基于语义单元表示树剪枝的高速多语言机器翻译*

高小宇[1], 高庆狮[1,2+], 胡 玥[1,2], 李 莉[2]

[1](中国科学院 计算技术研究所,北京 100080)

[2](北京科技大学 智能、语言与计算机科学研究所,北京 100083)

High Speed Multi-Language Machine Translation Based on Pruning on the Tree of Representations of Semantic Elements

GAO Xiao-Yu[1], GAO Qing-Shi[1,2+], HU Yue[1,2], LI Li[2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Institute of Intelligence, Linguistics and Computer Science, University of Science and Technology, Beijing 100083, China)

+ Corresponding author: E-mail: qsgao@public.bta.net.cn

Received 2004-08-10; Accepted 2005-02-03

Gao XY, Gao QS, Hu Y, Li L. High speed multi-language machine translation based on pruning on the tree of representations of semantic elements. *Journal of Software*, 2005,16(11):1909–1919. DOI: 10.1360/jos161909

Abstract: In this paper, a high speed multi-language machine translation approach based on pruning on tree representations of semantic elements is proposed. This is the multi-language machine translation with the following several characteristics: Chinese segmentation before translation into another languages is not necessary, and the translation time is $O(L)$ rather than general $O(LN)$, where L is the length of text, N is the number of semantic elements (i.e. number of language patterns) in SER-base, even if N is hundreds of thousands or millions.

Key words: natural language; machine translation; semantic element; representation of semantic element; semantic language

摘 要: 提出一种基于语义单元表示树剪枝的高速多语言机器翻译方法.此方法是一种将汉语翻译到其他语种不需要先进行汉语切分的多语言机器翻译方法.而且翻译时间为 $O(L)$而不是 $O(LN)$,其中,L 是文本的长度,N是语义单元库中语义单元的数量,一般有数十万或者数百万.

关键词: 自然语言;机器翻译;语义单元;语义单元表示;语义语言

中图法分类号: H085 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant No.60343010 (国家自然科学基金); the Foundation of Institute of Computing Technology, the Chinese Academy of Sciences under Grant No.20016250 (中国科学院计算技术研究所创新基金)

作者简介: 高小宇(1976－),男,福建漳州人,硕士,主要研究领域为数据库,自然语言与机器翻译,模糊集合,网络安全;高庆狮(1934－),男,教授,博士生导师,中国科学院院士,主要研究领域为大型、巨型计算机体系结构,并行算法与并行处理系统,自然语言及其处理,网络安全,模糊集合,人类智能及其应用;胡玥(1963－),女,博士,副教授,主要研究领域为并行算法,自然语言与机器翻译,网络安全;李莉(1980－),女,博士生,主要研究领域为自然语言处理.

半个多世纪以来,有许多没有夹杂语无伦次和正错混杂而且译文正确和比较符合目标语言习惯的机器翻译方法,包括句对句(sentence-to-sentence)、带变量的句子模式(sentence-pattern)和翻译记忆(translation-memory)法.但是它们的缺点是语言覆盖范围和语言知识库之比(简称覆盖率)过小.这些方法难以通用,因为我们需要覆盖的句子的数目非常巨大.

组块(chunk)是 Miller 教授从认知心理学角度提出的度量学习量和知识量的概念.Simon[1]指出,一个正常人的母语语言知识大约有 20 万个组块,留学生出国时所需的外语知识量大约是 1 万组块,出国攻读博士之后又工作了 15 年的留学人员,他们的外语知识量大约是 5 万个知识组块.

在 20 世纪 80 年代,中国科学院计算技术研究所提出了两个概念:“最小义元(meaning element)”和“语言模式(language pattern)”[2](分别相当于本文的“不可弃的不带变量的语义单元表示”和“带变量的语义单元表示”).例如,“bus pass”,“as 〈部分副词〉 as possible”,“〈数〉 times 〈数〉”,“Mr. 〈姓〉”,等等.其中,〈数〉,〈部分副词〉,〈姓〉是变量,〈〉内是该变量的类型,也是可代入的类型.变量可以由相同类型的(带变量或者不带变量的)语义单元表示所代入.这说明句子是由语义单元表示相互插入构成,而不是由线性排列构成.显然,不是所有的语义单元在所有的自然语言中都有它的表示.例如,“叶公好龙”在英语中就没有它的表示.

基于实例(examples-based)系统[3,4]增加了覆盖率,但能否只输出正确的、没有语无伦次、没有正错混杂的译文,有待于进一步发展.半个世纪以来的基本方法是基于规则的方法[5].但是,除了比较窄的领域外,翻译质量还很难达到实用要求.在 20 世纪 90 年代,基于统计(statistics-based)的方法[6]由于其翻译质量有明显飞跃而迅速受到广泛关注,但能否只输出正确的、没有语无伦次、没有正错混杂的译文,也有待于进一步发展.20 世纪 90 年代以来,模式(patterns)、模板(templates)和“组块(chunks)”[7−10]方法以及与统计方法结合,都得到了重视.“WORD-NET”[11]和“知网”[12]在消歧方面的作用也得到了广泛重视.

