知网机译系统中的语义计算

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

摘 要： 本文介绍的知网机译系统是一个直接以知网知识系统为其语言资源的基于规则的系统。本文在概要地介绍了该系统的基本组成之后，阐述了该系统的语义计算。文章指出该系统的核心是由作者提出的逻辑语义关系。这样的关系是该机译系统的源语言分析的目标，也是目标语生成的依据。文章给出了逻辑语义关系的定义,说明了它与句法关系的区别,并指出坚持知网机译所采用的逻辑语义关系的完备的必要性。文章通过介绍一些所采用的语义计算功能函数，显示机译系统的语义计算的深度，以及知识在机译研究中应有的位置和作用。

关键词： 知网，机器翻译，基于规则的机器翻译，语义计算，逻辑语义关系

Semantic computing in HowNet MT system

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

Abstract: *This paper describes HowNet English-to-Chinese machine translation system (HowNet MT). HowNet MT is a rule-based system, in which HowNet is used as its common-sense knowledge support. After giving a comprehensive outline of the system, the paper introduces logical semantic relationships (LSR) which function as the core of the HowNet MT. LSR is the goal of the analysis of the input English and the basis of the transfer and synthesis of the output Chinese. By giving fine examples the paper shows the semantic computing in HowNet MT and its depth. The paper presents some main functions employed in the system in general and the resolution of sense disambiguation and sense induction in particular.*

Keyword: *HowNet, machine translation, RbMT, semantic computing, logical semantic relationship*

1. 前言

知网-英汉机译（HowNet MT）是一个典型的、手工作业的、基于规则的系统，其研发始于 2011 年初，完成于 2013 年底。研发该系统的主要目的是为了进一步考核和完善知网，验证我们 30 余年前提出的关于建设常识知识库的设计、策略和方法，最后要回答一个问题，即有了完善的语义系统的支持是否能够大幅度提升机译系统的翻译质量和系统开发的效率。这里我们先介绍知网-英汉机译的概貌，以方便本文以后的阐述和讨论。知网机译系统是高度模块化的，主要模块有：

1. 知网词典：包含英中词条均在 11 万以上，义项各为 13 万余。词条上有十分详细的语法分析和转换生成所需的信息。
2. 公理规则库：现包含共性与个性规则计 1200 余条，用于动态地建立意义群落库，可供意义群落测定用。
3. 分析与转换生成规则库：包含句法驱动的共性规则约 4500 条，以及词汇驱动的个性规

则约 5400 条。粗略统计至少有 50%以上的规则用到语义信息。

4. 英语分析模块包括：

A. 查词典（Dictionary-Lookup），前后缀，大小写，未登录词信息获取与处理；

B. 意义群落测定器（Sense-Colony-Testing）消除歧义；

C. 句法语义分析包括如下子模块：

（1）未登录词处理（Unknown-Processing）-- 含 Unknown_Words，Prefix_Suffix 等 2 个子模块；

（2）"越...越..."结构的专项处理（the more...the better_Pattern Processing）-- 含 Yue_Lai_Yue1 个子模块；

（3）多词性判定（POS-Tagging）-- 含 Homo_Adj_Adv，Homo_Adj_Noun，Homo_Adj_Noun_Verb，Homo_Adj_Noun_Vxxx，Homo_Adj_Ved，Homo_Adj_Verb，Homo_Adj_Ving，Homo_Adv_prep，Homo_Adv_prep_wh，Homo_Noun_Ved，Homo_Noun_Verb（1/2），Homo_Noun_Ving，Homo_Prep_Wh，Homo_Root，Homo_That 等 16 个子模块；

（4）屏蔽特定词语（Masking）-- 含 Masking 1 个子模块；

（5）词语的句法功能判定（Syntactic-Functions）-- 含 Function_VBS，Function_Ved，Function_Ving，Function_Wh_Word 等 4 个子模块；

