意义群落测定在知网机译系统中的应用

董强，董振东，郝长伶
HowNet Technologies Inc.
Email:{dongqiang, dzd, support}@keenage.com

摘 要： 本文是《知网机译系统中的语义计算》的姊妹篇。本文提出和讨论了一种与被称为与意义群落相关的歧义的消除的方法和技术，以及它在机器翻译系统中的应用。该方法和技术被称为意义群落测定（Sense Colony Testing，SCT）。该方法是基于知网的，是英中双语的。它处理的文本可以是超句的。它可以独立运行，并作为其他类型的语言处理系统的应用工具。本文概要地说明了意群测定的原理， 介绍了它的组成部分以及实现方法，以及机器翻译系统应用此工具的适宜时机。在最后一节的讨论中，作者探讨了 SCT 所要涉及的特征的深度和精细度。

关键词： 自然语言处理，知网，机器翻译，意义群落测定，词义消歧，消除歧义

Sense Colony Testing in HowNet MT System

Qiang Dong, Zhendong Dong, Changling Hao
HowNet Technologies Inc.
Email:{dongqiang, dzd, support}@keenage.com

Abstract: *This paper is a companion piece of "Semantic computing in HowNet MT system". In terms of the relatedness to syntax, the ambiguity can be roughly divided into two types: syntax-related and non- syntax-related. The latter is closely associated with the discourse and topics. This paper proposes and discusses an original method and technique for solving the latter type of ambiguity, which is called sense-colony-related type. The technique is named Sense Colony Testing, SCT, which is wholly based on HowNet and can be operated bilingually, and can process text-level discourse. The paper describes SCT's linguistic principle and technical mechanism, and depicts the SCT as a tool used in HowNet machine translation, Lastly in the section of discussion the paper demonstrates the fineness of features that a NLP system have to employ..*

Keyword: *NLP, HowNet, machine translation, sense colony testing, WSD, disambiguation*

1．引言

30 多年前，在研发我们最初的机译系统的时候，就一直为这样一个句子里的一种歧义的消解而苦恼："When I got to the bank, I found I didn't bring the fishing rods. " 句中的"bank"，在我们具有基本常识的人们看起来是很容易判定的。想着要带上钓鱼竿去的地方，自然应该只能是"岸边"，而几乎不可能是"银行"，这是常识。当时研发的机译系统是基于规则的。如果就在全句范围内查找有无"fishiing rod"，当然是可以过得去了。但是如果遇到的不是"fishiing rod"，而是"bait"呢？是不是再写一条规则呢？另外，如果它们不处在当句呢，例如"We lost our way before we got to the bank at noon. When I unpacked my bag I found I didn't bring the fishing rod."那时候我开始想到机译需要一个常识知识库的支持。这就是研发知网的萌发点。我们想如果有一个常识库，就有可能大大提高机译译文质量。岂不知这一步竟然耗费了我们近 30 年的时间。真是既费时又费力（time-consuming, labour-intensive）！我们太固执

了，我们没有因此而改弦更张。因为我们想看一个究竟，看看是否是像我们所预想的。因此我们在知网已经建成之后，我们再重新开始了新的、基于知网的英-中机译系统的研发。如今机译系统也已基本完成了。这一新的系统仍然是英-中的，仍然是基于规则的，但是它与我们30年前的系统不一样了，它有知网这个常识库的支持了。它的译文质量、它的调试与完善的便利都是我们的30年前的系统所无法相比的。本文主要是介绍该系统所采用的、基于知网的一个消歧工具 - 意义群落测定器。 就是该工具提供了解决30年前困扰我们的那种歧义的方法。本文接下去的安排是：第2节说明什么是与意义群落相关的歧义；第3节讨论意义群落测定的认知理论基础 - 简单联想和复杂联想；第4节是本文的中心，介绍意义群落的原理和方法，意义群落系统的构成，以及如何运用于机器翻译系统；第5节是讨论，涉及知识库建设、语言处理中“特征（features)”的深度等。