1 语义语言[13,14]

语义语言的基本概念及形式定义已经在文献[13,14]中给出详细讨论,这里只作简单介绍.

1.1 语义语言的基本概念

使用不同自然语言的人们彼此可以沟通,不同自然语言可以互译是因为它们的句子有相同语义.一般来说,字和词不一定有对应关系.

句子的语义称为句义(SS).句义内表达一个意思的单元称为语义单元(SE).句义由语义单元组成.句义本身也是语义单元.语义单元在一个自然语言-I(例如,英语、汉语,等等)上的表示称为语义单元表示(SERi).一个自然语言-I 上的句子就是一个句义(SS)在该自然语言-I 上的表示(SSRi).从这个角度来看,句子(句义表示)是由(带变量和不带变量的)语义单元表示构成的,这样可以地清楚看到不同语言句子的对应关系.如果把句子看作由字的简单线性排列构成的,就看不到这种对应关系.

全部语义单元构成语义语言(SL).显然,语义语言包括全部句义.全部在 I-自然语言的语义单元表示构成自然语言-I.显然,自然语言-I 包括全部在 I-自然语言的句子.一个自然语言可以看成是语义语言的一个表示.两个自然语言(I,J)之间的翻译可以看作两种表示之间的转换.

自然语言理解就是对语义的理解.语义有 3 种不同的层次:语言层次、知识层次和语用层次.

例如,“把这杯水倒入那缸浓硫酸里.”这句话包含“这〈N〉”,“杯〈N〉”,“水”,“那〈N〉”,“缸〈N〉”,“浓硫酸”,“把〈N〉倒入〈N〉里”等 7 个语义单元表示.其中,N 表示名词.

如果一个机器人只有语言层次的理解,它就能够根据“这〈N〉”找到一个“标有‘水’的‘杯’ ”,再根据“那〈N〉”找到一个标有“浓硫酸”的“缸”,然后执行“把“标有‘水’的‘杯’ ”倒入“标有‘浓硫酸’的‘缸’ ”里”的命令.

如果一个机器人有知识层次的理解.它就不会马上执行“把“标有‘水’的‘杯’ ”倒入“标有‘浓硫酸’的‘缸’ ”里”的命令,而是警告主人:“危险”.因为它有知识层次的理解:“把‘水’倒入‘浓硫酸’里”会爆炸.

知识层次的语义本身又可以分许多层次.例如,“苹果”的知识层次的语义对小孩和对研究苹果的专家差别极大.对苹果专家而言,其语义的描述需要厚厚的一本书.

语用层次理解的例子.“今天是星期天”可以有多种语用层次的理解:仅仅是回答今天星期几,妻子提醒丈夫该休息休息,小孩提醒父亲该实现星期天带他到动物园玩的许诺,等等.

在本文中,如果不另加说明,语义仅限于语言层次的语义.

语义语言的形式定义、自然语言的形式定义、可弃语义单元与基本语义单元的优先次序从略,可参见文献[14].

1.2 语义单元表示的构成

不带变量的语义单元对应的在I-自然语言上的语义单元表示是由I-自然语言上的字构成的字串,我们称其为纯实量.例如,“中华人民共和国”是一个由7个汉字构成的字串,也就是一个汉语实量.“The People Republic of China”是一个由5个英文的字构成的字串,也就是一个英语实量.带变量的语义单元对应的在I-自然语言上的语义单元表示由 I-自然语言上的实量和由类型串构成的虚量混合构成.例如,“〈名词〉是由〈名词 2〉构成”是一个带变量的语义单元表示,它由两个实量(“是由”和“构成”)和两个虚量(“〈名词〉”和“〈名词 2〉”)构成.“〈名词〉 consists of 〈名词 2〉”是一个带变量的语义单元表示,它由一个实量(“consists of”)和两个虚量(“〈名词〉”和“〈名词 2〉”)构成.

1.3 在语义语言中的语义单元表示列举

例如,“先生”,“〈姓|名|姓名|别名〉先生”,“〈(女性代词|女性(名|姓名|别名))的所有格〉先生”是 3 个有关“先生”的不同的语义单元的汉语表示.其对应的英语表示分别是“sir”,“Mr. 〈姓|名|姓名|别名〉”,“〈(女性代词|女性(名|姓名|别名))的所有格〉(丈夫|老公)”.其中|表示或.