（6）紧密短语捆绑（Binding）-- 含 Bind_adj_adv，Bind_noun，Bind_other，Bind_verb 等 4 个子模块；

（7）语块组合（Chunking）-- 含 Chunk_adj_adv，Chunk_noun，Chunk_verb 等 3 个子模块；

（8）标点功能判定（Punctuation/Coordinate Relationships）-- 含 Bracket，CLSEGAND 等 2 个子模块；

（9）复杂并列结构关系确定（Siblings）-- 含名词并列结构、动词并列结构等共 3 个子模块；

（10）句法关系确定（Syntax Relationships）-- 含 Link_Adj，Link_Adv，Link_Infs，Link_Noun、Link_Verb，Link_wh 等 6 个子模块；

（11）深层语义关系确定（Deep-Semantic Relationships）-- 含 Deep_patient1 个子模块；

（12）介词短语关系确定（PP Attachment）-- 含 about 等 55 个英文介词子模块，以及 Prep_Phrase 处理短语性介词的 1 个子模块；

（13）复查校对（Proof_Reading）-- 含 Proof_Read 子模块；

5. 转换生成包括：

（1）动作性名词词组生成 -- 含 Trans_ActionNP 子模块；

（2）动词词组主动态生成 -- 含 Trans_Active 子模块；

（3）从句生成 -- 含 Trans_Clause 子模块；

（4）比较句生成 -- 含 Trans_Comparison 子模块；

（5）疑问句生成 -- 含 Trans_Interogative 子模块；

（6）被屏蔽词语生成 -- 含 Trans_Masked 子模块；

（7）非动作性名词词组生成 -- 含 Trans_NPS 子模块；

（8）否定意义生成 -- 含 Trans_Neg 子模块；

（9）名词词语否定生成 -- 含 Trans_Nominal_Neg 子模块；

（10）数词词组生成 -- 含 Trans_Numeral 子模块；

（11）动词词组被动态生成 -- 含 Trans_Verb_Passive 子模块；

系统一次运行将对输入原文文本进行 20 次扫描。据测算，系统翻译句长平均 20 个英文词语的运行速度为每句 25 毫秒。

2. 逻辑语义关系—知网机译的核心

逻辑语义关系作为机译的核心是董振东于 1978 年提出的，知网和知网机译均继承了这一思想和技术。在知网机译系统中，句法关系仅仅是知网机译分析的阶段性结果，而不是分析求解的最终结果，也不是转换生成的依据。 逻辑语义关系才是知网机译系统中原文分析的目标，是两种语言转换的平面，是译文生成的依据。逻辑语义关系包含了四个层次的关系，它们是：

（1）语言传输者和语言接受者之间的关系；

（2）语言者和语言片断内部其他实体之间的关系；

（3）语言内部除语言者外的各实体之间的关系；

（4）语言内部实体的附属特征。

我们把前两个层次归为一类，它包含：传讯、求讯、命令等；后者为另一类包含了 87 种关系。知网总共采用的语义关系是 91 种，两者基本一致。它们包含有：施事、经验者、领有者、受事、对象、内容等。这 91 种关系是经过用以标注中英文 22 万条记录、经过包括知网机译在内的 5 个机译系统验证了的。我们认为两种语言之间的转换（翻译）的本质是意义的转换。它们一定有共同的内容，特别是语言内实体间的关系。另外，众所周知，中文的主要语法手段是：词序和虚词的应用。那么中文的词序排列的依据是什么？根据我们长期以来的观察，认为主要是我们认定的逻辑语义关系。在知网机译系统的转换生成采用的是（1）共性的逻辑语义关系统排（简称“共性统排”），和（2）个性的逻辑语义关系的调整（简称“个性调整”）相结合的方法。共性统排是认为有一个假想的中文句子，它包含了所有的知网所认定的逻辑语义关系。在这个词序排列中，每一个父节点与其所辖的子节点的顺序都必须遵照此顺序。

特别要提醒的是：逻辑语义关系不仅是我们的机译系统的核心，也是知网知识数据库所包含的概念定义的核心。试看知网词典中以下几个概念的定义：

“犯罪” （commit a crime）

{engage|从事:RelateTo={police|警},content={fact|事情:modifier={criminal|刑事}
{guilty|有罪}}}

