

中国国家973课题“文本内容理解的数据基础”（2004CB318102）

第二届 机器学习 及其应用研讨会（2006年11月3日-5日，南京）

文学语言的机器理解研究

俞士汶 王治敏 曲维光（讲者）

北京大学计算语言学研究

Email: yusw@pku.edu.cn

2006年11月5日，南京大学

主要内容

- 计算语言学的研究目标
- 自然语言处理的主攻方向
- 丰富多彩的文学语言
- 隐喻的机器识别及求解初探
- 结语与致谢

主要内容

- 计算语言学的研究目标
- 自然语言处理的主攻方向
- 丰富多彩的文学语言
- 隐喻的机器识别及求解初探
- 结语与致谢

关于学科定位

《学科分类与代码表》（中国国家标准 GB/T13745--92 ）
语言学

——应用语言学

——计算语言学（740.3550）

（三级学科“计算语言学”属于文科）

计算机科学技术

——人工智能

——自然语言处理（520.2020）

（三级学科“自然语言处理”属于理工科）

两个三级学科的研究对象与内容基本相同，
新兴交叉学科在学科体系中已有一席之地。

（俞士汶 主编《计算语言学概论》，商务印书馆，2003年）



不同的术语与所指

计算语言学 (Computational Linguistics, CL)

自然语言处理 (Natural Language Processing, NLP)

自然语言理解——最高境界

语言信息处理

人类语言技术 (Human Language Technology, HLT)

中文信息处理 (Chinese Information Processing, CIP)

——汉字信息处理

——汉语信息处理

(Chinese Information Processing, CIP)

(大致相同，又各有侧重。)



“自然语言处理” 在研究什么？

- (1) 人工系统的自然语言界面
- (2) 机器（辅助）翻译
- (3) （深度）信息检索与信息提取
- (4) 文本分类、摘要、述评及术语与领域知识管理
- (5) 词典计算机辅助编纂
- (6) 面向语言本体研究与教学研究的应用

.....

对象与工具

字符信息（数据） —》语言信息（知识）

表现形式：字符串 —》词、句子、篇章

需要相应的理论、方法与技术，

需要对“语言”、“语言信息”、“计算语言学”以及
“自然语言处理” 有个概括的了解。

关于“语言”

英国《新科学家》周刊 2005年4月9日 的文章

——生命进化的十大奇迹：脑（第3项）和语言（第4项）

脑常常被视作进化过程中的最高成就，因为它赋予了人类一些高级特征，例如 语言、智慧、意识。语言是进化的终极发明。在令人类区别于动物的特征中，语言处于核心地位。语言也许称得上是人类的决定性特征之一。我们的祖先如何实现了语言从无到有的飞跃，这也许是科学史上最大的谜。语言是生物进化的最后一笔。这是因为语言令那些掌握了它的动物超越了纯生物的范畴。

语言系统是动物进化到人的两大标志之一。
语言能力可否作为机器学习的重要指标？



关于“自然语言处理”

语言信息处理是数值计算机

在非数值领域最早的应用（**MT, Turing**）。

语言学对计算机科学也有重要贡献（**Chomsky**）。

自然语言理解又特别困难：

- （1）依据对人类语言机制的认识
- （2）语言既是对象，又是工具
- （3）依据对当代计算机能力的认识
- （4）依据NLP技术发展的历史经验

汉语理解研究和其他语言一样困难，
汉语信息处理技术又有特殊的课题。
以汉语为母语的学者还有其独特优势。



关于“计算语言学”

为语言信息处理提供理论模型、实现算法、工程方案。

语言模型：

实际问题太复杂，需要根据应用的需要，进行简化、变换，使其成为可计算的形式，这就是模型化。

上下文无关语法就是一种语言模型，便于分析和生成符合规则的句子，可以覆盖相当一部分自然语言句型。

向量空间模型可以刻画文本的主要特征，可用于信息检索、文本分类。

常用算法：

基于规则的方法（词法、句法、语义、语用）

基于统计的方法（原始语料—加工了的语料）

相互融合，很可能殊途同归。

机器学习的重要地位（搭配、消歧、…）



回到学科定位



《学科分类与代码表》（中国国家标准 GB/T13745--92 ）

计算机科学技术

——人工智能里

——自然语言处理（520.2020）

（三级学科“自然语言处理”属于理工科）

语言学

——应用语言学

——计算语言学（740.3550）

（三级学科“计算语言学”属于文科）

文学

——

——

（文学与语言学是近邻。文学与信息科技无缘？如何相互促进？）



交 流

2002夏——两岸学者在北京大学酝酿

2003年12月在台湾元智大学成功召开
第一届“文学与资讯科技研讨会”