2. 与意义群落相关的歧义

我们这里所要论述的是词义（word sense）的歧义，不涉及句法结构关系的歧义。从与文本的句法关系上来观察，词义歧义可以分为两大类：即与句法相关的词义歧义（syntax-related ambiguity of words）和与意义群落相关的词义歧义（sense-colony-related ambiguity of words）。前者很容易观察，例如它们是因为词性的不同而发生的歧义如英文的″book″，有“书”（名词）和“订购”（动词）的歧义。 再例如它们是由于在句法关系上与不同的词语相关而发生的歧义如“bright”，同为形容词时，被修饰词语如果是“人”，则是“聪明”（bright children），而被修饰词语是“物”，则为“明亮”（bright room）。这类表面看上去很是依赖语义，但是它还是更加受句法上共现所制约。这类歧义在知网机译系统中，主要靠个性规则来解决。本文侧重于讨论后一类歧义。句法无关歧义是不依赖确定的句法关系（如词性、修饰关系或管辖关系的制约）的歧义，通常它们是依赖于更宽泛的语境的。除了在“前言”中举出的例子外，我们试看下面一段真实文本例子：

例 1： Duchess of Cambridge in labor

The Duchess arrived at the hospital around 6 a.m. London time at a back entrance, bypassing a huge group of media which had been staking out the front door for weeks. It is not clear if the Duchess has been induced. Royal sources have said the Duchess has planned a natural birth with Prince William to be at her side. The baby will be delivered by Marcus Setchell, the Queen's former gynaecologist.

这段文本中有三个词“labor”，“induce”和“deliver”都应归于典型的句法无关歧义，它们的词义都不是严格地受特定的句法关系制约的。而且它们是在一个超句的语境里。它们的解决只能指望常识库的支持，同时要创建一种是系统具有利用常识库的机制，来使机器具有某种利用、选择、判断的智能。因为任何基于规则的方法已无能为力了。让我们再举一个中文的例子:

例 2： 初二年级不仅课程任务最重，而且在很大程度上决定学生能否考入理想的高中。基础打好了，可以弥补初一学习的不足，更是初三强有力冲刺的保障。

此例中的“初二”， 也属于典型的句法无关歧义。虽然它可以处于某些固定的短语中而获得确定的意义，例如：“上初二”、“年初二”、“念初二”、“初二那天”、“初二那年”、“初二早上”、“大年初二”、“初二学生”、“初二第一学期”、“初二三班”等等。在知网的词典里，

对几乎所有的多义词的义项都会给出尽可能多的例子，其目的就是为了排除歧义。据我们考察现有的流行的机译系统，主要是依靠大规模短语对齐语料来解决此类歧义。应该说采用固定短语或习惯用语，来处理歧义不失为一种很好的方法。但是，短语毕竟难以穷尽。另一个方法就是依靠义项的出现频率。可是如果面对的歧义没有被对齐的短语涵盖，同时又并非是那个高频义项如“bank”的“岸边”、“crane”的“鹤”，“body”的“尸体”等，就不易解决了。句法无关歧义是一种大语境性的歧义，多半是与背景知识相关的。可以设想，如果一个小孩只有母亲领着去银行的经验，而从来不知道关于钓鱼的事情，那么当她遇到前面我们举出的例子，她也会把“bank”错当“银行”的。从研究的角度，从探索人的认知与举一反三的能力出发，我们还是要解决这个大语境的问题。我们认识到：（1）必须建立一个科学的、有效的具有语言处理能力的常识性知识库；（2）处理必须深入到概念，而不只是词语表记；（3）统计比较建立在当下的语境里，而不是仅仅对于词语的在全语言中的出现频度的依赖；（4）观察与处理必须是超句的广语境范围的，而不是仅仅限于一个句子的局部范围的。据我们长期的考察，目前流行的机译系统解决词义歧义的主要手段是：(1) 对齐的双语语料，(2) 依靠全语言范围的词频。它们的词义歧义排除失败主要是由于这两个手段的落空造成的。我们试看一个摘自 Wikipedia 的例子，按理似乎比我们上面举出的例子更简单一些：