“犯罪主体” （subject of a crime）：

{human|人:domain={police|警},{engage|从事:agent={~},content={fact|事情:
modifier={criminal|刑事}{guilty|有罪}}}}

“犯罪客体” （object of a crime）：

{human|人:domain={police|警},{suffer|遭受:content={fact|事情:
modifier={criminal|刑事}{guilty|有罪}},experiencer={~}}}

知网由词语到短语，再到句子和篇章的描述体系是一体化的。

2.1 逻辑语义关系与句法成分的区别

首先，我们严格地要把逻辑语义关系与句法成分区别开来。知网知识系统所采用 91 种语义关系是我们真正确立的逻辑语义关系，它们是与句法成分相区别的，它们是不依赖特定语言的。而知网机译所采用的 87 种关系因为它们的顺序将是中文的词序，因此其中有个别的实际上并不是我们认定的逻辑语义关系如：“带被-施事”、“带把-受事”、“意志”、“定语从句”、“状语从句”等。在研究逻辑语义关系，特别是如前所述的第（3）类，即语言内部的各实体之间的关系，或类似的诸如论元等时，要特别注意它们与句法成分的界限。究竟是

多一点好还是少一点好？究竟是不怕复杂还是追求简单好？我们的回答是该多少就多少，也不怕复杂，不怕难以区别，关键是一条：定义是否严格，是否有真实文本的语言现实的支持。举几个例子：“DurationAfterEvent 后延时段”、“duration 时段”等是否可以与“time 时间”合并呢？从下面的实例看，显然是不可以合并的。

他不知道老师已过世 3 年了 （DurationAfterEvent 后延时段 — 事件发生后起算的时段）

他打算在村子里呆 3 天（duration 时段 - 事件开始后起算的时段）

3 天前他还在村子里（time 时间 - 事件发生的时间点）

从把上面各句译成英文，可以知道把上述 3 个逻辑语义关系合而为一是不可取的。它们事实上也确实是不同的。

其次，应该遵循一个原则：对于句子内部关系而言，相同的句法关系或结构，可以有不同的逻辑语义关系；反之不同的句法关系可以有相同的逻辑语义关系。例如在英文里“the doctor’s diagnosis of the disease”与“the doctor diagnoses the disease”的句法结构和关系是完全不同的，但是它们有着相同的逻辑语义关系。如果一个系统或语言资源库标注结果不同，那么可以肯定其“论元”不够纯净。这两句在知网机译里的分析结果是一样的都是：“施事-行动-内容”。它们的中文译文表层不同，“医生对疾病的诊断”。“这位医生诊断疾病”。但是可以知道它们的逻辑语义关系是一致的。那么它们之间有没有差异呢？其差异在于上述第（1）类的关系，两者虽均为“传讯”，但第 1 句传达的是一个事实，而第 2 句传达的是一个事例。“事实”是静态的，非过程性的，而“实例”是动态的，具有过程性的。正因为如此，才导致它们（英文和中文）的表层结构有所不同。

2.2 逻辑语义关系的定义

我们很早就注意到有很多系统或语言资源的论元的定义不够精确，有的则与句法成分没有什么区别，例如“受事”，基本上就是“宾语”的另一个名称而已。

知网对于逻辑语义关系的定义是极其严格的， 它所采用的方法首先是对知网词典里的全部的事件义项标注它们可以管辖的是“patient”、“content”、“target”、“PatientProduct”、“ContentProduct”，还是“possession”。虽然它们都是事件的处置的对象，但是它们是有区别的。例如：

“PatientProduct”是“创造”的物理性结果如“build a **house**”

“target”是“表示情感”的目标、其他实体转移中的接受者，以及内容转移的接受者如“love **him**”，“sell **him** a toy”，“tell **her** a story”

“content”是“改变感知状态”的信息如“express our **condolences**”

“ContentProduct” 是“创造”的精神产品如“write a **book**”

“possession”是领有物如“I have a **book**”

