带有对译信息的惯用表示的收集

田中康仁　　穗志方（译）　俞士汶（校）
（日本兵库大学）　（北京大学计算语言学研究所）

摘要：本文论述了带有对译信息的惯用表示的数据收集、文件建立、惯用表示标准化以及相应译语等问题。进一步，又论述了惯用表示例句库文件的数据收集与检索方法，还分析了作为研究基础的数据的构造及检索方法。

关键词：惯用表示　数据收集　数据检索

The Collection of the Translation-Attached-Conventional Representations

Y.Tanaka　　Sui Zhifang(translator)　Yu Shiwen(proof reader)
(Hyogo University)　(Computational Linguistics Institute of Peking University)

Abstract: In this paper, a series of problems concerning the conventional representations are discussed, including data collecting, file creating, standardizing and translating. Furthermore, several approaches are put forward for data collecting and searching of the example database. Finally, the methods of data creating and searching are analysized, which is the fundation of the research.

Key words: conventional representation , data collecting , data searching

一、引言

目前，已有许多公司在开发机器翻译系统。然而，由于译文的质量还达不到要求，因此，这些产品还未能充分地使用。译文质量不高的原因在于，基于“句子是单词的集合”这一观念，单词词典现在是相当充实了，不过象词与词的搭配这一类信息现在还收集得很少，专用词语也收的不多。另一方面，基于“句子是通过学习相似句子而作成的”这一观念，正在研究根据句子的相似度来进行翻译，并准备将其实用化。然而，句子的集合达到什么程度就可以了？相似度达到多少就认为是相似了？这样的问题也很多。因此，笔者认为对于词与词的结合信息应该给予更多的关注。

机器翻译所需要的知识包括基础知识和高级知识。基础知识包括单词、复合词、专门用语、概念、概念与概念的关系以及惯用表示等。虽然以上所述各方面研究都有进展，但

笔者认为对于带有对译信息的复合词以及惯用表示的大量收集、研究以及体系化却少有成果，这种现状是不能令人满意的。高级知识象：“水在摄氏 0 度时结冰。”和“1995 年日本阪神地区发生大地震，死伤了很多人，倒塌了很多建筑。”。除此之外，笔者认为还必须考虑处在高级知识和基础知识之间的中间位置的知识。例如：谚语等。这里只考虑作为基础知识的带有对译信息的数据。

二、惯用表示

2 · 1 惯用表示的数量分析

就惯用表示（熟语）在多大的程度上使用这一问题，查看了有关的数量分析资料。旺文社出版的《英语熟语 1000 例调查》对于 16 年内日本全国大学入学考试试题中的惯用表示进行了调查，其中，不相同者总数为 14，368 条。按照 16 年内出现的次数列表如下：

1 次 … 8,083 (56%)　2 次 … 1,884 (13%)　3 次 … 926 (6.4%)

4 次 … 532 (3.7%)　5 次… 417 (2.9%)　共计：11，842 (82%)

笔者认为，必须收集到上述数据的 4 ~ 5 倍规模。如果考虑 5 倍情况，则必须收集 7 万个惯用表示。

2 · 2 考虑哪些惯用表示

通过考察收集惯用表示的若干本书的索引，发现惯用表示可有以下的存在形式：

（ i ）动词短语　例如 动词 + 介词的形式　例：look up

（ ii ）介词短语　例：in front of

（ iii ）副词短语　例：as soon as

（ iv ）be 动词 + 动词短语　（ v ）含有 not 的否定形式　（ vi ）其他

2 · 3 英语的介词

通过对英语的介词的调查，发现有三种类型。这里引用首藤训宏、中岛武著《介词 · 惯用语的英语会话》

介词分以下几类：1 · 单纯介词　2 · 复合介词　3 · 分词介词

单纯介词: at, but, by, down, *ere, for, forth, from, in, like, of, off, on, out, over, per, save, through, till, to, up, via, with （ere 是“在…前”的意思，现在仅在诗歌中使用）