又如,汉语句子“张先生是工程师”的句义就是“姓张的先生的职称是工程师”.这个句义(SS)是一个语义单元(SE),可以用“I_{sTP}(Mr. (Z_{hang}),$e_{ngineer}$)”来表示.其中,“I_{sTP}(〈关于人的名词或代词〉,〈职称〉)”,“Mr. (〈姓|名|姓名|别名〉)”,“Z_{hang}”和“$e_{ngineer}$”是4个语义单元,I_{sTP}(X,Y)的参量X的类型是与人有关的名词、代词对应的事物义,I_{sTP}(X,Y)的参量Y的类型是与职称有关的事物义,M_r. (X)的参量X的类型是姓、名或者姓名.这个句义和4个语义单元在汉语上的表示分别是“张先生是工程师”,“〈关于人的名词或代词〉是 $_{TP}$〈职称〉”,“〈姓|名|姓名|别名〉先生”,“张”和“工程师”.这个句义和4个语义单元在英语上的表示分别是“Mr. Zhang is an Engineer”,“〈关于人的名词或代词〉 is_{TP} a 〈职称〉”,“Mr. 〈姓|名|姓名|别名〉”,“Zhang”,“engineer”.它们还可以如表1和表2所示写成“1(2(3),4)”,“1(X_{PS},X_{TP})”,“2(X_{NA})”,“3”,“4”.

Table 1 Semantic elements and their representations

表 **1** 语义单元及其表示

Semantic element	Number and type of parameter	Representation of SE in Chinese	Representation of SE in English	Representation of SE in Japanese
1 (#N_{PS} #N_{TP})	2, N_{PS}, N_{TP}	X_1 是 X_2	X_1 is a X_2	X_1 は X_2 です
2 (#N_{NA})	1, N_{NA}	X 先生	Mr. X	X さん
3	0	张	Zhang	张
4	0	工程师	engineer	技师

Table 2 Remembrance approaches to represent SE

表 **2** 语义单元便于记忆的写法

SE	One approaches	Remembrance approaches to represent SE
1 (#N_{PS}, #N_{TP})	1(X_{PS}, X_{TP})	是$_{职称}$($X_{1\,人}$, $X_{2\,职称}$) = Is_{TP}(X_{1HU}, X_{2TP}) = です ($X_{1\,人}$,$X_{2\,职称}$)
2 (#N_{NA})	2(X_{NA})	先$_{生}$(X) = $M_{r.}$(X_{LN}) = さん(X $_{姓}$)
3	3	张= C_{Zhang} = 张
4	4	工$_{程师}$ = $e_{ngineer}$ = 技$_{师}$

2 基于语义语言的多语言翻译系统的翻译过程

两种自然语言(I,J)的句子或者文本的翻译可以用两步来实现.第1步,把自然语言-I通过“在自然语言-I上的语义分析”变为语义语言的句子或者文本.在语义语言的句子称为句义表达式SSE.第2步,把语义语言的句子或者文本,通过简单地“在自然语言-J上的代入展开”成为自然语言-J的句子或者文本.第2步十分简单,执行时间

可以忽略不计.所以,同时翻译成多种自然语言的执行时间仍然可以忽略不计.例如,句义表达式 1(2(3),4)(即 $I_{sTP}(M_r.(Z_{hang}),e_{ngineer})$)展开成为汉语的过程:1(2(3),4)=>1(2(张′),工程师′)=>1(张先生′,工程师′)=>张先生是工程师.展开成为英语的过程:1(2(3),4)=>1(2(Zhang′),Engineer′)=>1(Mr. Zhang′,Engineer′)=>Mr. Zhang is an Engineer.其中,*X*′表示语义单元,表示 X 对应的语义单元.

翻译时间主要是第 1 步"在自然语言-I 上的语义分析"的执行时间.本文提出一种快速的"在自然语言-I 上的语义分析"方法,其执行时间 $T=O(L)$而不是 $T=O(LN)$.其中,L 是句子或者文本的长度,N 是在语义单元库中的语义单元的总数,一般说来,N 可以超过数百万.

3 语义单元表示树

3.1 语义单元表示的标准形式

令 S 表示语义单元表示中的实量串,X 表示语义单元表示中的虚量串(也称为变量串).那么,所有的语义单元表示都可以由以下 4 种基本形式(S,SX,XS 和 XSX)和连接符($)构成(见表 3).例如,英文"The People Republic of China"是一个由 5 个英文字组成的实量串,基本形式是 S;汉语"中华人民共和国"是一个由 7 个汉字组成的实量串,基本形式也是 S;"〈X 整数(整数)〉〈Y 人(关于人的名词代词)〉"是一个由两个虚量(变量)组成的虚量串,基本形式是 X,其中,( )中是变量的类型.英语"〈X_N(名词或代词)〉 consists of 〈Y_N(名词或代词)〉"是 XSX 形式,汉语"〈X_N(名词或代词)〉是由〈Y_N(名词或代词)〉组成"是 XSXS 形式.除了纯虚量(变量)串以外,所有的匹配比较总是从实量开始,然后是实量前的虚量,最后才是实量后的虚量.符号|表示开始,连接符$表示两个基本形式之间的连接.