知网作为一个知识系统，从认知的角度，必须回答：“写书”、“烧书”、“读书”、“爱书”、“有书”中动作与“书”之间的关系有无差别？如果在这些语境中的“书”都是“patient”，是否合乎事实？当然，在面对不同的应用时，可以有所取舍，可以删繁就简。

2.3 语言内部实体的附属特征

语言内部实体的附属特征表示的不是语言内部两个实体间的关系，也不表示语言者与语言实体的关系。它们是附属在语言内部实体身上的某种特征。这些特征在机译系统中的作用是不可忽视的，因为它们对于原文分析与转换以及译文生成都会产生很大的影响。限于篇幅我们不可能一一详述这些特征，我们只是将它们分列在附录中。这里只是强调一点，即这些特征往往不是孤立的，它们会多个交叉共现，进而会加大分析和生成的难度。例如：

There are no better electric stoves than this type.

看来简单，但是因为存在着“比较”、“否定”等附属特征，加之英中两种语言在这些方面差

异又较大，使得它无论对于统计机译还是规则机译都比较困难。我们不得不在这些附属特征上多下功夫。

2.4 一个具体实例

从下面的实例看逻辑语义关系在知网机译中的应用，请参见附录 2。

原文： He mentioned nothing to me about his previous job in the university.

知网译文： 有关他的以前在大学的工作，他对我没有提及任何东西。

（1）该句中原文分析的结果：

全句中心 – mention	2 个主要特征：	past 和 pred(icate)
	4 个子节点：	he – agent
		nothing - content
		to (me) - aim
		about (job) – concerning

子节点 – nothing	1 个主要特征：	neg(ative)
	3 个子节点：	his – possessor
		previous - time
		in (unversity) - location
子节点 – university	1 个子节点：	the -- demo

（2）转换生成后译文结果：

全句中心 – 提及	2 个主要特征：	past 和 neg(ative)
	4 个子节点：	他 – agent
		任何东西 – content
		对 (我) – aim
		有关 –concerning - sentential
子节点 – 任何东西	3 个子节点：	他的 – possessor
		以前的 – time
		在 (大学) – location
子节点 – 大学	1 个子节点：	（） – demo

实例显示：知网机译系统的分析结果，原文句子各节点的逻辑语义关系，以及系统依此逻辑语义关系来转换和生成译文。译文的句子层词序排列是：

Sentential agent aim head content

有关…的工作 他 对我 提及 任何东西

译文的短语层词序排列是：

Possessor time location head

他的 以前 在大学的 工作

如果短语层将三个限定语笼统地定为定语，那么应该如何排列它们的词序呢？请与下列流行机译系统的这部分的译文比较。

系统 a：他在大学以前的工作。

系统 b：在大学里他以前的工作。

系统 c：他过去的工作在大学

系统 d：他的早先工作在大学。

3. 知网的语义计算函数

知网的语义计算函数在知网系统所附的手册中已经载明，本文不再一一赘述。本文要介绍和讨论几个与语义计算函数相关的关键性问题。

大家知道知网与海内外许多语言资源不同，它描述的对象是概念，它对概念的描述不仅仅是给出一个分类标记（class），或再加上一个分类层级体系（taxonomy）。知网的最主要的独特之处在于它采取了以义原和语义关系为基础的结构化语言来定义每一个概念。 因此它的语义计算在深度和广度以及复杂度上都是独特的。

首先，我们介绍提取概念定义的函数。如果提取的是完整的定义，当然比较简单。当我们要提取的是定义的一部分我们用的函数是：

（1） 关于“>”如*DEF> -- 表示概念定义中应包含有列于{ }中的内容，例如：

```
suspect     CW[^absolute];L1[*DEF_inDic>{fact|事情:modifier={guilty|有罪}}]
              $@replace(CW,noun);CW[*HY=`嫌犯`]@bind(CW,L1).
```