> “Bank fishing is fishing from river banks and shorelines. People typically do this by casting fishing bait or lures into the water in an attempt to catch fish. Bank fishing is usually performed with a rod and reel but nets , traps , and spears can also be used. People who fish from a boat can sometimes access more areas in prime locations with greater ease than bank fishermen. However many people don't own boats and find fishing from the bank has its own satisfying advantages. Bank fishing has its own requirements, and many things come into play for success, such as local knowledge, water depth, bank structure, location, time of day, and the type of bait and lures.”

流行机译系统的 2014-08-04 的译文：

> “银行钓鱼是从河岸和海岸钓鱼。人们通常这样做铸造渔饵，引诱入水，企图钓到鱼。通常执行与杆和卷轴银行钓鱼，但网，陷阱，和长矛也可以使用。人们从一船鱼谁有时可以访问更多的地区的黄金地段更容易比银行渔民。 然而，许多人没有自己的船，发现从银行钓鱼有自己满意的优势。 银行钓鱼都有自己的要求，许多事情开始发挥作用的成功，如当地知识，水深，银行结构，位置，时间，和鱼饵和诱惑的类型。”

然而，就是这个例子早在 2008 年我们就拿来测试过流行的机译系统， 以后我们每年都继续跟踪，可惜翻译结果没有什么实质的改进，“bank fishing”始终被译为“银行捕鱼”。奇怪的是不止一个的流行系统，也多年来一直是“银行捕鱼”!

3. 简单联想和复杂联想

上面“bank”的例子，给我们提出了一个问题：即人是怎样使“岸”与“捕鱼”联系起来的呢？为什么当我们遇到“捕鱼”这一概念时，就会确定“河岸”会与之共现呢？从“岸”联想到“水域”，似乎比较容易理解和接受，而“岸”，因为“水域”，也许可以联想到“水”，也还比较容易接受。但是由“岸”可以联想到“鱼”，或进一步联想到“捕鱼”，以至“渔具”，似乎太远、太复杂了。完成这样的远距离联想，肯定是需要某种习得机制了。我们认为不同人的个体的联想能力是不同的，不是本能性的，而是需要学习的，这就是经验和知识的获得与积累。我们又应该如何做能让计算机也习得这些联想能力呢？

知网把可以直接由概念的定义（DEF）直接生成的联想，称作“简单联想”，它们的生成工具叫“概念相关计算器（Concept Relevance Calculator, CRC）”。知网把需要事前由人经过推理指导而生成的联想，称作“复杂联想”，它们的生成工具叫“推理机（Inference Machine）”。它们都是知网知识系统的重要的组成部分。无论是概念相关计算器还是推理机都是基于概念的，而不是基于词语表记的，因此都是可以独立于特定语言的，也即是完全语义的。推理机生成的数据是本文将要介绍的意义群落测定的主要基础。

4. 意义群落测定

4.1 意义群落测定的概念与原理

在生物学领域，有一个概念叫菌落（colony of bacteria）。不同细菌在一定的培养条件下形成的菌落具有一定的特征（morphology），包括菌落的大小、形状、光泽、颜色、硬度、透明度等等。菌落特征对菌种鉴别是有意义的。

词语包括词（words）和词组（multiword expressions, MWEs）所含有的意义或者所指代的概念，也有类似于细菌的菌落的现象，不同的概念在特定的语境中表现出它们的特有的形态。

这些特有形态的共性使它们形成一个特定的区别于其他概念的意义群落（sense colony）或简称意群。同一词语具有的不同的概念或意义在特定语境中所形成的意群的规模可能不同，也即所包含进意群的概念的数目不同，这样将会给出某一概念的相关性的计数。同时也可以体现对与之相关的最近的概念所处位置的距离。于是我们就可以根据相关性计数（Relevancy Count）和距离值（Distance Value）来确定这一被测定的概念的所属，这就是我们所谓的意义群落测定（sense colony testing）。意义群落测定的工具，我们称之为意义群落测定器（sense colony tester）。试看以下两例：

英文： He sat on the bank of the Yellow River and watched the current.