Table 3 SER consists of four basic types (S,SX,XS, and XSX) and their connections ($)

表 3 语义单元表示由 4 种基本形式(S,SX,XS 和 XSX)和连接符($)构成

Types of SER	Realization (match begin from |)	Types of SER	Realization (match begin from |)
		X	|X
S	|S	XS	X|S
SX	|SX	XSX	X|SX
SXS	|SX $|S	XSXS	X|SX $|S
SXSX	|SX $|SX	XSXSX	X|SX $|SX
SXS...XS	|SX $|SX $... $|S	XSXSX...S	X|SX$|SX $... $|S
SXS...XSX	|SX $|SX $... $|SX	XSXSX...SX	X|SX$|SX $... $|SX

3.2 实量从**"more"**开始的语义单元表示树的一个例子

表 4 是一个英语语义单元表示树的例子.*表示语义单元的类型和序列号等等.

Table 4 An example of SER-Tree starting with real-string "more"

表 4 一个实量从"more"开始的英语语义单元表示树的例子

SER									
V	more	N	*						
J,	more	J	*						
	more	than	SL	*					
	more	than	N2	$'s	*				
	more	than	J	*					
	more	than	N	$can	describe	*			
	more	than	N	$can	shake	A	stick	at	*
	more	fire	in	N	$'s	bed	straw	*	
	more	N	$than	N2	$'s	*			
	more	N	$than	J	*				
	more	A	$than	A2	*				
	...	...	...	...	...				

3.3 语义单元表示第**2**个及其后的实量串的处理

表 4 中的所有第 2 个及其后的实量串之前都加一个符号$,见表 5.

进一步地,由两个以上实量串构成的语义单元表示可以分解成为一个主表示和子表示集.主表示是在(见表 5)原来的树中用语义单元表示序号代替$之后的部分,见表 6.子表示集是由各语义单元表示以及$之后的部分构

成的,见表 7.

Table 5 The SERs which have two or more real-strings

表 5 两个以上实量串的语义单元表示

more	than	N2	$'s	*	
more	than	N	$can	describe	*
more	N	$than	N2	$'s	*
more	N	$than	J	*	
more	A	$than	A2	*	

Table 6 Main-Representation

表 6 主表示

more	than	N2	$1	*	
more	than	N	$2	*	
more	than	N	$3	*	
more	fire	in	N	$4	*
more	N	$5	*		
more	N	$6	*		
more	A	$7	*		

Table 7 Sub-Representation

表 7 子表示

No.	Sub-Representation					
1	$'s	*				
2	$can	describe	*			
3	$can	shake	a	stick	at	*
4	$'s	bed	straw	*		
5	$than	N2	$1	*		
6	$than	J	*			
7	$than	A2	*			

3.4 语义单元表示树的形成

begin

〈把自然语言 I 的语义单元表示库中全部有两个及两个以上实量串的语义单元表示分解成为主表示加子表示集,并且用主表示修改语义单元表示库,同时把子表示集并入库中〉;

〈用特殊标志代替自然语言 I 上的语义单元表示库中的全部变量串(即虚量串)〉;

〈对自然语言 I 上的语义单元表示库中的全部语义单元表示进行排序,特殊标志按空处理〉;

〈排序后,所有特殊标志还原回原来的变量串〉;

〈把语义单元表示库转换成树形(称为语义单元表示树)〉;

end

例如,表 4 可以转化成表 8.

Table 8 SER getting from Table 4 after resolution

表 8 由表 4 分解后形成的语义单元表示

No.	SER					
	V more	N	*			
	J, more	J	*			
	more	than	SL	*		
	more	than	N2	$1	*	
	more	than	J	*		
	more	than	N	$2	*	
	more	than	N	$3	*	
	more	fire	in	N	$4	*
	more	N	$5	*		
	more	N	$6	*		
	more	A	$7	*		
1	's	*				
2	can	describe	*			
3	can	shake	a	stick	at	*
4	's	bed	straw	*		
5	than	N2	$1	*		
6	than	J	*			
7	than	A2	*			

表 8 可以转化成为以下 4 个树形表示:

more (N-V*,J-J,*,than (SL*,N2·$1*,J*,N($2*,$*3)),fire·in·N·$4,N($5*,$6*),A·$7*);

's(*1,bed·straw*4);

can (describe*2,shake·a·stick·at*3);
than (N2·$1*5,J*6,A2*7).

