基于信息熵的候选实例模式检索算法*

张孝飞 陈肇雄 黄河燕 俞旸
中国科学院计算机语言信息工程研究中心 北京 100083

摘要：EBMT 系统通常都需要有一个非常大的实例模式库，如何从中高效地选出一定数量的对后续类比翻译最有帮助的候选实例，是任何实用 EBMT 系统所必须解决的一大难题。文章基于句子的词表层特征和词信息熵提出了一种多层次候选实例模式检索算法。通过在实际系统上的运行测试，结果表明该方法较好的解决了候选实例模式检索这个难题。

关键词：EBMT 实例模式库 候选实例 词表层特征 信息熵

Retrieval Approach of Candidate Translation Examples Based on Entropy

ZHANG xiao-fei CHEN zhao-xiong HUANG he-yan Yu yang
Research Center of Computer & Language Information Engineering,
Chinese Academy of Sciences, Beijing, 10083

Abstract: EBMT system often requires a large corpus of translation examples. So the difficulty how to fast and effectively retrieve an amount of candidate translation examples which are useful for latter translation by analogy reasoning from the corpora must be resolved for any application EBMT system. In this paper, a multi-layer retrieval approach of candidate translation examples is proposed based on word surface features and word entropy. The approach is tested on an application MT system, and the test result show that the approach effectively resolves the problem of the retrieval of candidate translation examples.

Key words: EBMT, corpora of translation example, candidate translation example, word surface features, entropy

1. 引言

*基金项目：国家自然科学基金资助项目（60272088）

基于实例的机器翻译 EBMT（Example-based machine translation）由于其知识获取比较容易、建造成本比较低、译文质量比较高等特点，正越来越受到机器翻译界的重视。基于实例的机器翻译 EBMT 的基本思路是：预先构造由双语对照的翻译单元对组成的语料库，然后翻译过程选择一个搜索和匹配算法，在语料库中寻找最优匹配单元对，最后根据例句的译文构造出当前所翻译单元的译文。一般的，EBMT 系统包括候选实例模式检索、语句相似度计算、双语词对齐和类比译文构造等几个步骤[1,2,3,4]。

通常，实用的 EBMT 系统所需要的实例模式库都非常大。例如卡内基梅隆大学的 PanEBMT 系统，其语料库中包含了 280 多万条双语句对。因此，如何从这样大的一个实例模式库中高效地检索出一定数量的足够相似的候选实例，提供给后面的类比翻译处理模块，是影响 EBMT 系统翻译能否成功的关键因素之一。文献[2]的方法是通过对实例库中所有出现的词汇建立完全索引，然后利用词索引找出所有匹配的翻译实例，再从中选出 5 个最近建立的翻译实例作为候选实例。这种仅仅考虑翻译实例建立时间的方法，很可能会遗漏以前建立的但对后面类比翻译却非常有用的一些候选实例。文献[5]采用多层次约束检索，句法功能词、句法特征、领域知识、语义功能词等约束能力依次加强。这种方法建立在 IMT/EC-863 系统[6]已有的强大的规则分析能力基础之上，具有明显的优势。但一般的 EBMT 翻译系统缺乏强大的规则分析能力，因此这种检索方法实现起来难度较大，对系统的规则分析能力要求较高。

本文在 IHSMT 系统[5]的基础上，提出了一种新的基于词表层特征和词信息熵的多层次候选实例模式检索方法。

2. 检索策略需要考虑的一些问题

候选实例模式检索算法的设计，首先也是最重要的应该是能把最相似的、最有利于后续类比翻译的实例检索出来。因为如果检索不到相似实例或检索出来的实例相似度过低，都会导致后续类比翻译的失败。

其次，检索出来的候选实例不能太少，候选实例太少很容易遗漏一些对后续类比翻译非常有用的实例，导致翻译失败或译文质量不高；候选实例也不能太多，否则会包含一些与输入的待翻译句子不太相似、对后续类比翻译作用不大的实例，导致系统性能下降，很难达到实时文档翻译所要求的速度。根据我们的经验，理想的候选实例应该在 5 个左右。

第三，具体 EBMT 系统的实例模式检索策略需要与后续的处理策略如句子相似度计算、词对齐和类比译文生成等算法通盘考虑，以取得整个系统的最优化。

3. 基于词表层特征的粗选实例模式集

粗选实例模式集中应尽可能多的包括一些对后续类比翻译有用的实例。在处理过程中，

通过把句子表示成单词的集合，并在其上定义句子的词表层特征相似度。

定义 1：句子的词集合表示为

$\pi(S)=\{W_1, W_2, \ldots W_n\}$ ..... (1)