在我们把要测定的文本输入知网意义群落测定器后，测定器显示了该文本经由知网词法处理器断词后的每一个词语的每一个义项，同时也显示它们的相关性计数和距离值，例如：

```
bank          （岸）    ：  Count=0002  Value=0.2428571429
              （银行）  ：  Count=0000  Value=0.0000000000
Yellow River  （水域）  ：  Count=0002  Value=0.2750000000
current       （水流）  ：  Count=0002  Value=0.1178571429
              （电流）  ：  Count=0000  Value=0.0000000000
```

	bank (岸)	bank(银行)	Y-River（黄河）	current（水流）	current（电流）
bank (岸)			1	1	0
bank(银行)			0	0	0
Y-River（黄河）	1	0		1	0
current（水流）	1	0	1		
current（电流）	0	0	0		
	2	**0**	**2**	**2**	**0**

表 1

表 1 显示“bank”的两个概念、“Yellow River”一个概念和“current”两个概念分别所形成的意义群落，以及它们各自所得到的支持的相关性计数。

有鉴于此，知网翻译系统将此句英文翻译为：

“他坐在黄河的**岸边**，同时观看**水流**。”

假如我们把句子改为

英文： He sat on the bank near the power station and tried to measure the current

看上面相关的计数和距离值将如何变化？

```
bank          （岸）     ：  Count=0001  Value=0.0333333333
              （银行）   ：  Count=0000  Value=0.0000000000
power station （电站）   ：  Count=0001  Value=0.0500000000
current       （电流）   ：  Count=0001  Value=0.0500000000
              （水流）   ：  Count=0001  Value=0.0333333333
```

	bank (岸)	bank(银行)	P-station（电站）	current（水流）	current（电流）
bank (岸)			0	1	0
bank(银行)			0	0	0
P-station（电站）	0	0		0	1
current（水流）	1	0	0		
current（电流）	0	0	1		
	1	**0**	**1**	**1**	**1**

表 2

表 2 显示“bank”的两个概念、“power station”一个概念和“current”两个概念分别所形成的意义群落，以及它们各自所得到的支持的相关性计数。相关词语的概念的相关性计数和词语距离值都有了相应的变化。 “current”一词的两个概念分别与“bank”的“岸”概念以及“power station”的“电站”概念构成了意群，分别得到的相关性计数都是 1。 “current”的“电流”义项与“power station”构成了意群并且它们之间的距离值是 0.0500000000，而同样的这个“current”的另一个义项“水流”是与“bank”构成了意群，它们之间的距离远，其值为 0.0333333333。 SCT 在最后判定和选择义项的依据是：相关性计数大的优于相关性计数小的；如果相关性计数相等，则优选距离值大的。 因此改动后的句子的知网翻译的译文则是：

“他坐在这岸边靠近发电站，同时尝试测量电流。”

下面再来看一个中文的例子：

“11 月 18 日晚 7：00，初二第一次学生家长会在各班教室举行，各班班主任经过认真细致的准备，在精心布置过的教室里迎接各位家长的到来。各班家长带着希望的心情前来参加这次家长会，教室里济济一堂，座无虚席。” SCT 给出的测定结果显示：

```
初二          （教育等级）  ：Count=0011  Value=0.8429159536
              （日期）      ：Count=0007  Value=0.1441666667
```

该测定文本中，支持“初二”的“教育等级”义项的共有 11 个概念，其代表的词语应是：“学生”(1 次)，“家长会”(2 次)，“班”(3 次)，“教室”(3 次)，“班主任”(1 次)，“前来”等。而支持“初二”的“日期”义项的共有 7 个概念，其代表的词语是：“月”(其中的 2 个概念)，“日”(其中的 3 个概念)，“晚”(其中的 2 个概念)。