这是一条判别“suspect”词性的个性规则。 在知网词典并没有一条词条的 DEF 是“{fact|事情:modifier={guilty|有罪}}”。但是实际上其 DEF 中包含着这样两个单元，即“fact|事情”作为其首义原，在其定义的其他位置上有“modifier={guilty|有罪}”的词条共有 132 条。例如：

```
“bigamy” ：          {fact|事情:CoEvent={GetMarried|结婚:frequency={again|再}},
                        modifier={guilty|有罪}}
“cyber-terrorism” ：  {fact|事情:CoEvent={TerrorAttack|恐怖袭击:
                        means={internet|因特网}},modifier={guilty|有罪}}
“fraudulence” ：     {fact|事情:CoEvent={cheat|骗},modifier={guilty|有罪}}
```

（2） 关于“inDic”如 *DEF_inDic -- 表示只要某词条在词典里确实含有一个列于{ }中的内容的， 例如：

```
in      L1[adj,*QF=*CW];R1[*DEF_inDic>{part|部件:whole={place|地方}}]
          $@replace(R1,{part|部件:whole={place|地方}})
          CW[*HY=`在`,*LOG=`location`]@link(L1,CW).
```

这是一条介词“in”的规则。设我们要处理的句子是“The situation is more dangerous in this area.”“area”在词典中有含有 5 个概念或称义项。这一规则是说只要其中有一个概念符合规则的内容，即匹配成功。

含有“inDic”的函数对于实现两个概念的语义互感是非常有用的，在多数情况下甚至是必须的。例如：“reduce”和“jaw”它们各自都是多义词，但它们之间有着一种词义互感的关系。在机译运行过程中的某个阶段，是难以断定它们各自当前是处于哪一个义项。这时候我们就采用含有“inDic”的函数来处理它们。知网机译系统的个性规则有一条：

```
reduce      CW[verb,absolute,^GY0,*DP=`patient`];
              Z1[*DEF_inDic>{part|部件:domain={physiology|生理学}}]
              $@replace(CW,{doctor|医治});
```

@replace(Z1,{part|部件:domain={physiology|生理学}}) CW[GY0].

这条规则的含义是：当“reduce”作为动词，且如果有“受事”，其义项中有“动物体部件”，那么“reduce”按“复位”义项的词条置换，同时那个“受事”则按“动物体部件”义项的词条置换。

前面我们曾说明逻辑语义关系与句法成分的区别。这里我们将作进一步比较。上面的关于“reduce”的规则中，有一个表达是“*DP=patient”，这是什么意思呢？这是说如果“reduce”这个词在分析时表明它有“patient”的话。知网所指的“patient”在英文的句法层面有哪些结构表现呢？它应该包括如下：

1. The surgeon reduced his dislocated jaw.	这名外科医生为他的脱臼的颌复位。
2. His dislocated jaw was reduced by the surgeon.	他的脱臼的颌被这名外科医生复位。
3. His jaw that was reduced yesterday was dislocated again.	昨天被复位的他的颌再一次脱臼。
4. His jaw the surgeon reduced yesterday was dislocated again.	这名外科医生昨天复位的他的颌再一次脱臼。
5. The jaw reduced by the surgeon was dislocated again.	这名外科医生复位的颌再一次脱臼。
6. The surgeon had two more dislocated jaws to reduce today.	这名外科医生又有两个脱臼的颌今天要复位。
7. The professor examined his jaw to be reduced in the clinic.	这位教授检查他的要在诊所里被复位的颌。
8. His jaw was easy for the intern to reduce.	由这名实习生复位他的颌是容易的。

上面 8 个英文句子虽非源于真实文本，但它们的句法和语义关系结构都是合法的。这些句子中“reduce”和“jaw”之间的句法关系都不相同，但它们的深层语义关系都是相同的。由此可以认识到知网机译语义计算的深度。知网把这样的“受事”叫做“深层受事（deep-patient）”。上述各句中的“reduce”和“jaw”的歧义判别都是通过执行上面的“reduce”个性规则完成的。这里有一点要提醒：第一，在上面的规则的结论部分不仅规定了“reduce”的应判定结果，也同时要求置换其“深层受事”应该判定的结果。如果不对“深层受事”规定，其结果就有可能出现这样的结果如：He reduced her foot.译文却是“他为她的英尺复位”。这就是我们前面提及的概念的相互感应和制约的现象。利用概念感应是消除歧义的有效方法之一。知网机译系统的个性规则和介词短语管辖关系与意义的判定规则都大量利用了概念感应。