其中 S 表示句子，W_i 为句子中的单词。对于英语，词 W_i 需要事先进行形态还原；对于中文，句子 S 需要事先进行词切分处理。要注意的是 $\pi(S)$ 与通常数学概念上的集合有些不同，即 $\pi(S)$ 中的元素可以有重复。

比如句子 S_1：*If you were a manager would you decentralize authority?*

表示为：$\pi(S_1)=\{$if, you, be, a, manager, will, you, decentralize, authority, ?$\}$

定义 2：句子 S_1 和 句子 S_2 的表层相似度：

$Sim_s(S_1, S_2)= \Gamma(\pi(S_1) \cap \pi(S_2))$ ..... (2)

其中 $\cap$ 表示集合的求交运算。同样要注意的是与严格数学意义上的集合求交运算不同，即求交运算的结果中允许出现两个或多个同样的元素。Γ 运算符表示求集合中的元素个数。

两个句子的表层相似度越大，则输入的待翻译句子与翻译实例相同的单词就越多，后续类比译文构造过程对翻译实例所要做的修改量也就越少。这说明表层相似度的计算方法从总体上是符合系统的要求，即有利于最终生成高质量的译文。比如有下列两个句子：

S_1：*The city is becoming more and more prosperous.*

S_2：*BeiJing is becoming more and more beautiful.*

则我们根据定义 1 和定义 2 有：

$\pi(S_1)=\{$*the, city, be, become, more, and, more1, prosperous, .*$\}$

$\pi(S_2)=\{$*BeiJing, be, become, more, and, more1, beautiful, .*$\}$

$\pi(S_1) \cap \pi(S_2)=\{$*be, become, more, and, more1, .*$\}$

$Sim_s(S_1, S_2)= \Gamma(\pi(S_1) \cap \pi(S_2))$

$= 6$

为了得到粗选实例模式集，将待翻译的句子与库中的实例逐个比较，计算出它们之间的表层相似度；然后按相似度大小降序排列；选择最前面的 20 个翻译实例作为粗选实例模式集。如果总共不到 20 个实例，则全部选作为粗选实例模式集。

例如，输入待翻译句子：*It is important for children to read some simple books that can improve the view of almost whole world.* 则得到如下的粗选实例模式集（注：我们这里使用的语料库总共包括 162，918 条英汉双语句对，大约 200 万词的英文和 200 万词的中文）。

1. *It is essential for child to read some books that can develop his view of future life.*
2. *To read some books that can improve the view of whole life is of consequence to every child.*
3. *It is important for students to do some physical exercises that can improve their health.*
4. *To play some games that can improve the intelligence is of consequence to every child.*
5. *It is important for leaders to adapt his view to some others of contrary.*
6. *For authors, to write the books about beauty of world is important than to about scandal.*

7. *It is not convincible that one book can change one's views of all aspects of life.*
8. *The world will be simple if people can develop his prejudice of others.*
9. *It is not books but games almost every child likes.*
10. *Extensive reading is a simple method to improve our effectiveness of study?*

限于篇幅，这里只列出了前 10 个翻译实例。可以看出，大部分检索出来的翻译实例与待翻译句子比较相似，对后续的类比翻译很有帮助；但也有一些翻译实例与待翻译句子并不太相似，对后续的类比翻译作用不大，这样的翻译实例应该滤去。

系统实现时，如果真是逐个从库中取出实例来计算相似度，效率是非常低的。通过事先对库中出现的每个词汇建立完全索引，索引形式为：

<Word，SID_1*n_1，SID_2*n_2，... ，SID_N*n_N>

比如<book, 20*5, 70*20>，表示单词 book 在句子编号为 20 开始的 5 个连续句子（即编号为 20、21、22、23 和 24 的句子）和编号为 70 开始的 20 个连续句子中都出现。具体编码时，通过链表数据结构来实现，如图 1 所示。这样就可以仅利用索引来进行快速的表层相似度计算。