这里我们要特别再次提醒：知网所研究和利用的对象是概念或是义项，而不是词语。对于知

网，词语只是进入概念计算的媒介。

4.2 意义群落测定器的构成

意义群落测定器由下列部件构成，它们各自的功能以及它们之间的关系如下，参见图 1：
（1）知网基础数据库 （HowNet Dictionary） - 它提供意义计算的基本数据，包括词语代表的概念的定义（DEF），以及分类层级体系（taxonomies）。
（2）知网公理规则库 （HowNet Axiomatic Ruls Base）- 该库中的规则是由开发者编写的，这些规则表示了各种概念之间可能的公理性规则。规则有两类：共性规则，例如：

~Process_Attribute　　0000$@extract_def(cd,c0,ExtractMode_Class)
@search(c0,SearchMode_OfValue)!

此规则的意义是：这是关于属性的规则：它要求提取与输入的这一属性的概念相对应的属性值，例如：“速度” - “快”、“高速”、“一溜烟”、“慢”、“低速度”、“老牛破车”等等。以及个性规则，例如：

disease|疾病　　0000$c0[def={Safety|安危}]@search(c0,SearchMode_First)
@search(c0,SearchMode_OfValue)!

此规则的意义是：这是一条个性规则，如果 “概念”是“疾病”，要求提取属性“安危”以及与之相对应的属性值，例如：“平安”、“无恙”、“万无一失”、“危险”、“高风险”等等。
（3）推理机 （HowNet Inference Machine） - 在公理规则库的基础上建立起基于概念的意义群落，因此推理机也被认为是一个意义群落的激活器或生产器。
（4）词法处理器 （HowNet Word Processor） - 包含中英两种语言的词法处理器，对输入文本做断词。
（5）意义群落测定器 （Sense Colony Tester） - 根据输入文本的每一个词语计算其所属各概念的意义群落，比较、统计、计算获取“相关性计数”和“相关词语距离值”，进而判断优先度和进行消歧。

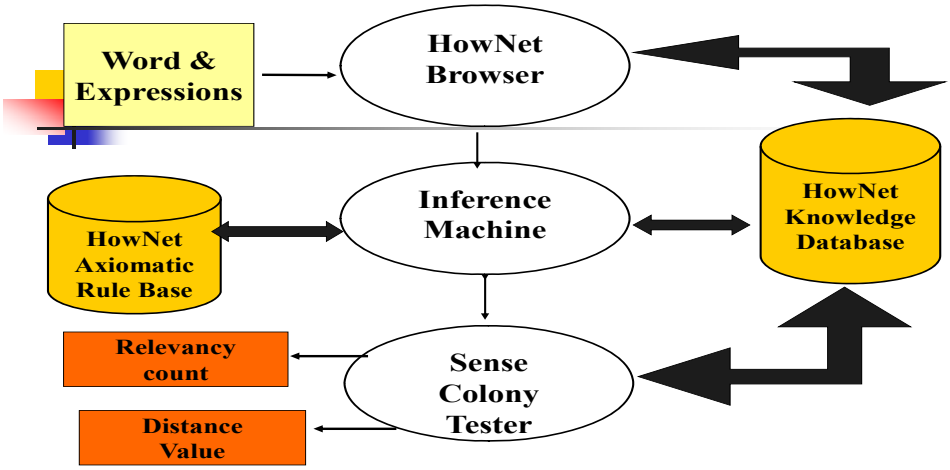

图 1

4.3 意义群落测定应用于机译系统

意义群落测定器是作为一个独立的工具嵌入知网机译系统的，用户可以通过开关来选择开启或关闭。在机译系统中工作顺序是：查词典（包括断词）- 意群测定 - 规则匹配（也即分析、转换、生成等）。我们的实验表明，如果把加工顺序变换为规则匹配先于意群测定，是不可以的。很明显规则的强制性、约束性高于现有的意群测定。如果把意义群落测定器嵌入其他类型的机译系统，同样也必须注意其介入的时机。最重要的原则是：把强制性处理（例如词性判定）的置于其后，要允许其后的更强制性的处理有最后决定权。