4 基于语义单元表示树剪枝的快速语义分析算法

4.1 基于语义单元表示树剪枝的快速语义分析算法

```
begin
  while 〈自然语言-I 上的篇章非空〉
  begin
    〈从自然语言-I 上的篇章中取一句〉;
    while 〈句子非空〉
    begin
      〈从句子中相续地取一个字(w)〉;
      〈取以该字(w)为开始的最多一棵语义单元表示树〉;
      〈对已经取出的所有语义单元表示树,根据字(w)进行剪枝,把与(w)不同的枝剪掉〉
      〈对类型流进行剪枝处理〉
    end;
      〈从剪枝后没有被剪掉的语义单元表示中求句义表达式 SSE,并输出句义表达式 SSE〉
  end
end
```

4.2 基于语义单元表示树剪枝的快速翻译算法

```
begin
  while 〈自然语言-I 上的篇章非空〉
  begin
    〈从自然语言-I 上的篇章中取一句〉;
    while 〈句子非空〉
    begin
      〈从句子中相续地取一个字(w)〉;
      〈取以该字(w)为开始的最多一棵语义单元表示树〉;
      〈对已经取出的所有语义单元表示树,根据字(w)进行剪枝,把与(w)不同的枝剪掉〉
      〈根据类型流处理规则对类型流进行处理〉
    end;
      〈从剪枝后没有被剪掉的语义单元表示中求句义表达式 SSE〉;
      〈基于语义单元表示库或者语义单元表示树库,把 0 个,1 个或者多个句义表达式通过代入求出一个
        或者多个目标语言的译文,并且输出一个或者多个目标语言的译文〉
  end
end
```

4.3 实量串之前及之后的虚量串的处理次序

实量串之前及之后的虚量串的处理次序是先实量串,然后是实量前的虚量串,最后是实量后的虚量串.表 9 给出了几个例子.

Table 9 Process order

表 9 处理次序

SE					Process order (real-string first, then virtual-string before RS, VS after RS at last)			
V	more	N	*		more	V	N	*
J,	more	J	*		more	J,	J	*
SL,	more	or	less	*	more	or	less	SL,*

4.4 例 子

He has more books than Tom's:他有比汤姆多的书.此句在英语上的语义单元表示树见表 10,其剪枝过程见表 11.

Table 10 SER index tree in English

表 10 在英语上的语义单元表示树

No.	SER in Language-I (English)							Language-J (Chinese)	Type	Illustration
1		he						他	Nr	
2	N	has	N_2					N 有 N_2	J	
3		more	N	$than	N_2	$'s	*	比 N_2 多的 N;	N	3 can be replaced by 3′
3′		more	N	$4′						3′ includes 4′
4′		than	N_2	$5′						4′ includes 5′
5′		's								
6		books						书	N	
7		Tom						汤姆	Nr	

Table 11 The process of pruning

表 11 剪枝过程

	Sentence	Tree of SER	Type	Process of pruning
1	He	He	Nr	**1**: Nr
2	has	N |has N	J	2:V ↑1N: NV -# ↓N **2:J**
3	more	more (...|N (...|$4′)|...)	N	3=3′, -↓|4' -#□N **-□N =3**
6	books	Books	N	**6**:N **more 6**
4′	than	than(...|N(... |$5′)...)		4′, -↓|5′ -#□N **-□N =4**
7	Tom	Tom (|′)	Nr	**Tom** **7**: Nr
5′	's	's		**5**′

在表 11 中,|表示实量开始,↑表示向上处理实量前的虚量,↓表示向下处理实量后的虚量,↓|表示处理下一个实量,□表示返回处理两个实量之间的虚量,#表示等待,下横线表示匹配成功,-表示尚需.

说明:"he"(1),"books"(6),"Tom"(7)和"'s"(5′)这 4 个实量串的匹配立即成功,虽然"Tom"从未用到.接着,"Tom's"(7)和'more 6'="more books"匹配成功.然后,3=3′="more N $4′"'匹配成功.最后,2="N has N"匹配成功.

4.5 类型流处理规则

表 12 中列举了一些类型流处理规则.

Table 12 The example of rules about type-flow processing

表 12 类型流处理规则列举

Left side	Right side
S+N measure unit	Semantic meaning of quantity SL
SL+N	Semantic meaning of thing N
A+N	Semantic meaning of thing N
Determinant of degree F+A	Semantic meaning of adjective A
…	…

4.6 高速语义分析的计算时间T

设 L 是待翻译文本的长度,l 是平均语义单元表示的长度,m 是平均语义单元树结点的分支数目,N 是在语义单元表示库中语义单元的数目.

因为取语义单元表示树的时间是取 L 棵“语义单元表示树”的时间,每棵树中语义单元表示平均长度为 l,每个节点平均分叉数为 m.所以,取“语义单元表示树”的总时间=$O(Llm)$.

因为剪枝的时间是根据 L 个字逐一对已经取出的语义单元树进行剪枝,每个字要对已经被取出的数目为 x 的树进行剪枝,其中 $x<l$,l 是每棵树中语义单元表示的平均长度.对每棵树的一个节点平均比较 m 次.所以,总剪枝时间是$(0+1+2+...+l+l+...+l)m \le O(Llm)$.

