基于浅层句法信息的翻译实例获取方法研究

陈　鄞　　赵铁军　　杨沐昀　　李　生
(哈尔滨工业大学国家教育部微软重点实验室　哈尔滨　150001)

摘　要：翻译实例库是基于实例的机器翻译系统的主要知识源。本文采用基于浅层句法分析的方法进行翻译实例的获取。首先根据浅层句法信息划分源语言和目标语言的翻译单元，然后在词对齐结果的指导下，利用统计对齐模型确定源语言和目标语言翻译单元之间的映射关系，从而获取翻译实例。通过与几种较具代表性的翻译实例获取方法进行对比实验发现，无论是对翻译实例库直接评测，还是通过 EBMT 系统进行间接评测，该方法都获得了令人满意的效果。

关键词：翻译实例库；基于实例的机器翻译；浅层句法分析
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2009)02-0310-04

Extraction of Translation Example Based on Shallow Parsing Information

Chen Yin　　Zhao Tie-jun　　Yang Mu-yun　　Li Sheng
(*MOE-MS Key Laboratory of Natural Language Processing and Speech*,
Harbin Institute of Technology, *Harbin* 150001, *China*)

Abstract: Translation example base is the main knowledge source of example-based machine translation system. In this paper, a shallow parsing information based approach is proposed to extract translation examples. First, translation units of source and target language sentences are segmented respectively according to shallow parsing information. Then, guided by word alignment result, an statistical model is used to align translation units between source and target translation units, and thus translation examples are extracted. Experiment result shows that the proposed method achieves satisfying result in both direct evaluation of example base and indirect evaluation by EBMT system.

Key words: Translation example base; Example-based machine translation; Shallow parsing

1　引言

EBMT 系统的主要知识源是双语对照的翻译实例库。翻译实例通常是从双语语料库中获得的，获取方法主要分为两类：基于结构分析的方法和基于概率统计的方法。

基于结构分析的方法通常是分别对两种语言进行句法结构分析，然后根据一定的映射方式进行双语的结构对齐如文献[1]。基于结构分析的方法在语言学知识的指导下，理论上可以获得准确率较高的对译片段，但是，由于单语句法分析技术目前尚不十分成熟，分析结果还很难保证。

基于概率统计的方法中，包括一类基于统计机器翻译词对齐结果的短语对译片段抽取方法，如文献[2]。另外，还有基于概率词典的方法[3]、基于词对齐的方法[4]、基于词语关联度和双语统计同现测度的方法[5]等等。基于概率统计的方法大多具有较强的通用性和灵活性。但是，抽取出来的短语通常不是语言学意义上的短语，因此，在译文分析时缺少一定的直观性。

另外，还有一些研究者采用了基于标志词方法[6,7]。但是，由于标志词主要是一些预先定义好的虚词，据此获取的翻译片段通常粒度过大，实例过长，造成匹配效率低下，而且对于汉英这样结构差异较大的语言，虚词的对齐实际上是一个很困难的过程。

针对以上几类方法的优缺点，本文采用一种折衷的方法——基于浅层句法分析与概率统计词典的方法。与完全句法分析相比，浅层句法分析技术已经比较成熟，准确率都能保证在 90%以上。而且，EBMT 本身的特点也不提倡对句子作深入的句法分析。在此基础上，通过概率统计词典来辅助完成翻译等价单元的映射过程。

2007-07-25 收到，2007-12-31 改回
国家自然科学基金(60375019)资助课题

2　基于浅层句法信息的翻译实例获取方法

翻译实例的获取过程分为 3 个步骤，总体流程见图 1。首先，分别对源语言句子和目标语言句子进行浅层句法分析，从而分别获得源语言和目标语言的翻译单元；然后，使用词汇对齐工具对双语句对进行词语对齐；最后在词对齐结果的指导下，在源语言和目标语言的翻译单元之间确定映射关系，从而获取翻译等价单元。

2.1 浅层句法分析

在英语浅层句法分析方面，本文采用基于多 Agent 的浅

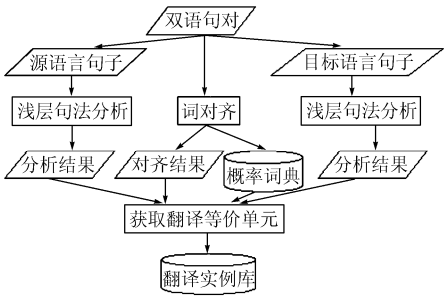

图 1 基于浅层句法信息的翻译实例获取方法

层句法分析技术[8]。在该方法中，按照短语的敏感特征把短语分到不同的 Agent 中，每个 Agent 使用适宜的模型和算法，通过 Agent 间的协调配合完成英语语块识别。该方法充分考虑每种短语的自身特点，同时还避免了对所有短语都使用“词”特征而导致的数据稀疏；对于语块识别中非常重要的基本名词短语的识别，采用了边界统计和词性串校正相结合的英语基本名词短语识别方法。