在知网机译系统中，意义群落测定器是由用户可选的。我们无数次的实验证明，采用意义群落测定（SCT）的翻译译文质量高于不采用的，另一方面，采用后翻译质量反而被破坏的情形也是有的，但多数会在句法-语义分析之后加以矫正的。 采用 SCT 对于知网机译系统基本上是加分的。试看下例，译文 1 是不采用 SCT 的，译文 2 是采用 SCT 的：

原文：The courts of Chinese Emperors were among the largest and most complex of all, with the Manchu dynasty occupying the whole Forbidden City and other parts of Beijing. In the late first millennium, the Japanese developed an exquisitely refined court that played an extremely important role in their culture.

译文 1：中国皇帝的法庭属于所有之中最大和最复杂，随着满族朝代占领整个紫禁城和北京的其他部分。在第一个千年晚些时候，日本人发展了在他们的文化中扮演一个端重要的角色的一个精巧不俗的法庭。

译文 2：中国皇帝的宫廷属于所有之中最大和最复杂，随着满族朝代占领全紫禁城和北京的其他部分。在第一个千年晚些时候，日本人发展了在他们的文化中扮演一个极其重要的角色的精巧典雅的宫廷。

5. 讨论

5.1 完全基于语义

知网的意义群落测定器是一个纯粹的语义计算的工具，与特定语言的句法无关。它可以完全独立于特定的语言。只要知网词典所涵盖了的语言它都可以同样的进行处理。它对意义的计算是基于概念，而不是基于词语表记的。在完善、扩充和升级时，不需要任何形式的对数据的标记。所要做的工作有两件：（1）修订、完善和增加知网词典的内容；（2）扩充、修订知网公理规则库的内容，特别是个性规则。这些个性规则是常识性、公知性的。通常我们在扩充规则时，是按照不同的主题，一些关键的概念，以及不同的领域等专题进行的。表 1 给出了由推理机激活的部分专题意义群落的词语分布统计。

尸 体	3218	身 体	2966	战 争	5409	打 仗	5288	生 病	2421
购 物	2765	旅 游	2385	股 票	1085	银 行	4266	餐 饮	5287
犯 罪	3616	罪 犯	4117	房 屋	1138	建 筑	5 3 6	家 庭	1023
灾 害	1526	地 震	2551	海 啸	1422	空 难	6837	大 学	1991

表 1

实际上，知网中的任何一个英文或中文的词语所代表的任何一个概念，都有它自己可能被激活的意义群落。同时意义群落是动态的，它将随着概念数据和推理机的规则的变化而变化。

5.2 处理是文本级的

知网的意义群落测定的处理范围是超句、文本级的。通常文本越完整，上下文语境越全面，测定表现的精确性越高。请看下例，这是一段完整的新闻报道摘抄，其中包括一句标题，四个完整的句子。

"Boston bombing suspect's widow wants **body** released
The widow of one of the Boston Marathon bombing suspects will ask the Massachusetts medical examiner to release his **body** to his family, her attorney said Tuesday. Attorney Amato DeLuca said in a statement that Katherine Russell wants Tamerlan Tsarnaev's remains released to the Tsarnaev family. Tsarnaev, 26, died after a gunfight with authorities. Police said he ran out of ammunition before his brother, 19-year-old Dzhokhar Tsarnaev, dragged his **body** under a vehicle while fleeing the scene."