因此,高速语义分析的计算时间 $T=O(Llm)=O(L)$,而不是 $O(LN)$,其中 L 是文本的长度,N 是在语义单元表示库中语义单元的数目,N 可以是数十万、数百万,甚至更多.对英语而言,树的结点平均分支数目 m 约等于 1.717,语义单元表示的平均长度 l 约等于 4.38;对汉语而言,l 约等于 4.17(根据从一部英汉辞典初步提取的 30 多万语义单元表示库进行的统计).

4.7 歧义求解列举

例子:He saw a girl with a telescope.

表 13 给出了例句剪枝后的 8 个语义单元及其在英语和汉语的语义单元表示,表 14 给出了求句义表达式的过程.

Table 13 There are eight SEs and their SERs in English and Chinese after pruning

表 **13** 剪枝后的 8 个语义单元及其在英语和汉语的语义单元表示

SE	1(N,N,N)	2(N,N)	3(N,N)	4	5	6(N)	7	8
SER in English	N_1 saw N_2 $8	N_1 with N_2	N_1 sawN_2	he	Telescope	a N	girl	with N_3
SER in Chinese	N_1 用 N_3 看见 N_2	带着 N_2 的 N_1	N_1 看见 N_2	他	望远镜	一 L(N) N	女孩	
Type of SE	J	N	J	N	N	N	N	for 1

Table 14 The processing of SS-expression solution (In practice, SS-expression solution and pruning can be unified)

表 **14** 求句义表达式过程(实际上,SS-表达式的求解过程和剪枝过程可以合并)

	SE number-SE-SER in English-SER in Chinese-type	The processing of SS-expression solution
He	4-he-他-N	4:N
saw	1-N_1 saw N_2 with N_3-N_1 用 N_3 看见 N_2-J 3- N_1 saw N_2-N_1 看见 N_2-J	1(4,6(7),6(5)):J*| 3(4,2(6(7),6(5))):J*
a	6-a N-一 L(N)N-N	6(7):N
girl	7-girl-女孩-N	7:N
with	2-N_1 with N_2-带着 N_2 的 N_1-N	2(6(7),6(5)):N
telescope	5-telescope-望远镜-N	5:N 6(5):N

实际上,求句义表达式过程与剪枝过程可以同时进行.

可以求出两个句义表达式:1(4,6(7),6(5))和 3(4,2(6(7),6(5))).

两个句义表达式只对应一个英语句子:“He saw a girl with a telescope.”.

两个句义表达式对应两个汉语句子:

“他用一个望远镜看见一个女孩” (对应于句义表达式 1(4,6(7),6(5)));

“他看见一个带着一个望远镜的女孩” (对应于句义表达式 3(4,2(6(7),6(5)))).

4.8 切分歧义和未登录名的识别

由于本翻译算法不需要事先切分,切分歧义如果不能形成合法句子,就自动被剪枝剪掉;如果形成合法句子,就输出歧义译文句子.关于未登录名的识别,将另文讨论.

5 翻译系统新评价标准、翻译难易和实验选择和预测

5.1 翻译系统新评价标准

通常评价标准是翻译正确给正分,翻译不正确不给分.每一页都有错,实际上无法使用的译文仍然可以得到 99%~99.5%的高分.

新的评价标准是:如果原文正确,翻译出正确结果,给正分;翻译不出来,不给分;翻译出语无伦次的译文给负

高分(例如,负 5~10 倍);翻译出错误结果给负极高分(例如,负 50~100 倍).

5.2 文本类别的翻译难易及实验选择

众所周知,科技、合同、法律类比较容易,因为歧义少;生活类次之,最常用的几百字,用法很活;小说比较难;诗最难,往往不是需要单纯翻译,而是需要再创造.

因此,实验采用科技、合同、法律类,偏容易;采用小说,又偏难.所以,采用生活类比较折衷.

5.3 实验及结果

一个 5 000 语义单元的英汉双语库,1 万个与双语库内容相近的日常生活句子.实验结果:因为库中的知识不够,近一半翻译不出来.即语义单元及其表示不全,不能满足求解该句子的需要.其中 6 处出现不正确和语无伦次,其原因是语义单元及其表示提取不全正确.例如,“It is cold in outside”译成为“是户外的感冒”是库中语义单元表示有错:“cold:感冒”应该改成“cold:寒冷的:(A)”;“a cold:感冒:(N)”译文应该是:“户外是寒冷的”.经过排除错误的语义单元以后,实验的例子中的不正确和语无伦次,可以排除.

附录给出了几个例子的译文对比.其译文正确性的关键是语义单元的正确提取.

6 结 论

本文提出了一种基于语义单元表示树剪枝的快速多语言机器翻译方法.这种方法将汉语翻译成其他语种不需要先进行汉语切分的多语言机器翻译方法.而且翻译时间为 $O(L)$而不是 $O(LN)$,其中,L 是文本的长度,N 是语义单元库中语义单元的数量,一般可能超过百万.初步实验说明,它是值得探索的一种新方法.