复合介词: abaft, aboard, about, above, across, adown, after, against, aloft, along, amid, amidst, among, anent, around, aslant, athwart, before, behind, below, beneath, beside, besides, between, betwixt, beyond, despite, except, inside, into, onto, outside, since, throughout, toward(s), under, underneath, until, unto, upon, withal, within, without

分词介词: *barring, bating, concerning, considering, during, excepting,

*notwithstanding, past, pending, regarding, respecting, saving, touching

(* barring=except "除…之外"。 notwithstanding "尽管…")

这些词中，如果只研究单纯介词，对上述旺文社的资料进行调查也就可以了。

2 · 4 对英语介词的研究

以英日对齐的平行语料库作为分析对象。然而，这种平行语料库很不容易得到。如果能够得到，则以英语为中心考虑。首先建成文中关键词索引 KWIC，如下所示:

英语的上文——介词——英语的下文——日语译文

这样作成的 KWIC 按包含介词的右侧语句排序，介词都集中到一起了。

同时，也了解到按包含介词的左侧排序的情况，这样进行动词短语的研究就比较容易。这里也必须考虑人工调查、编辑调查、自动（半自动）抽取等。

英语中有以下的构造: （ 动词 + 【前置词） + 名词短语】

一种可能为动词与介词结合，组成动词短语。另一种可能为介词与名词结合，成为介词短语。所以存在歧义情况。在这种情况下，需要有人工的辅助。

2 · 5 利用语料库建造 KWIC

使用华尔街日报的光盘建造介词的 KWIC。建成了按这些词前面的词语排序的表和按后面的词语排序的表。

这些 KWIC 对动词短语的分析、介词的分析都很有效。这些介词的 KWIC 的数如下所示:

in 634,857 for 325,876 by 195,710 on 177,050 from 138,031

at 136,823 but 24,325 out 21,088 over 18,299 per 9,542

off 4,355 down 3,877

这些词中， in, for, by, on, from, at 这六个前置词特别重要。

为了抽取惯用表示，也考虑使用对译语料库进行抽取的方法。

还有一种方法就是从词典中抽取惯用表示。

2 · 6 以什么样的词典或书为抽取对象

现在有大厚本的、高质量的惯用语词典。然而，这些书虽然好，但也存在以下问题:

i ）著作权的问题; ii ）书的出版周期长，使用时会有些过时; iii ）因为书的作者不同，对于惯用表示的选择、译语的选择都有所不同。这有好的方面，也有不好的方面。

对于书的引用，常常会引起著作权的纠纷，这一点也要提请注意。

按以下的书或杂志为对象来进行讨论。

把从高中学生所使用的参考书、杂志、指南等实用书籍中抽出的惯用表示（英语）以及译文放入存储器，然后抽出需要的内容。

这样的书，也存在一些问题，如商业主义的趣味性，以实用主义为中心等；也有为了写成一本厚书而带来的问题。然而，对于象英语这样巨大的对象进行完全的研究是很难的。因此，将很多人的知识一点一点地收集，并将其体系化，这种做法是可取的。

这里，不抽取例句。例句从用其他方法建成的语料库中抽取，可用于对译语进行验证。

三、惯用表示的收集

3 · 1 以什么样的形式来输入

输入形式如下:

1) No.编号; 2) 英语惯用短语; 3) 日译（可能有几种译法）; 4) 书籍编号; 5) 页号

采用反斜杠来将这些项隔开。实例如下:

例 1 :
\ 1 \ 05515
\ 2 \ be unable to [A]
\ 3 \ [A]することができない
\ 4 \ F
\ 5 \ 12

例 2:
\ 1 \ 05501
\ 2 \ call at
\ 3 \ 访ねる
\ 4 \ F
\ 5 \ 10

3 · 2 输入数据的数量

从 8 本书中抽出惯用表示并进行了输入、整理。对于一个英语的惯用表示，从有多少种译语的角度进行调查，结果如下:

1 种… 6,170　2 种… 2,888　3 种… 478　4 种… 88　5 种… 21　6 种… 6　合计: 9,653

正在对以上的内容加以整理、研讨。现在，仍有 6 本书的惯用表示需要追加，最终变成约 2 万的数据。

3 · 3 惯用表示的标准化

惯用表示在从句子中抽出后，必须对其进行如下的标准化:

1) 动词的过去形变为原形　was able to → be able to; is, was, were … → be
2) 所有格的问题　in his way home → in one s way home
3) 反身代词→ oneself　dried himself off → dry oneself off
4) 复数形式变为单数形式　soft contact lenses → soft contact lense
5) The 变为 a ,特殊的保留不变　例, The Japanese → a Japanese
6) 缩约形式复原　don t → do not; I ve → I have
7) 对于分离词组，中间按顺序插入[A] [B]等符号　launch [A]in [B]
8) 动词 + ing →动词　driving wasps off → drive wasp off
9) 其他

对以上所述的一般的标准化规则总有许多例外。对于例外情况，必须进行个别处理。

3·4 惯用表示和多种译语

对于一个惯用表示可以有多个译语。例: best season 最好的季节; 好季节

然而，实际上使用以下的译语: The cherry is best season. 樱花盛开。

因此，惯用表示的译语应采用有代表性的译语。

然而，实际上代表性的译语（也存在若干种情况）在多大程度上使用？不同的译语（表记稍有不同，意义也有不同）如何使用，使用条件是什么？这些也必须加以调查。

英语的动词短语与日语的复合动词也有相似的方面。因此，动词短语有歧义性。为此，应明确使用条件。

四、含有惯用表示的例句的收集

考虑在什么程度上收集含有惯用表示的例句。

· 收集惯用表示约 5 万个

· 对于一个惯用表示平均收集 5 个例句，5 万 × 5 = 25 万（例句）

所以有必要收集约 25 万个带有译文的例句。这是一个必须做的非常庞大的工作。

五、对于惯用表示例句的检索

在专门讲惯用表示的书中，惯用表示和它的例句的对应是平常的事。在这里，笔者提出用另外一种文件形式来进行管理。

带有对译信息的例句文件，除了惯用表示以外，还可以广泛服务于以下目的:

（1）机器翻译中作为测试数据使用;

（2）抽出句型例句;

（3）利用例句作为开发翻译系统的基本资料;

（4）作为惯用表示的多种译语的材料;

（5）其他等。

然而，带有对译信息的数据年年增加，所以例句和惯用表示文件最好放人不同的文件。

惯用表示文件和惯用表示例句文件的关系如图 1 所示。

惯用表示的抽取程序前面已有论述，这里就不再涉及。

例句检索程序从已标准化的惯用表示文件中检索与实际使用的表示相符合的形式。实际中使用的变形表示的例子有: 动词的各种变形; be 动词有 is，was，are，were ...；单、复数形的变化等。

例句检索程序、惯用表示抽取程序必须通过以下的表和词典来实现:

