双语词对齐;

⑤双语词典抽取:使用 SilkRoad1.0 词典抽取模块[18]从词对齐语料中抽取双语词典。

3.1.2 语义单元自动获取实验

选取 27800 个链语法正确分析的句对自动抽取语义单元。从中随机选择 235 个句对作为测试语料,采用准确率(P)、召回率(R)和 F 值人工评价语义单元自动抽取的性能,其表达式如下:

$$P = \frac{N_r}{N} \times 100\%$$

$$R = \frac{N_r}{N_a} \times 100\%$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\%$$

其中,N_r 表示自动抽取出来的正确的基本语义单元数,N 表示自动抽取出来的基本语义单元数,N_a 表示人工抽取的基本语义单元数。

3.1.3 基于语义单元的机器翻译实验

对实验①自动获取的语义单元进行人工检查。由于人力的限制,检查并选取了 10000 个正确语义单元进行翻译测试。从实验①之外的的双语句对中随机选取 1000 对进行汉英翻译。由于很多句子没有被语义单元覆盖,只有 234 句能翻译。以此 234 句作为测试集,采用 NIST 官方网站发布的 mteval-v11a-cmufix _ b.pl[20]进行评测。选择开源统计机器翻译 SilkRoad 1.0 的 CARAVAN[21]作为基准系统。

3.2 实验结果及分析

(1)实验①评测结果及分析

自动抽取测试结果见表 2。其中 *ESER* 为英语语义单元表示构造结果, *SER _BO* 为优化前双语语义单元抽取结果, *SER _AO* 为优化后的抽取结果。

表 2 语义单元抽取和词语对齐实验结果

	P	*R*	*F*
ESER	88.53%	95.69%	91.97%
SER _BO	60.80%	65.72%	63.16%
SER _AO	74.36%	73.78%	74.06%

从表 2 可以看出,英语单语的语义单元表示抽取的 F 值达到了 91.97%。双语语义单元的自动获取的 F 值为 63.16%。运用优化策略对语义单元自动获取进行优化之后,准确率和召回率均有较大提高,分别提高了 13.56% 和 8.06%, F 值也提高 10.90%。结果说明语义单元自动获取方法是有效的,优化策略在很大程度上避免了词语对齐错误的干扰。

(2) 实验②的结果及分析

语义单元翻译系统对比实验结果见表 3。其中 DUT-SEMT 是我们基于语义单元的机器翻译实验系统,CARAVAN 是基准系统。

表 3 对比实验结果

	NIST	*BLUE*
DUT-SEMT	8.7632	0.8578
CARAVAN	6.9876	0.5599

从表 3 可以看出,DUT-SEMT 翻译的性能优于统计机器翻译系统 CARAVAN。初步结果说明基于语义单元的机器翻译系统是有潜力的。

3.3　相关工作比较

与相关工作的不同主要表现在两个方面。获取的内容和获取的方法。语义单元理论认为翻译的基础是意义相等,而语义单元是意义的单位。语义单元在不同的语言中有不同的表示,在两种语言中的表示有确定的对应关系。语义单元的表示形式上看是实量和虚量串,但是它的构成是稳定的,是一种语言知识的表示。如一个动词所带的参数通常与它的配价有关。不同于翻译模板[5-11]和 E-Chunk[12],而且 E-Chunk 没有考虑变量的情况。

Kaji[5]利用双语的句法分析和双语词典的对应寻找翻译模板,但是需要健壮的双语的句法分析器,并且受词典的限制。刘群[7]改进了这种方法,利用约束规则成功地获取双语的句法规则转换模板。Lv[6]使用英语的句法分析器和一种双语平行语法获取句法模板。Liu[8]利用汉语的句法分析器和双语的词对齐获取句法树到词串的翻译模板。Quirk[10]利用依存树和双语词对齐获取翻译模板。吕[12]采用链语法获取其中的词串(短语)作为英语的 Chunk,再利用词语对齐寻找汉语的 Chunk。这几种方法和本文有些相近,但是,本文获取的是语义单元结构,目前缺少对句子进行语义分析的工具。因此本文采用英语的链语法分析器对句子作链语法分析,然后根据单词之间的链关系构造一种浅层的语义层次分析,并在此基础上获取英语的语义单元。然后利用词语对齐寻找汉语的语义单元表示。另一点不同的是本文采用了一系列的优化策略,对于词语对齐有一定的容错能力。