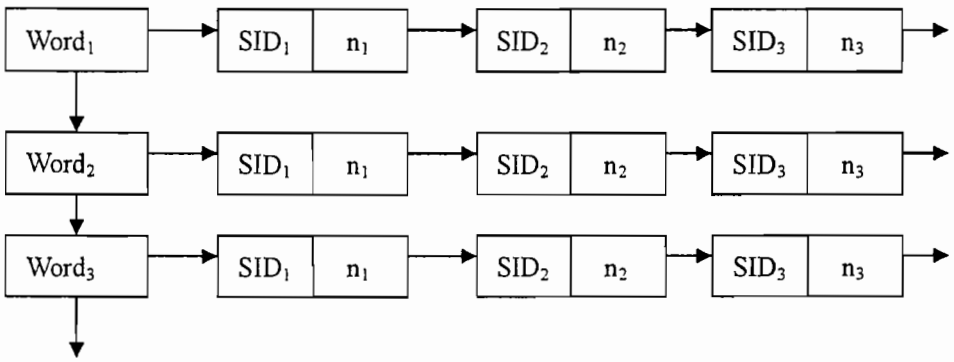

图 1. 词索引结构图

4. 基于信息熵的精选实例模式集

前面得到的粗选实例模式集，仅仅考虑了词表层特征即词形特征，而没有考虑语言本身的任何语法、语义特征和统计特征。其中包括了一些与输入的待翻译句子其实并不太相似，对后面的类比翻译译文构造作用并不大的实例模式。如果将这些实例模式全部提交给系统的后续模块进行处理，将会导致系统效率低下。因此，对于这些实例模式我们得想办法将其滤去。

基于规则的语法、语义特征计算比较复杂、繁琐、费时，如果在检索阶段对所有粗选实例模式都进行语法、语义特征的计算，一方面会严重影响系统性能；另一方面极可能会选择错误的候选模式，因为语法、语义分析本身非常复杂，有许多问题尚没有完全解决。

比如判断一个英文句子是主动态还是被动态，就不是一件容易的事，我们在系统上曾做过试验，准确率只有80%左右。

近年来，统计方法在自然语言处理领域逐步得到了广泛的应用。基于统计的方法可以避免规则方法的许多缺陷，例如，它利用的知识主要是统计数据，可以从语料库中利用有指导或无指导的学习方法得到，从而避免了人工获取规则的繁琐过程。同时，获取的知识具有客观性好、一致性强等特点。但要采用统计的方法，前提条件是研究人员手中得要有个大的语料库。尤其是大型的高质量句对齐双语语料库不容易建立。

我们考虑采用统计的方法，对粗选实例模式集利用简单的统计计算进行进一步的精选。

定义 **3**：词信息熵

$$H(w) = \lg(M / m) \quad \text{.....(3)}$$

其中 w 表示词，M 表示训练集（语料库）中的句子总数，m 表示出现了词 w 的句子数。词的信息熵越大，说明该词在语料库中的出现频度越低，对区分句子的作用也就越大。

定义四：句子 S_1 和 句子 S_2 的信息熵相似度：

$$Sim_H = \sum H(w_i) \quad \text{.....(4)}$$

其中 $w_i \in \{\pi(S_1) \mathrm{I}\ \pi(S_2)\}$，运算符 π 和 I 的含义参见前面定义 1 和定义 2。

两个句子的信息熵相似度越大，则从概率上来讲，输入的待翻译句子与翻译实例在语义上更相似。同时通过信息熵的计算方法，对一些特别常用的词比如{the、a、and、of}等起到了一定的抑制作用。因为这些特别常用的词不仅没什么语义意义，同时由于这些词几乎出现在每一个句子中，其对句子的区分能力也不大。表 1 列出了词频及其信息熵的分布情况。表 2 列出了部分词在语料库中的出现频次及其相应的信息熵值。

表 1. 词频及信息熵分布

词频（次）	1~4	5~50	51~100	101~500	501~1000
信息熵	5.21~4.61	4.51~3.51	3.51~3.21	3.21~2.51	2.51~2.21
词数（个）	27707	11269	1435	1559	209
词率（%）	65.4	26.6	3.4	3.7	0.5
词频（次）	1001~5000	5001~10000	10000~	*词汇总数*	*词次总数*
信息熵	2.21~1.51	1.51~1.21	1.21~	—	—
词数（个）	155	16	19	***42369***	***1993937***
词率（%）	0.4	0.04	0.05	—	—

表 2. 部分单词的出现频次及其信息熵

单词	出现频次	语料库的总句子数	信息熵
The	82365	162918	0.30
Of	49854	162918	0.51
And	36489	162918	0.65
It	16627	162918	0.99
Blend	50	162918	3.51
Vortex	5	162918	4.51
Borax	1	162918	5.21

