概念作为深度学习的基础的准备

提纲

- 概说
- 知网的变与不变
- 应用带动和检验研究

概说

- 知网是什么
- 知网的30年的历程
- 知网的变与不变

知网是什么

- 定义： 知网（英文名称为**HowNet**）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识系统。
- 核心：义原、概念、关系组有机组合的知识系统。

知网的30年的历程

- 1988年设计探索
- 1997年工程实施
- 2000年公布推出
- 2006年知网专著出版《HowNet and the Computation of meaning》,并同年由世界科学出版社出版发行。
- 2012年12月荣获"钱伟长中文信息处理科学技术奖"一等奖。
- 2014年我们推出了基于知网的英汉机器翻译系统。
- 2016年我们推出了基于知网的中文文本分析系统。

知网的30年的历程

近20年来得到海内外同行业界的可贵支持

近日倪光南院士在给董振东的信中写道：

“董老师：

知网已经发挥了很好的作用，现在语言信息处理发展很快，它的应用也越来越普遍，我相信你们的平台会受到大家的关注。”

“知网做得很好，我试了一下，大家都能享用自然语言处理的成果了！

由于将高深的知识赋予给计算机，使自然语言计算成为可能，我相信这将会推动该领域的发展和扩大成果的推广应用。”

知网的变与不变

非常幸运：　该变的，变了；
　　不该变的，没变！

- 知识词典信息变化
- 中英文词语同步增加和更新，与时俱进

知网规模的变化

- Chinese character 20898
- Chinese word & expression 126330
- English word & expression 117704
- Chinese meaning 143997
- English meaning 139110
- Definition 33904
- Record 228663

知网信息的变化

- 中、英文词语增加了句法信息，以适应中文句法分析的需要

NO.=097966
W_C=看
G_C=verb [2 MustObj] [kan4]
S_C=PlusEvent|正面事件
E_C=~病，医生~病人，一天要~几十个病人，医生给病人~伤口，去医院~眼睛
W_E=treat
G_E=verb [7 treat verb -0 vt,sobj,ofnpa]
S_E=PlusEvent|正面事件
E_E=
DEF={doctor|医治:domain={medical|医}}
RMK=

中英文词语同步增加和更新

- 中英文始终保持同步

  如：吐槽- belittle 、小三-Lolita、点赞- acclaim

- 增加新的句法特性标识

  如：quasivt 不及物可带宾语

银行变身妖股；抢滩上海；投资中国；登陆中国；吴恩达辞职百度；王海峰掌舵百度AI平台；潜伏你我身边；美对土总统保镖施暴抗议者强烈不满；

知网的不变

核心不变：

哲学、义原、关系、体系架构、知识描述语言不变

这才体现了知网的设计的正确性，系统的稳定性，与时俱进的适应能力。

面向应用的研发

- 相关性计算；
- 义群测试；
- 英汉翻译；
- 中文文本分析；
- 英文文本分析；

应用带动和检验研究

- 从义原说起
- 什么是真正的知识系统

从义原说起

- 什么是义原？真的有义原吗？
- 知网如何认定义原？
- 知网如何认定关系？
- 知网如何利用义原和关系？

什么是义原？

- 义原是最小的意义单位，它独立地具有特定的意义，也可以与其他义原组合成复合的概念。
- 一个以汉语为母语的人一生中认识和应用多少汉字？大约3000，那么所有的意义不是都在这里了吗？尤其是常识中的意义应该都被覆盖了。

知网如何认定义原？

知网观察探寻义原大体有如下步骤：

- 选定4000个汉字，将它们的义项一一列出来；
- 对列出的义项同类归并；这样例如汉字中的“医”、“治”、“看”、“瞧”、“疗”所具有的“医治”暂定为一个义原。
- 对归并后余下的义原，观察它们可能存在的关系，研究它们是什么关系。例如，我们发现了它们主要体现了“横向”关系，如：“医”、“药”、“患”、“病”等

义原

- 义原 2086
  - 实体 152
  - 事件 802
  - 属性 245
  - 属性值 887
- 次要特征 128

知网如何认定关系（1）？

- 关系是知网的灵魂，因为关系是知识的灵魂；
- 义原和概念组合成一个关系的系统，才是知识。
- 知网就是要反映这个系统，并且要使其能够为计算机所认识、理解和运算。

知网如何认定关系（2）？

1. 汉字（单音节）选定义原；
2. 汉语词语（多音节）解决关系；

如：事件-施事：行人、教师、患者

工具-事件：枪击、口试、笔耕

方式-事件：严惩、力挺、暴打

部件-修饰：口臭、

事件-处所：诊所、学校、卖场

什么是真正的知识系统（1）

- 义类词典（thesaurus）是吗，如同义词词林？否
- 同义词集（synset）是吗，如WordNet？否。

  第一，如果它们不是可机读的或数字化的，自然就不在我们讨论之列。

  第二，它们的初衷是为了人们写作和翻译用于查找更加适宜的词语的，即是一种word-finder。

  即使它们是可机读、数字化的，它们也不是我们所指真正意义上的知识，或常识。它们只是的知识的一个方面，是语言学范畴内的知识。
- 当下流行的“知识图谱”是吗？只能算是部分的是，但却是一个“黑箱子”，它还不是体系化的，对人不够透明，不可控。

什么是真正的知识系统（2）

- 知网的基于义原的概念定义，举例：

“国家” -- {place|地方:PlaceSect={country|国家},domain={politics|政}}

“夫人” -- {human|人:belong={family|家庭},modifier={female|女}{spouse|配偶}}

“总统” -- {human|人:HostOf={Occupation|职位}, domain={politics|政},
modifier={HeadOfState|元首},{manage|管理:agent={~},
patient={place|地方:PlaceSect={country|国家},domain={politics|政}}}}