（1）代名词的变化表;（2）不规则动词的变化表;（3）词典 名词的单、复数形式; 规则动词的词干、词尾的变化。

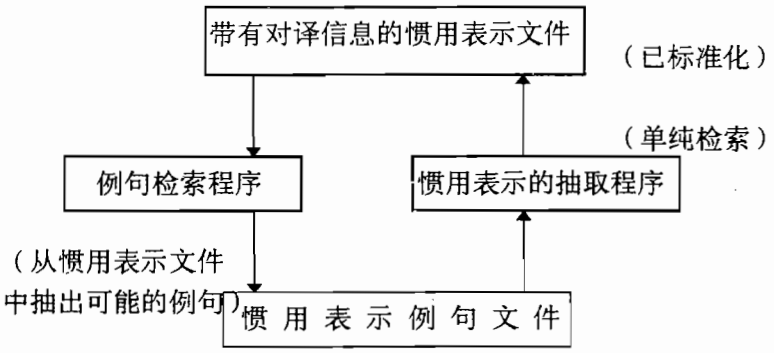

图 1

这些程序在抽取、检索惯用表示的同时统计例句的使用频度。

例句检索程序抽出英语的表示，同时研究日语的多个译语一致的情况有多少，完全不同的情况又有多少。另外，要分析不同的译语产生的原因。

尽管有经过了标准化的惯用表示例句文件，但仍然可能存在检索不到的情况。对于这些，应调查单词（原形）的使用频度，检索用语的原形或标准形的变化。通过惯用表示例句文件的组成单词的原形或标准形的逻辑运算来检索，监视遗漏的情况，并随时修改程序。

六、惯用表示用例句文件

惯用表示用例句文件是日英对应的，其中最好使用其惯用表示。如果不使用惯用表示，仅作为一般文件来用，也没有什么问题。 包括以下项目：

1）No. 编号； 2）日语句子； 3）英语； 4）参考文献（记号）； 5）页码； 6）检索关键词

例 1：\ 1 \ 0001
\ 2 \ 彼女は君の野心に轻蔑しているぜ。
\ 3 \ She is contemptuous of your ambition .
\ 4 \ B 000
\ 5 \ 27

此外，应考虑使用计算机处理或者有一部分手工编制，将英语惯用表示的检索关键词作为一个项目。这样可以提高计算机处理的效率。

这样，就可以方便地加注反斜杠区分记号。

七、惯用表示例句文件的建造方法

1）人工输入：一人一年输入 5000 句， 10 人一年可输入 5 万句。因此，经过 5 年可以建成约 25 万例句的语料库。

2）CD - ROM 或电子书籍（数字书籍）的利用： CD - ROM 化的英日 · 日英词典的重新编辑采用的是同一种方法。

另外，最近电子书籍可以很便宜地买到，它可被个人计算机阅读。这种软件也作为电子书籍在出售。

可将电子书籍或数字书籍的内容作为文本文件使用的软件在日本售价为 3,980 日元。同免费软件相比，可以得到更多的用于研究与翻译的资料。

SONY 还出售一种可以独立使用电子书籍的便携机器，不需要另外的软件。

另外，电子书籍可在日本的大书店买到。这些书店正通过各种流通途径出售，约有 300 份，今后还会扩大。

3 ）通过 OCR 扫描书再进行编辑：考虑将一本书或杂志通过 OCR 扫描后再进行编辑。

以前， OCR 的精度，文字识别的精度以及订正的方法还不太好，现在，这些技术已有提高，从总体上可达到实用水平。

但是，必须注意著作权的问题。供个人研究使用或作为测试数据使用是可以的，但作为商品开发就会有问题。

八、结束语

以上已讨论了收集惯用表示的一些基本问题。今后，应考虑应用时所遇到的问题以及译语选择问题。

进行语言研究，要充实包含惯用表示和惯用表示例句的文件，以此作为基础资料，稳步前进。为了在科学上、工程上处理语言，数据的准备是重要的。

在思考带有对译信息的惯用表示的问题的同时，笔者已开始收集数据。希望这些数据今后对于提高机器翻译系统的质量能发挥更大的作用。

我相信并愿意实践这样一句哲学上的话：“量变引起质变”。

参考文献

1 ）花木金吾 《英语熟语 1000 例调查》 旺文社 ISBN 4 - 01 - 030754 - 4

2 ）首藤训宏、中岛武《介词 · 惯用语的英语会话》 ISBN 4 - 478 - 98011

3 ）田中康仁、吉田将 惯用表示的收集与整理 情报学基础 5 - 1 情报处理学会 1987 · 6

4 ）田中康仁、吉田将 概念词典的作成 自然言语处理 75 - 12 情报处理学会 1990 · 1