在汉语浅层句法分析方面，本文采用基于最大熵马尔科夫模型的浅层句法分析技术[9]。在该方法中，语块标记的转移概率(transition probability)通过马尔科夫模型进行估算，而条件概率通过最大熵模型进行估算。

2.2 双语句对的词汇对齐

词汇对齐工具采用了哈尔滨工业大学机器翻译实验室开发的英汉词汇对齐工具[10]，该工具使用 N-gram 统计和语言学知识相结合的解决策略，在降低了汉语分词错误的影响的同时能够发现复合词、新词及术语翻译。在统计翻译词表获取中，使用基于迭代策略的词表抽取算法，有效地解决了间接相关问题，例如：“Sigma-西格玛”，“bus stop-公共汽车站”等，并且原来词典里没有这些词对和短语，经过对齐之后，这些片断被正确地对齐了。

2.3 翻译等价单元的获取

本文在词对齐结果的指导下，进行双语语块的对齐，从而获取翻译等价单元。语块对齐过程可以描述为：给定源语言语块序列 s 和目标语言语块序列 t，$|s|=I,|t|=J$，寻找使概率 $P(a\mid s,t)$ 最大的对齐方式 $\hat{a}$：

$$\hat{a}=\arg\max_a P(a\mid s,t)=\arg\max_a P(a,t\mid s) \qquad (1)$$

设 $a=(m_1,n_1)(m_2,n_2)\cdots(m_k,n_k)$ 为一种语块对齐方式，$\forall i\in[1,k]$，$m_i\in[0,I]$，$n_i\in[0,J]$，$m_i=0$ 和 $n_i=0$ 分别代表对空的源语言语块和目标语言语块。这里近似认为

$$P(a,t\mid s)\approx\prod_k P(m_k,n_k,t\mid s)=\prod_k P(t_{m_k}\mid s_{n_k}) \qquad (2)$$

本文根据词语对齐结果，确定候选的语块对齐方式，从而降低了式(1)的搜索空间。首先，根据词对齐结果为每一个源语言语块 s_i 确定候选对译语块集合。这里规定，如果目标语言语块 t_j 中存在与 s_i 中词语对齐的词语时，t_j 即为 s_i 的候选对译语块。例如，给定如图 2 所示的双语句对，其中，源

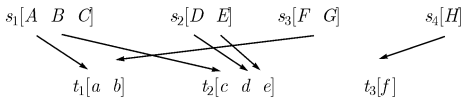

图 2 基于语块的翻译单元实例

语言包含 4 个语块，分别是 s_1，s_2，s_3，s_4，目标语言包含 3 个语块，分别是 t_1，t_2，t_3。根据它们之间的词对齐关系，我们可以建立一张语块对应表，如表 1 所示。

表 1 候选语块对应表

源语言语块	候选对译语块
s_1	t_1，t_2，$t_1\ t_2$，Null
s_2	t_2，Null
s_3	t_1，Null
s_4	t_3，Null

在语块前后相邻的情况下，允许一对多的对齐方式。例如：由于 s_1 有两个候选对译语块 t_1 和 t_2，而且 t_1 和 t_2 相邻，因此，可以将 t_1 与 t_2 合并起来作为 s_1 的候选对译语块。根据表 1，可以列出语块对齐的所有可能方式，如图 3 所示。

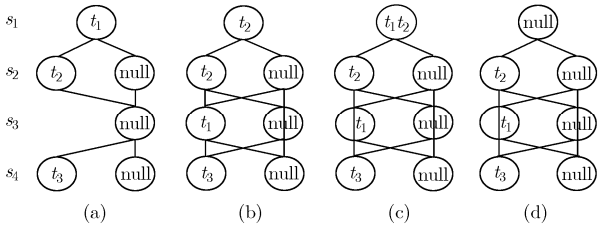

图 3 语块对齐示意图

用文字描述，分别是

a_1: { s_1 - t_1、 s_2 - t_2、 s_3 -null、 s_4 - t_3 }　　(图 3(a))

a_2: { s_1 - t_1、 s_2 - t_2、 s_3 -null、 s_4 -null}　　(图 3(a))

a_3: { s_1 - t_1、 s_2 -null、 s_3 -null、 s_4 - t_3 }　　(图 3(a))