当我们对全文做 SCT，3 句中的全部 3 个"body"的结果是：

"body"（第 1 句） （尸体） ： Count=0005 Value=0.0710754741
"body"（第 1 句） （身体） ： Count=0002 Value=0.0043478261
"body"（第 2 句） （尸体） ： Count=0005 Value=0.1194814815
"body"（第 2 句） （身体） ： Count=0002 Value=0.0080000000
"body"（第 3 句） （尸体） ： Count=0005 Value=0.0132791952
"body"（第 3 句） （身体） ： Count=0002 Value=0.0054054054

5.3 语言处理的特征挖掘深度

英语的"body"是个多义词，基本的常用的词义包括"身体"和"尸体"，前者在 WordNet 语料库出现次数为 536，后者为 266。它们的语义分类在 WordNet 中是完全相同的，即它们有着完全相同的各层级上位，因此它们在 wordNet 语义分类上是无法区分的。这在知网中也是一样，它们的语义主类都是"部件"。 但是知网还有一个概念定义使其区分开来了，同时知网的概念定义是结构化的，是面向计算机的，例如：

"身体"：DEF={part|部件:PartPosition={body|身},domain={physiology|生理学}，
whole={AnimalHuman|动物}}
"尸体"： DEF= {part|部件:PartPosition={body|身},domain={physiology|生理学},
whole={AnimalHuman|动物:{die|死:experiencer={~}}}}

也就是说知网是用"特征"而不仅仅是用"类"，将"body"（中文的"身体"和"尸体"）的两个义项区别开来了。有"die|死"这一特征的形成了独特的、区别于无"die|死"的语义群落，从而达到了排除歧义的功效。由此，人们也许可以体会到我们一贯强调的："分类宜粗不宜细，但属性的描述则宜细不宜粗"的原则了。另外，如此深度，如此细微的特征，可不同于猫的形象特征也不像马克杯的形状特征，如果也真的能通过无监督的神经网络而自动学到，那可是太神奇了！

6. 结论与未来的工作

前文我们说过机译系统所面临的歧义是多种多样的。要注意善用不同的技术处理不同类型的问题。像 SCT 解决的歧义是文本级的、与文本内容、主题、领域相关的。那些与句法关系相关的或与搭配和习语性的，则应该利用另外的机制和技术来解决。

基于知网的意义群落测定是一个自足的、独立的工具。本质上它自身就是一个文本内容的检

测系统。它不仅详细扫描文本，而且它还对文本进行了意义分析。我们把这样的检测称作为文本的计算机断层扫描，因为它把文本所有可能的意义都揭示出来了，并且进行了意义群落的分析。 因此除排歧以外，它还可以应用于各种语言信息处理系统，例如文本聚类，文本分类，文本内容比较、查重，文本主题抽取等等。

未来的研究与开发重点是不断通过机译系统的测试来发现意义群落测定系统所依赖的意义群落的不平衡性，例如我们已经发现并已调整的某些内容，其中最突出的是“人”这一概念所涉及的特征太宽泛等。

感谢

我们在本篇基本完稿后，曾寄送李维博士，请他过目。李维博士在百忙之中详细地审阅了拙稿，不仅对稿件本身的内容提出了许多宝贵的意见，还对我们的研究提出了十分有见地的建议。我们在此对他表示由衷的感谢。

参考文献

Zellig S. Harris. 1954. Distributional structure. Word,10(23):146–162.

Marine Carpuat and Dekai Wu. 2005. Word Sense Disambiguation vs. Statistical Machine Translation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05), pages 387–394, Ann Arbor, Michigan,June. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation Using Word Sense Disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 61–72.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A Topic Similarity Model for Hierarchical Phrase-based Translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 750–758, Jeju Island, Korea, July. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2013. A Topic-Based Coherence Model for Statistical Machine Translation.In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13), Bellevue,Washington, USA, July.

Deyi Xiong and Min Zhang, 2014. A Sense-Based Translation Model for Statistical Machine Translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1459–1469, Baltimore, Maryland, USA, June 23-25 2014.

Van de Cruys and Apidianaki, 2011 Latent Semantic Word Sense Induction and Disambiguation. In

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1476–1485, Portland, Oregon, June 19-24, 2011.

Zhendong Dong, Qian Dong, HowNet and the Computation of Meaning, World Scientific, 2006