除了深层受事外，知网还挖掘一种“深层修饰（deep-modified）”。试看下面的个性规则：

strong CW[adj,^GY0];FH[*DEF>{drinks|饮品:{addict|嗜好:patient={~}}}]
$@replace(CW,{TasteStrong|味浓})@replace
(FH,{drinks|饮品:{addict|嗜好:patient={~}}});CW[*HY=`凶`,GY0].

这条规则的含义是：当“strong”作为形容词，且其深层被修饰词语，亦即它的深层语义父节点的概念定义是“酒”，则应按“味道浓重”置换词典中的义项。例如：

1. The liquor is strong.	白酒很凶。
2. I have some very strong liquor.	我喝一些很凶的白酒。
3. The chemical can make liquor stronger.	这样的化学剂能使白酒更凶。
4. The liquor was made stronger by the chemical.	这瓶白酒因化学剂变得更凶。
5. This liquor, although a bit strong, tastes very nice.	这瓶白酒，虽然有点儿凶，吃起来很可口。
6. Although a bit strong, this liquor tastes nice.	虽然有点儿凶，这瓶白酒吃起来可口。
7. All the liquor that became stronger was thrown away.	变得更凶的所有的白酒被丢弃。

8. The liquor that I threw away was too strong. 我丢弃的白酒太凶。

4. 讨论

历经近 40 年，先后研发过 4 个英中机译系统，值得安慰的是有所进步，特别是最近的这个系统。有一些问题值得讨论。

第一，无论哪一种类型的机译系统，有三种知识是不可少的，即语言内的知识、语言外的知识，以及语言间的知识。无论它们是显性的还是隐性的。迄今为止无论是基于规则的还是基于统计的系统，这三种知识都是欠缺的。真正要很好研究的是如何让系统有更全面、更多的知识，以及如何使它们相互配合，合理地、最大限度地得到运用。

第二，在系统的研发过程中，“按下葫芦起了瓢”的捉襟见肘的困境并不是基于规则系统的固有的现象。之所以会产生这样的现象是因为过去语言资源，特别是知识资源不足。知网机译系统在开发过程中，今天改昨天编写和调试好规则是不多见的。 知网机译的调试主要的是增加规则，修订某些不合理的或者不完备的数据。知网的所有的程序是非常稳定的。它的解释程序的解释能力是非常强的。这也是避免“按下葫芦起了瓢”的基本保障。

第三，知网的词典是双语、通用的，同时知网没有采取利用大规模文本进行词频统计，这样就会遇到选择义项的困难。例如：中文自身有词语“父母”、“爸妈”、“老爸老妈”、“二老”、“椿萱”等，而它们的英文对译词是“parents”。我们不得不对所有的多义项词语进行逐一考察，加以优先选用或暂不选用的标记。因此我们知道了双语双向的词典跟双语单向的词典是不同的。由此说明，通用的知识资源在应用于特定目标时，调整或适应是必须的。

最后但可能是更应注意的，愿以刘涌泉的 2014 年初的文章《机器翻译一甲子（1954-2014）》的几句话结束本文：刘涌泉写道：“16 年前，我曾写过一篇文章“机器翻译归根到底是个语言学问题””。他还指出：“为了使机器翻译更上一层楼，有必要扭转忽视语言研究的倾向。”即便是统计机器翻译，其最大的进步是它从最初宣称可以完全不用语言学家的知识到如今已认认真真地吸取句法、语义信息，进而从“纯净的”统计发展为“混血的”了。

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以致文字提出了许多宝贵的意见、还对我们的研究项目提出了颇有见地的建议。我们在此对他表示由衷的感谢。