4　结　论

本文提出一种基于转换和映射的双语语料的语义单元自动获取方法。该方法在缺少对句子完全语义结构分析的情况下,借助链语法解决了英语句子中词语语义层次的问题,抽取语义单元的英语表示,有效地解决了语义单元结构获取的困难。针对从语义单元的英语表示到汉语表示过程中词语对齐错误,提出了一套优化策略,该策略较好地解决了语义单元在不同语言中表示对应的困难。初步结果显示该方法是有效的,而且基于语义单元的机器翻译系统很有潜力。下一步工作将针对链语法的不能正确处理的结果进行研究,提高语义单元自动获取的性能,并对语义单元理论进行深入的研究。

参考文献

[1] 高庆狮, 陈肇雄, 李堂秋. 类人机译系统原理. 计算机研究与发展, 1989,26(2):1-7

[2] Gao Q S, Hu Y, Li L, et al. Semantic language and multi-language MT approach based on SL. *Journal of Computer Science & Technology*, 2003, 18(6):848-852

[3] 高小宇, 高庆狮, 胡玥等. 基于语义单元表示树剪枝的高速多语言机器翻译. 软件学报, 2005, 16(11):1909-1919

[4] Fang M, Gao Q S, Yu Z B. A semi-automatic extraction of the SERB in machine translation based on SL. In: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2005.398-403

[5] Kaji H, Kida Y, Morimoto Y. Learning translation templates from bilingual text. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, 1992.672-678

[6] Lv Y J, Zhou M, Li S, et al. Automatic translation template acquisition based on bilingual structure alignment. *Computational Linguistics and Chinese Language Processing*, 2001, 6(1): 83-108

[7] 刘群. 汉英机器翻译中若干关键技术研究:[博士学位论文]. 北京:北京大学计算机系, 2004

[8] Liu Y, Liu Q, Lin S X. Tree-to-String alignment template for statistical machine translation. In: Proceedings of the 21 st International Conference on Computational Linguistics, Sydney, Australia, 2006.609-616

[9] Chiang D. A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43 rd Annual Meeting of the Association for Computational Linguistics, Michigan, USA, 2005.263-270

[10] Quirk C, Menezes A, Cherry C. Dependency treelet translation: syntactically informed phrasal SMT. In: Proceedings of the 43 rd Annual Meeting of the Association for Computational Linguistics, Michigan, USA, 2005.271-279

[11] Hu R L, Zong C Q, Xu B. An approach to automatic acquisition of translation templates based on phrase structure extraction and alignment. *IEEE Trans on Audio, Speech, and Language Processing*, 2006, 14(5):1656-1663

[12] 吕学强, 陈文亮, 姚天顺. 基于连接文法的双语 E-Chunk 获取方法. 东北大学学报(自然科学版), 2002, 23(9):829-832

[13] Sleator D, Temperley D. Parsing English with a Link Grammar. Pittsburg: Carnegie Mellon University, 1991

[14] 张华平. ICTCLAS. http://www.nlp.org.cn/project/project.php? proj_id=6:CNLP,2002
[15] SMT team. TokenizeE.perl.tmpl. http:// www.clsp.jhu.edu/ ws99/projects/mt/toolkit/:JHU,1999
[16] Temperley D, Sleator D, Lafferty D. Link grammar parser 4.0. http:// www.link.cs.cmu.edu/link/:CMU,2004
[17] Och F J. GIZA++. http://www.fjoch.com/GIZA++.html: ISI, 2003
[18] 何彦青. SilkRoad 短语抽取模块. http://www.nlp.org.cn/ categories/default.php? cat_id=21:CNLP,2006
[19] 卢伟. 英汉平行语料库. http://www.xmuoec.com/gb/hanyu/hanyu/data/corpus/index.htm:XMU,2006
[20] Papineni K. mteval-v11a-cmufix_b.pl. http://www.nist.gov/ speech /tests/mt/resources/scoring.htm:NIST, 2007
[21] 陈毅东. CARAVAN. http://www.nlp.org.cn/ categories/default.php? cat_id=21:CNLP,2006

Automatic extraction of semantic elements based on transformation and mapping

Fang Miao, Cao Jingxiang
(Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024)

Abstract

The paper proposes an approach to automatic extraction of semantic elements (SEs) from the English-Chinese bilingual parallel corpus. Its work consists of three parts: first, constructing a set of rules from the chain of English sentences parsed by the link parser to determine the relative logical hierarchy of each word to generate the English semantic element representation (ESER) based on the transformational algorithm, then, mapping the ESER into the Chinese semantic element representation (CSER) according to the statistical word alignment so as to obtain bilingual semantic element representations (SERs), finally, extracting SEs by optimizing the bilingual SERS through a series of strategies like SERs competition, unaligned word pair recall. This approach can extract SEs without complete semantic analysis from the bilingual parallel corpus. The experiments showed that the F-measure of SE extraction reached 74.06% and the SE based machine translation systems featured in their high accuracy.

Key words: semantic element, automatic extraction, link grammar, word alignment