References:

[1] Simon HA. Cognitive Psychology, Lecture. Beijing: Peking University, 1983.

[2] Gao QS, *et al*. The principle of human-like machine translation. Computer Research and Developement, 1989,26(2):1−7 (in Chinese with English abstract).

[3] Brown RD. Example-Based machine translation in Pangloss system. In: Proc. of the COLING'96. Copenhagen, 1996. 169−174.

[4] Brown RD. Automated generalization of translation examples. In: Proc. of the COLING 2000. Sarbrucken, 2000. 125−131.

[5] Zhao TJ. The principle of machine translation. Harbin: Harbin Institute of Technology, 2001 (in Chinese).

[6] Och FJ, Weber H. Improving statistical natural language translation with categories and rules. In: Proc. of the 35th Annual Conf. of the Association for Computational Linguistics and the 17th Int'l Conf. on Computational Linguistics. Montreal, 1998. 985−989.

[7] Och FJ, Tillmann C, Ney H. Improved alignment models for statistical machine translation. In: Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park: University of Maryland, 1999. 20−28.

[8] Kaji H, Kida Y, Morimoto Y. Learning translation templates from bilingual texts. In: Proc. of the COLING'92. 1992. 672−678.

[9] Abney SP. Parsing by chunks. In: Berwick R, Abney S, Tenny C. Principle-Based Parsing. Kluwer Academic Publishers, 1991. 257−278.

[10] Cicerkli I, Guvenir HA. Learning translation templates from examples. Information Systems, 1998,23(6):353−363.

[11] Miller GA. WordNet. 1998. http://www.wordnet.princeton.edu

[12] Dong ZD, Dong Q. How Net. 1999 (in Chinese). http://www.keenage.com

[13] Gao XY, Gao QS, Hu Y. The multi-language machine translation approach based on Semantic language. Report of Invent Patent, ZL01131689.6, 2001 (in Chinese with English abstract).

[14] Gao QS, Hu Y, Li L, Gao XY. Semantic language and multi-language MT approach based on SL. Journal of Computer Science & Technology, 2003,18(6):848−852.

附中文参考文献:

[2] 高庆狮,等.类人机译原理.计算机研究与发展,1989,26(2):1−7.

[5] 赵铁军.机器翻译原理.哈尔滨:哈尔滨工业大学出版社,2001.

[12] 董振东,董强.知网.1999. http://www.keenage.com

[13] 高小宇,高庆狮,胡玥.基于语义语言的机器翻译系统和方法.发明专利报告, ZL01131689.6, 2001.

附录: 译文对比的几个例子.

原	**I want my steak well-done.**	**“Away the mare! ”he said to his friend setting sail for Hankou.**	**Away with your nonsense.**
1	我想要我的牛排做得好.	“离开母马!”他对他的朋友(出发去 Hankou)说.	离开用你的废话.
2	我希望我的牛排做得出色的.	他对他的朋友设置帆说离开为汉口母马的.	废除你的胡说八道.
3	我想要我的牛排熟透.	“离开牝马!”他向他的为 Hankou 起航的朋友说.	离开与你的胡说八道.
4	我想要牛排干得出色的.	“这个牝马!”他为 Hankou 对他的朋友设置的帆(航行)说了	对于你(们)的胡说.
本	我的牛排要全熟的.	他对动身到汉口去的朋友说:“前途珍重!”	不要胡说,滚开.
原	**“Away with him!” he said.**	**Put the receiver closer to your mouth. I can't hear you.**	**She shouted, “Away with it!”**
1	“离开与他!”他说.	放更接近于你的嘴的接收者.我不能听到你.	她大叫,“离开用它!”
2	“带走他"他说	把收件人靠近你的嘴.我不能听你.	她“废除它”喊叫!
3	她呼喊了,“离开与它!”	把接收装置放近你的嘴.我不能听见你.	“离开与他一起!”他说.
4	“与他!”他说了.	放更紧(更近)你(们)的嘴的接收器.我不能听到你(们).	她用它(这)大声说了,“!”
本	他说:“赶他出去!”	把话筒放近你的嘴.我听不见.	她大声喊道:“把它拿走!”
原	**I cannot away with his reproaches.**	**But will it be better than what we see since 1973?**	**There seems no trace of water here.**
1	我不能离开用他的责备.	但是它将是与我们从 1973 起明白什么相比好吗?	似乎没有这里水的痕迹.
2	我废除他的责备不录制.	但是它将是与我们从 1973 起明白什么相比好吗?	似乎没有这里水的痕迹.
3	我不能离开与他的责备.	但是它将比我们自从 1973 看见的好一些?	不似乎是这里的水的踪迹在那里.
4	我不能以他的责备.	但是意愿它与我们从 1973 起看见的的相比,是更好的吗?	那里好像不水的踪迹这里.
本	我不能忍受他的申诉.	但是它将是比我们从 1973 年以后所看到的更好吗?	看来这里没有水的痕迹.
原	**I spoke to manager himself.**	**Will it be as good as the golden-age of the fifties?**	**Now it’s what you do on the internet.**
1	我说话与经理他自己.	它将是实际已经五十的黄金时代吗?	现在它是在因特网上你所做的.
2	我自己跟经理说话.	它请是实际已经的五十黄金时代?	现在它是在因特网上你所做的.
3	我讲对管理人自己.	将它像金色年龄一样好 50?	现在它是你在因特网上做的.
4	我和这个管理人员说自己.	它该多好像这五十的黄金时代一样好吗?	现在这是什么你在因特网上做.
本	我对经理他本人说过.	它将像五十多岁的黄金年龄那样好吗?	这就是你现在在互联网上所做的事.
原	**He avows himself (to be) an artist.**	**The baby, awakened from the sleep, smiled at the sight of its mother.**	**You should not avoid his company.**
1	他自己公开承认一位艺术家.	婴儿,由于睡眠唤醒,一看见它的母亲就微笑.	你不应该避免他的公司.
2	他承认他自己(是)一位艺术家.	从睡眠唤醒小孩对看见它的母亲微笑.	你不应该避开他的公司.
3	他声明他自己(是)一个	婴儿,从睡觉醒来了,看见那光景它的母亲	你应该不避免他的公司.