⋮

需要指出的是，在语块前后相邻的情况下，我们还允许多对一和多对多的组合方式。例如在图 3(b)中，由于 s_1 和 s_2 相邻，而且它们具有共同的候选对译语块 t_2，因此，它们可以同时和 t_2 对齐。但是，在图 3(a)中，虽然 t_1 是 s_3 的候选对译语块，但是由于 s_1 已与 t_1 对齐，而 s_3 与 s_1 又不相邻，因此，s_3 不能再与 t_1 对齐。接下来需要考虑式(2)中的概率 $P(t_j\mid s_i)$，采用如下的模型：

$$P(t_j\mid s_i)=\sum_{a_c}P(a_c,t_j\mid s_i)\approx\max_{a_c}P(a_c,t_j\mid s_i)$$
$$=\prod_k\max_l P(t_{j_l}\mid s_{i_k}) \qquad (3)$$

其中 a_c 为语块间的词对齐方式。本文使用公开软件包

GIZA++[1)]训练公式(3)中的词汇翻译概率。GIZA++是一个翻译模型工具，实现了 IBM 公司提出的 5 种复杂程度递增的数学模型[11]，并且加入了一些新的特色。为了充分利用词对齐的信息，对式(3)进一步改进，即对于任何一个词 s_{i_k}，如果存在与其对齐的词语 t_{j_x}，则直接取概率 $P(t_{j_x} \mid s_{i_k})$；如果不存在与其对齐的词语，则取概率 $\max_l P(t_{j_l} \mid s_{i_k})$。

3 实验结果及分析

实验选取了几种较具代表性的翻译实例获取方法作为基准比照系统(baseline)。

(1)基于词对齐信息　该系统采用基于词对齐的方法获取三类翻译实例，分别是原子对译片段、对齐闭包和平行扩展对译片段[4]。

(2)基于 phrase-based SMT　该系统采用基于短语的统计机器翻译中的短语译文获取方法[2]。这里使用公开软件包 PHAROH[2)]中的短语译文训练工具获取短语译文，从而构建翻译实例库。

(3)基于标志词　该系统采用基于标志词对齐的翻译实例获取方法[7]。

实验中的各个系统均从 IWSLT2004 提供的 BTEC (Basic Travel Expression Corpus)训练集中抽取翻译实例，该集合由 23496 个汉英句对构成。本文从两个方面对翻译实例库的获取方法进行评价。一种方法是直接评价翻译实例库本身的译文质量。另一种是通过 EBMT 系统的性能来评价翻译实例库的质量。

3.1 直接评测

通过人工评测的方法，从实例库中随机抽取 100 个翻译实例进行人工打分，打分标准分为优、良、中、差 4 个等级[7]。实验结果如表 2 所示：

表 2 直接评测结果(%)

	基于词对齐	基于 phrase-based SMT	基于标志词	基于浅层句法信息
优	65.9	77.5	57.0	67.5
良	13.2	10.0	11.6	13.0
中	12.1	7.5	17.4	13.0
差	10.0	5.0	14.0	6.5

从直接评测的实验结果可以看出，基于浅层句法信息的翻译实例获取方法性能好于基于词对齐和基于标志词的翻译实例获取方法，但是等级“优”所占的比例低于基于 phrase-based SMT 的翻译实例获取方法。通过分析发现，这主要是因为基于 phrase-based SMT 的方法获取的实例“重叠”现象较为严重。例如由汉英句对“This is my first time diving./这是我的第一次潜水。”得到如下一些翻译实例：这——>This，这是——>This is，这是我的——>This is my，是——>is，是我的——> is my，是我的第一次潜水——>is my first time diving，我的——>my，我的第一次潜水——>my first time diving，第一次潜水——>first time diving。也就是说只要一个较长的实例翻译正确了，与其相近的实例基本上也都是正确的。但是这种方法得到的翻译实例粒度过碎，在实际应用中哪种方法更实用，还有待于通过间接评测来进一步考查。