2004年3月罗凤珠主编

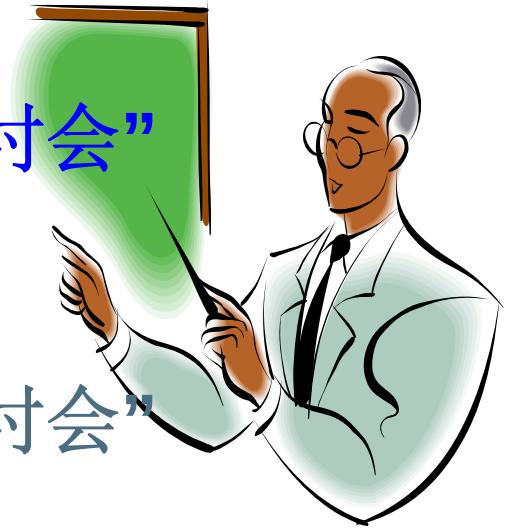
《语言，文学与资讯》在台湾出版

2005年12月8日-10日

第二届“文学与信息科技国际研讨会”
在北京大学召开

2007年3月

第三届“文学与信息科技国际研讨会”
将在日本东京学艺大学召开



主要内容

- 计算语言学的研究目标
- 自然语言处理的主攻方向
- 丰富多彩的文学语言
- 隐喻的机器识别及求解初探
- 结语与致谢

自然语言理解困难——实例之一

关于自动升降晾衣架的对话

妻子：“嘿，过了一年才坏。”

丈夫：“什么呀，才一年就坏了。”

丈夫理解了妻子的意思吗？

——虚词词义：才（数量词前后，意义不同）

——背景知识：保修期

——知识激活机制？

自然语言理解困难——实例之二

关于“沙漠化”的文章

“几年前由于种植籽瓜有利可图，使大批的种植者就到过渡带来开垦，……。

在这样的绿洲和沙漠过渡带开垦，极易造成风蚀。”

——<今日民航>2001年9月号

就/ 到/ 就到/ 到/ 到过/ 过/ 过渡/ 带/ 来/ 带来/

（未登录词的识别——知识背景——认知机制）

自然语言处理主攻方向——歧义消解

词语切分：白天鹅

可能的切分：白天鹅/---白/ 天鹅/---白天/ 鹅/---白/ 天/ 鹅/

计算机程序可以按某种算法实现这种切分，给出一种或多种结果。对否？

白天鹅飞过来了——白/ 天鹅/ 飞/ 过来/ 了

白天鹅可以看家——白天/ 鹅/ 可以/ 看/ 家/

白天鹅在湖里游泳——白/ 天鹅/ ？ 白天/ 鹅/ ？

读音不同的词的词性标注：只—量词 q [zhi1] /副词 d[zhi3]？

这只会测水温的鸭子

——这/ 只/ 会/ 测/ 水温/ 的/ 鸭子/ （切分无歧义）

——这/r 只/q 会/v 测/v 水温/n 的/u 鸭子/n ， 挺有用的

——这/r 只/d 会/v 测/v 水温/n 的/u 鸭子/n ， 没什么用

（意义决定词性，还是词性决定意义？）

自然语言处理主攻方向——歧义消解

读音相同的“连”也有不同的词性（意义）：

一个连有三个排——“连”是名词 **n**

我们兄弟心连心——“连”是动词 **v**

苹果可以连皮吃——“连”是介词 **p**

短语结构的歧义：**m + q + n + “的” + n**

三个大学的老师 三/**m** 个/**q** 大学/**n** 的/**u** 老师/**n**

——[[三/**m** 个/**q** 大学/**n**] 的/**u** 老师/**n**]