参考文献

Francisco Guzm´an，et al. (2014), Using Discourse Structure Improves Machine Translation Evaluation Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 687–698,

Corbí-Bellot, M. et al. (2005) "An open-source shallow-transfer machine translation engine for the romance languages of Spain" in Proceedings of the European Association for Machine Translation, 10th Annual Conference, Budapest 2005, pp.79–86

Tsujii, J., Linguistic Knowledge, World Knowledge and Conceptual Knowledge, in : Proceedings of the International Symposium on Electronic Dictionaries, in Tokyo, November 1988

董振东，1981 逻辑语义及其在机译中的应用，中国的机器翻译，刘涌泉编，知识出版社，北京，1984，pp. 25-45

董振东，董强，2001，面向信息处理的词汇语义研究中的若干问题，语言文字应用，第 3 期，pp. 27-32

刘涌泉 2014，机器翻译一甲子(1954-2014)，语文建设通讯 第 105 期，2014-01，pp. 1-5

附录 1

时间特征：过去、现在、将来

动词体特征：与时间相关(完成、进行)

动词趋向

动词进程（假定、发端、进展、延续、完结）

动词结果（达成、未达成、有能力、没能力、有可能、没可能、试试）

动词态 （主动、被动）

名词数特征：（单数、复数）

名词体特征：（静态、动态）

否定：全否定、半否定

比较：正比较，负比较

中文特征：特殊结构（把字结构、被字结构、是…的结构）

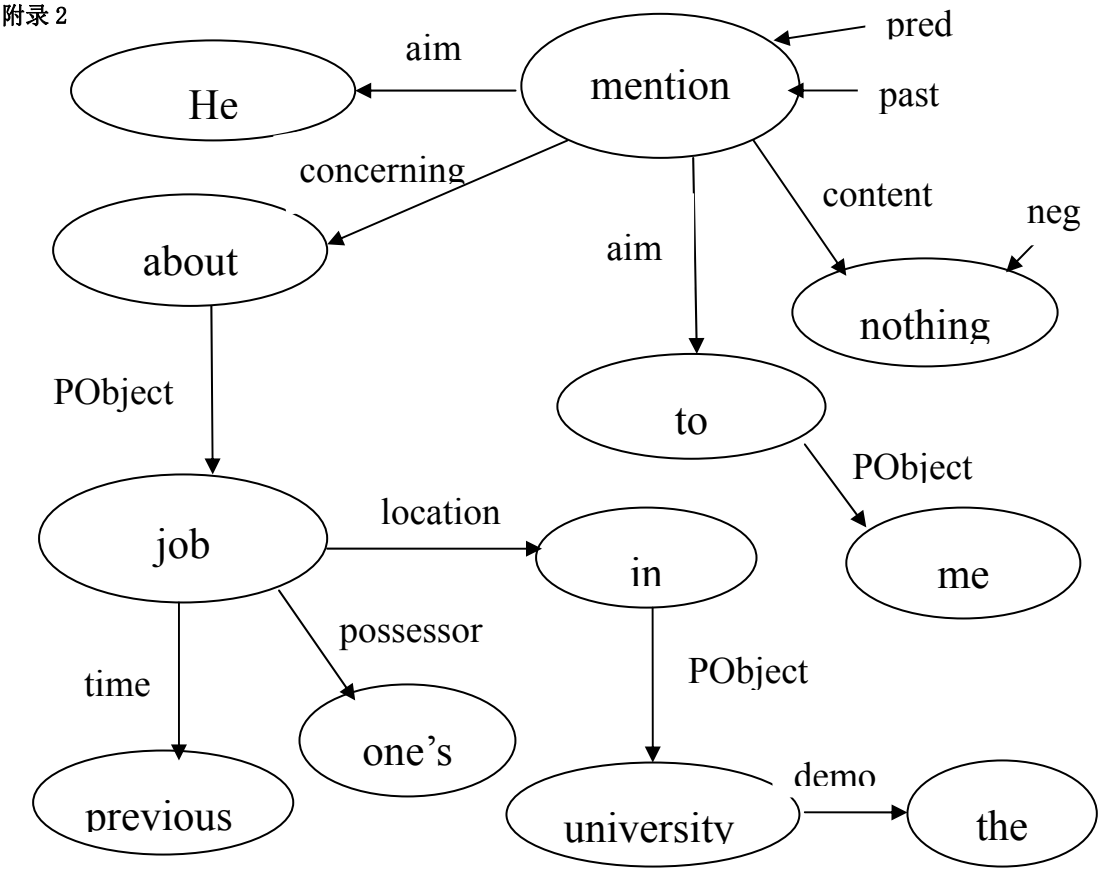
词语汉字音节数

句子种类：简单句、复杂句

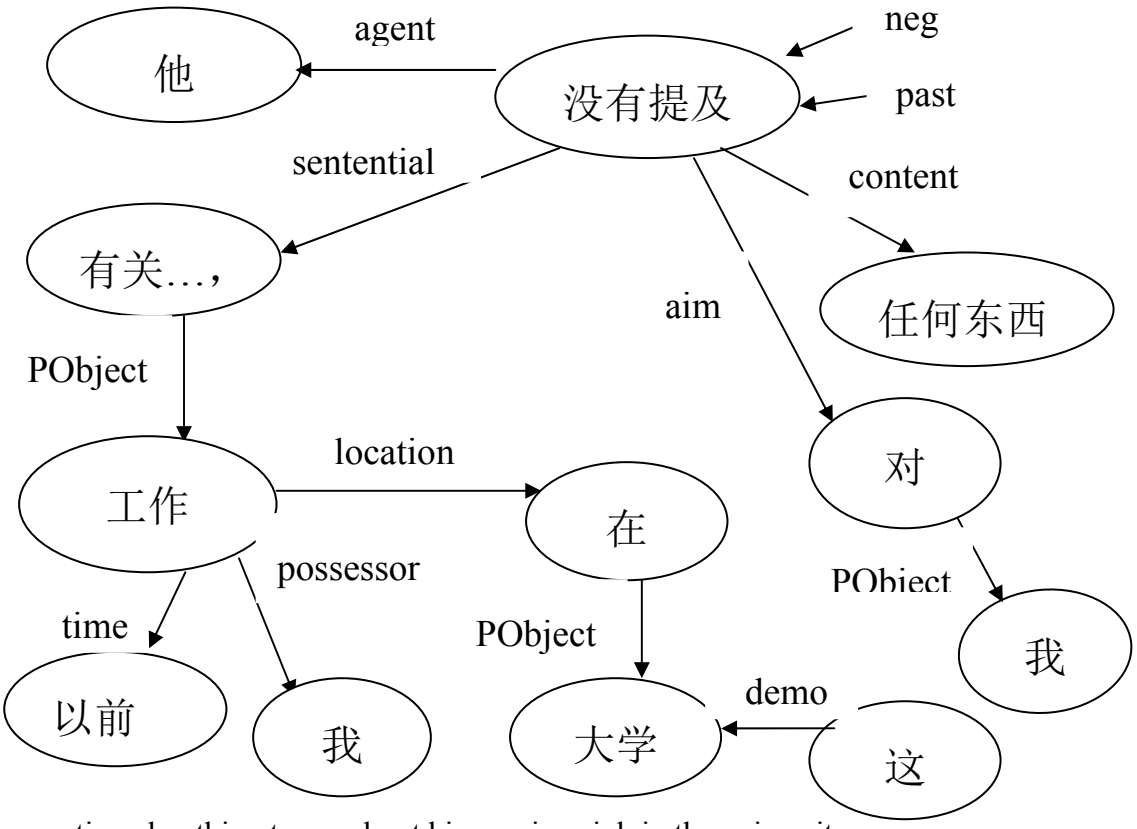
句子层次：主句、从句、短语层、句层

词特征：全大写、首字母大写、缩略语、未登录词、前缀、后缀

附录 2



****



He mentioned nothing to me about his previous job in the university.
有关他的以前在大学的工作，他对我没有提及任何东西。