3.2 间接评测

考虑到构建翻译实例库的最终目的是要应用在 EBMT 系统中，于是很自然的想法，就是通过使用同一个 EBMT 系统所表现出的不同性能来评价不同的实例库的性能。系统的译文质量通过译文自动评价标准 Nist 和 BLEU 分值来衡量。本文采用作者所在研究室开发的一个 EBMT 系统[4]作为实验平台。与大多数 EBMT 系统相似，该系统的翻译过程分为 3 步：首先根据翻译实例库对输入句子进行切分，通过特定的启发式规则选出最优的切分结果；接下来，采用基于对译片段相似度的译文选择模型选出最优译文；最后，根据 N 元模型对实例顺序进行调整，从而生成最终译文。实验过程中，测试语料来自 IWSLT2004 提供的 BTEC(Basic Travel Expression Corpus)测试集，该集合由 500 个汉语句子及其英语译文(506×16)构成。实验结果如表 3 所示。可以看出，基于浅层句法信息方法获取的翻译实例库使得系统译文的 Nist 分值达到了 6.5155，比其它几种方法分别高出了 9.4%，14.7%和 4.9%。这说明浅层句法信息对整个译文质量起到了积极的作用。基于 phrase-based SMT 的方法虽然在直接评测时翻译实例的质量较高，但在间接评测时并没有表现出同样好的效果。前面已经提到，该方法抽取出的翻译实例有许多重叠现象，这种现象相当于将一个较长的翻译实例又进一步分解为多个长度较短的翻译实例。虽然在理论上说这些翻译实例具有较强的灵活性，但在实际应用中，过碎的翻译实例给 EBMT 系统带来了过多的实例组合方式，从而也更容易引入错误的组合方式，导致译文质量下降。就 Bleu 分值来说，基于浅层句法信息的方法对应的系统译文质量好于基于 phrase-based SMT 和基于标志词的方法，但与基于词对齐的方法对应的系统译文性能差不多。事实上，Bleu 侧重于译文的词序，而 NIst 更测重对词汇的选择(通过词汇信息权重)，Zhang 和 Riezler 等人均通过实验验证了 Nist 对译文结果的细微差别更加敏感[12,13]。因此，根据这两种方法(基于浅层句法信息的和基于词对齐的)对应的 Bleu 和 Nist 分值可以看出，虽然两者在译文词序上的性能基本相当，但前者在译文词汇的选择上更加恰当一些。

4 结束语

本文采用了浅层句法信息与概率统计词典相结合的方

1) http://www.isi.edu/~och/GIZA++.html

2) http://www.isi.edu/licensed-sw/pharaoh/

表 3 间接评测结果

	Nist5	Bleu4
基于词对齐	5.9554	0.2099
基于 phrase-based SMT	5.6795	0.1940
基于标志词	6.2083	0.1829
基于浅层句法信息	6.5155	0.2032

法进行翻译实例的获取，并将其应用于 EBMT 系统。通过与几种较具代表性的翻译实例获取方法进行对比实验发现，无论是对翻译实例库直接评测，还是通过 EBMT 系统进行间接评测，该方法都获得了令人满意的效果。

参 考 文 献

[1] Kaji H, Kida Y, and Morimoto Y. Learning translation templates from bilingual texts [C]. Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, 1992: 672–678.

[2] Koehn P. PHARAOH Training Manual. http://www.iccs.informatics.ed.ac.uk/~pkoehn/smt/training-manual.ps, 2006: 3,3.

[3] Simões A and Almeida J J. Combinatory examples extraction for machine translation[C]. In Proceedings of the Tenth Workshop of the European Association for Machine Translation (EAMT-06), Oslo, Norway, June 2006: 27–32.

[4] Yang M, Zhao T, and Liu H, *et al*.. Auto word alignment based Chinese-English EBMT[C]. Proc. of International Workshop on Spoken Language Translation. Kyoto, Japan, 2004: 28–30.

[5] 程洁, 杜利民. EBMT 系统中的多词单元翻译词典获取研究[J]. 中文信息学报, 2004, 18(1): 55–61.

[6] Gough N and Way A. Robust large-scale EBMT with marker-based segmentation [C]. Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation. Maryland,USA. 2004: 95–104.

[7] 刘海洁. 基于标志词的汉英 TBMT 研究[D]. [硕士论文], 哈尔滨工业大学, 2005.

[8] 梁颖红. 基于多 Agent 的英汉文本语块识别技术研究[D]. [博士论文], 哈尔滨工业大学, 2006.

[9] Sun Guang-Lu, Huang Chang-Ning, and Wang Xiao-Long, *et al*.. Chinese chunking based on maximum entropy markov models [J]. *Computational Linguistics and Chinese Language Processing*, 2006, 11(2): 115–136.

[10] 吕雅娟. 基于双语语料库对齐的翻译知识自动获取技术研究[D]. [博士论文], 哈尔滨工业大学, 2003.

[11] Brown P F, Pietra S A D, Pietra V J D, and Mercer R L. The Mathematics of statistical machine translation : Parameter estimation[J]. *Computational Linguistics*, 1993, 19(2): 263–311.

[12] Zhang Ying, Vogel S, and Waibel A. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system [C]? Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004: 2051–2054.

[13] Riezler S and Maxwell III J T. On some pitfalls in automatic evaluation and significance testing for MT [C]. Proceedings of the 43th Annual Meeting on Association for Computational Linguistics, USA, 2005: 57–64.

陈　鄞：女，1978 年生，博士生，研究方向为自然语言处理、机器翻译等.

赵铁军：男，1962 年生，教授，博士生导师，中国中文信息学会理事，研究方向为自然语言处理、机器翻译等.

杨沐昀：男，1971 年生，副教授，研究方向为自然语言处理、机器翻译等.