	艺术家.	微笑了.	
4	他承认自己(是)一个艺术家.	这个宝宝,从睡眠唤醒(起),对看见它的母亲笑了.	你(们)不应该避免(避开)他的公司.
本	他自称为艺术家.	那个刚醒的婴儿一看见他的母亲就笑了.	你不应和他疏远.
原	**Would you please pass the salt?**	**Great achievements have been made in this field over the past few years.**	**He has come here only for a few days.**
1	可以请你递给盐吗？	大的成就在过去几年中已经在这个领域制造.	他已经只是几天来这里.
2	请你通过盐好吗?	在过去几年的时间中伟大成就已经这田野制造.	他这里已来拿一点几天仅.
3	请您传递食盐？	大成就在过去的很少年上在这田野被做了.	他仅仅有一些天的时间来这里了.
4	(请)你递过盐吗?	在这个领域(田地)关于过去极少年已经制造了很大成就.	他来(了)这里仅仅几天.
本	请把盐递给我好吗?	在这个领域取过去几年中得了很大的成就.	他已经来这儿只呆几天.
原	**She didn't answer for several minutes.**	**You're heading for an accident if you drive after drinking alcohol.**	**The medicine won't take action for hours.**
1	她没对几分钟负责.	如果在吸收酒精之后,你驾驶,你是善于一事故的头脑.	药许多小时将不采取行动.
2	她不做在几几分钟中答案.	如果在吸收酒精之后,你驾驶,你前往一事故.	药将不采取行动好几个小时.
3	她若干分钟没负责.	如果你在喝酒精以后开车,你是为一个事故的标题.	药不会有小时的时间采取行动.
4	她没对几分钟负责.	如果你们在喝酒以后开车(动),你们是朝事故进发.	药将不对数小时吃.
本	她几分钟后才回答.	如果你酒后开车,你注定是要出事的.	这药要过几小时才能生效.
原	**You're wanted on the telephone.**	**She often avoids him like a leper.**	**He is awaiting your convenience in the sitting room.**
1	你被在电话上想要.	她经常像一名麻疯病人一样回避他.	他正在起居室里等待你的便利.
2	你在电话上被寻找.	她像一个麻风病患者常常避开他.	他等待你的在起居室中方便.
3	你是有电话的.	她经常像一名麻风病患者一样避免他.	他在客厅正在等候你的便利.
4	你(们)在电话上(被)要(求).	她经常避免(避开)像一个麻风病患者那样的他.	他在起居室中在等待你(们)的方便.
本	有人找你接电话.	她时常远远避开他.	他正在起坐室里等你.
原	**He awaked her from ignorance.**	**You'll awaken a sleeping dog by doing so.**	**He awakened to the importance of that matter.**
1	他由于愚昧无知唤起她.	你将通过如此做唤醒一条睡的狗.	对那件事情的重要性他唤醒.
2	他从无知唤醒她.	你将通过做那样唤醒一条睡觉狗.	他向那物质的重要性醒来.
3	他从无知唤醒了她.	你将由这样做唤醒一条睡觉的狗.	他醒来了到那件事的重要性.
4	他出于无知唤醒(起)了她.	你(们)通过做如此将唤醒(起)一条睡觉狗.	他对那物质(事情)的重要(性)醒着.
本	他启发她.	你如此做会无事生非的.	他察觉了那事的重要性.