“第一夫人” -- {human|人:***belong***={family|家庭}
{human|人:HostOf={Occupation|职位}, domain={politics|政},
modifier={HeadOfState|元首},{manage|管理:agent={~},
patient={place|地方:PlaceSect={country|国家},domain={politics|政}}}},
modifier={female|女}{spouse|配偶}}

知网的基于义原的概念定义的关键点

- 结构化；
- 可多层次嵌套；
- 利用处理时，可拆解，可单独提取任何一个义原；
- 不同层次可设不同的权重；
- 第一个义原为类义原，享有最高权重；

继承关系

“总统”和“第一夫人”，都是{human|人}，因此他们还具有“人”的所有的特征，以及继承{human|人}的上位的所有的特征；

├{thing|万物} {entity|实体:{ExistAppear|存现:existent={~}}}
│ ├{physical|物质} {thing|万物:HostOf={Appearance|外观},{perception|感知:content={~}}}
│ │ ├{animate|生物} {physical|物质:HostOf={Age|年龄},{alive|活着:experiencer={~}},{die|死:experiencer={~}},{metabolize|代谢:experiencer={~}},{reproduce|生殖:agent={~},PatientProduct={~}}}
│ │ │ ├{AnimalHuman|动物} {animate|生物:HostOf={Sex|性别},{AlterLocation|变空间位置:agent={~}},{StateMental|精神状态:experiencer={~}}}
│ │ │ │ ├{human|人} {AnimalHuman|动物:HostOf={Name|姓名}{Wisdom|智慧}{Ability|能力},{think|思考:agent={~}},{speak|说:agent={~}}}

上面是知网的分类体系中的记载。“总统”，管理国家，是他的个性特征；他会思考、说话，是他作为”人“的特征；他也会”生“、”死“，作为”生物”的特征。

这些不就是“常识”吗？

从这里可以看到：知网是一个知识体系，而绝非一部义类词典或同义词词典！

应用举例（1）

试看下面的知网的中文词条“初二”的两个概念定义：

初二（1）-- {Rank|等级:domain={education|教育}, host={InstitutePlace|场所:domain={education|教育}, modifier={intermediate|中等},{study|学习:location={~}},{teach|教:location={~}}}}

初二（2） -- {time|时间:TimeSect={day|日}, modifier={specific|特定}}

输入测试文本（部分的）：

11月18日晚7：00初二第一次学生家长会在各班教室举行。

结果：初二（1） 0.4722222222

初二（2） 0.3371428571

知网义群测试器

输入测试文本（全文本）：

"11月18日晚7：00初二段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。"

初二（1） 0.5503191615

初二（2） 0.3371428571

中文文本分析器

分析结果显示表各栏目说明

000	分-合词后词号
	加工后屏蔽信息
000	加工后新词号
expression	词语
FH	父节点词号
Son	子节点词号
ES	姐节点词号
YS	妹节点词号
DP	深层语义父节点词号
Deep Son	深层语义子节点词号

中文文本分析器

语知科技

关于我们 核心优势 产品体验 文档中心 联系我们

NOID_1	MASKID	NOID_2	EXPRESSION	FH	SON	ES	YS	DP	DEEPSON	LOG	DEEPLOG	UNITID
000		000	ROOT									
1	*	1	合肥	005						location		073938
4	*	4	一	005						quantity		187300
5		5	女子	000	004001			012			referent	123298
8		8	为	000								167093
9		9	证明	000	015019							204797
12	*	12	自己	015					005	relevant		212815
15	*	15	是	009	016012					content		148333
16	*	16	女性	015				019		isa	agent	123228
19	*	19	花费	009	023				016	content	attributive	076851
22	*	22	7	023						quantity		003847
23	*	2	年	019	022					duration		121582
24		24	。	000								010366

截图(Alt + A)

分析结果显示表各栏目说明（续）

log	逻辑语义关系
DeepLog	深层逻辑语义关系
POS	词类
UnitID	对应知识词典的义项
GramInfo	知识词典的句法信息
TempInfo	加工中的临时信息 1
Aspect	加工中的临时信息 2

关于逻辑语义关系

语言分析的重要任务是得出两个节点之间的我们称之为逻辑语义关系，如agent、patient等，有人问为何要100多个？

1. 她大后年将从医学院毕业了 -- 时点 time
2. 两年后她医学院该毕业了 -- 事件前时段 DurationFE
3. 他在医学院读了8年 -- 事件时段 duration
4. 不到两年她就该从医学院毕业了 -- 事件前时段 DurationFE
5. 他毕业都3年了 -- 事件后时段 DurationAE

我们认为NLU是为了最大限度提取文本所含的信息。

为此，知网的中文文本分析结果包括两大类逻辑语义关系：

（A）与句法关系平行的语义依存，在结果文件中置于“log”中；

（B）深层逻辑语义，置于“DeepLog”中。请看下例：

看下面句子的分析演示

合肥一女子为证明自己是女性花费7年。

请看“DeepLog”栏目中的各项：

”referent“(指代)，012号词”自己“是005号词”女子“的父节点；

”agent“（施事），019号词”花费“是016号词”女性“的父节点；

”attributive“（定语），016号词”女性“同时被019号词限定

这个分析结果对于下列问题可轻松作答：

（a）谁花费7年时间？ -- ”女子“ （agent）

（b）为的是什么？ -- ”证明“ （purpose）

（c）证明什么？ -- ”是“ （content）

（c-1） -- ”自己“ （relevant）

(C-2) -- “女性“ （isa）

（d）花费多长时间？ -- ”7年“ （duration）

与深度学习结合的思考

- 用概念或者义原代替关键字
- 降低维度
- 减少歧义
- 不可控成为可控

谢谢大家

欢迎提出

宝贵的意见与建议！