分析表 1 可知，在近 2 百万词次规模的语料库中，总共包括了 42369 个词汇。其中出现 500 次以上的词汇占总词汇的比率还不到 1%，出现 50 次以下的词汇占了总词汇的 92%。尤其是只出现 1~4 次的词汇竟占了总词汇的 65.4%。这些统计结果也正是我们检索候选实例模式时，粗选时只选择 20 个和精选时选择 5 个候选实例模式的重要依据之一。当然，随着语料库的变化，粗选和精选实例模式的数量要做适当的调整。

从表 2 可以看出，常用词的出现频率比较高，其相应的信息熵比较低；而非常用词的出现频率比较低，其相应的信息熵比较高。同时我们注意到只出现一次的词（如 borax）其信息熵是最常用词{the}的信息熵的 5.21/0.3=16.8 倍。这说明，如果在整个语料库中直接利用信息熵的大小来筛选候选模式，则会给一些非常用词以过大的比重，结果会导致选出来的翻译实例在句子结构上与输入的待翻译句子相差很大，不利于后续的类比译文构造。这也正是我们首先在整个语料库中进行基于词表层特征的粗选，然后再在粗选实例模式集中进行基于信息熵精选的原因。

为了得到精选实例模式集，将待翻译的句子与粗选实例模式集中的翻译实例逐个比较，计算出它们之间的信息熵相似度；然后按相似度大小降序排列；选择最前面的 5 个翻译实例作为精选实例模式集，提交给系统后续的类比翻译模块。

例如，前面的例子（输入的待翻译句子为 *It is important for children to read some simple books that can improve the view of almost whole world.*）经精选后得到如下实例模式集：

1. *It is essential for child to read some books that can develop his view of future life.*
2. *It is important for students to do some physical exercises that can improve their health.*
3. *To read some books that can improve the view of whole life is of consequence to every child.*
4. *It is not convincible that one book can change one's views of all aspects of life.*
5. *It is important for leaders to adapt his view to some others of contrary.*

可以看出，精选后的翻译实例与输入句子都比较相似，对后续的类比翻译和译文构造都非常有帮助。因此，将它们全部提交给系统后面的类比推理翻译模块进行进一步的更复杂的处理。

5. 结束语

高效的候选实例模式检索策略，是任何一个 EBMT 系统都必须解决的一个难题。检索不出一定数目的足够相似的候选实例，结果只会导致 EBMT 系统翻译的失败。本文提出的基于词表面特征和信息熵的多层次检索方法，比较有效地解决了这个难题。我们在 IHSMTS[5]系统上测试本文方法，确能较大幅度的提高系统中 EBMT 部分的翻译覆盖面（即提高了 EBMT 在整个系统中的贡献率）。具体的实验设计和实验统计数据，我们将在其它文章中详细论述。

但由于算法对实例模式库中所有出现过的单词都做了完全索引，以及实行了多层次检索策略，所以算法的时间和空间开销都比较大。我们下一步将设法优化该算法，以提高其效率，使之能达到文档实时翻译的要求。

参考文献

[1] Ying Zhang, Ralf D. Brown, and Robert E. Frederking. Adapting an Example-Based Translation System to Chinese. In Proceedings of HLT 2001: First International Conference on Human Language Technology Research, p. 7-10. San Diego, California, March 18-21, 2001

[2] Ralf D. Brown, Example-Based Machine Translation in the Pangloss System. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), P. 169-174. Copenhagen, Denmark, August 5-9, 1996

[3] H. Maruyama and H. Watanabe. Tree Cover Search Algorithm for Example-Based Translation. In Proceeding of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation （TMI-92）, p. 173-184, Montreal, 1992

[4] Sergei Nirenburg. The Pangloss Mark III Machine Translation System. Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University), Information Sciences Institute (University of Southern California). Issued as CMU technical report CMU-CMT-95-145.

[5] 黄河燕、陈肇雄、宋继平，一种人机互动的多策略机器翻译系统IHSMTS的设计与实现原理，国际机器翻译与计算机语言信息处理会议论文集，中国北京，1999年6月26-28日, p. 270-276,

[6] 陈肇雄，高庆狮，智能化英汉机译系统 IMT/EC，中国科学，A 辑，第 2 期，1989