——[三/**m** 个/**q** [大学/**n** 的/**u** 老师/**n**]]

三所大学的老师——[[三/**m** 所/**q** 大学/**n**] 的/**u** 老师/**n**]

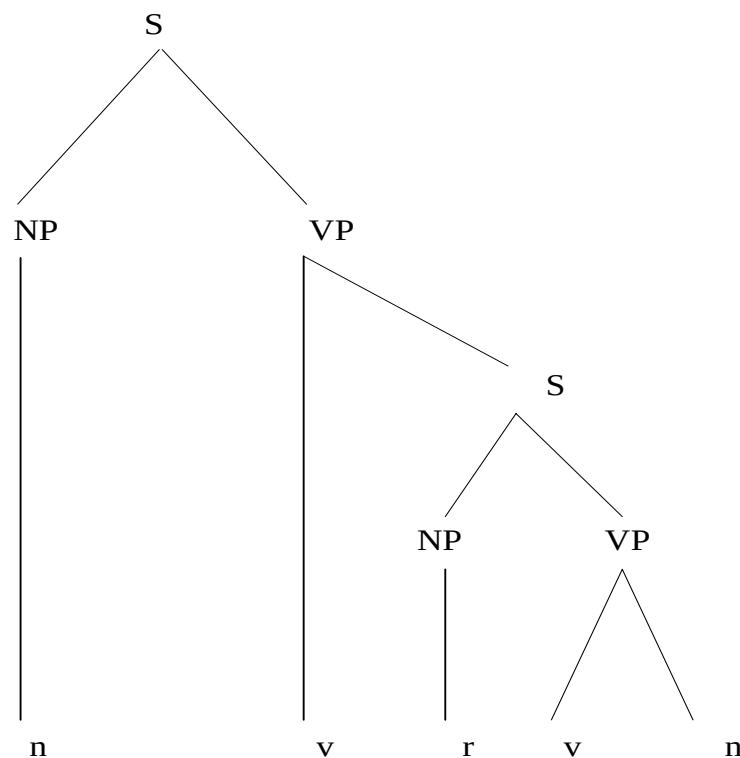
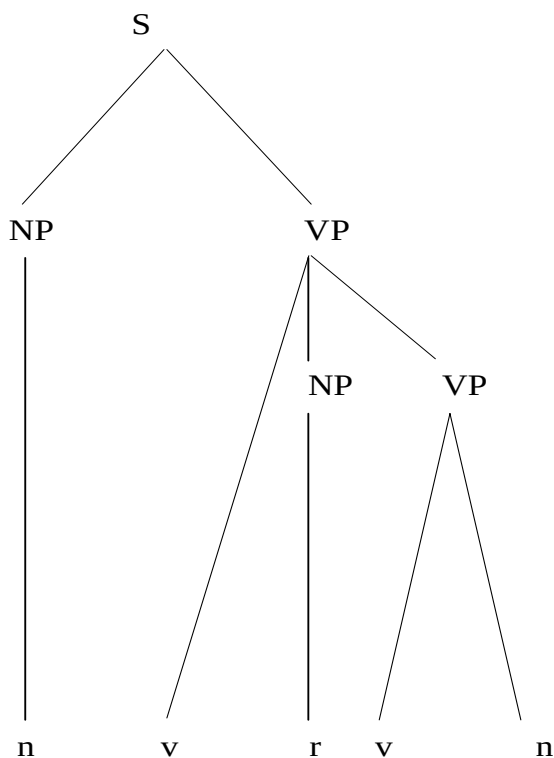
三位大学的老师——[三/**m** 位/**q** [大学/**n** 的/**u** 老师/**n**]]

句法结构的歧义

例1 会员 选举 他 当 主席

例2 学生 认为 他 是 校长

n v r v n (切分、标注无歧义)



句法结构（树）不同

语义歧义以及依赖语境的歧义消解

汉语语义分析（切分、标注、句法分析都无歧义）

熊猫/n 吃/v 竹笋/n

学生/n 吃/v 食堂/n

民工/n 吃/v 大碗/n

老师/n 写/v 毛笔/n

汉语语义指向分析

写/v 好/a 了/u （文章）

写/v 累/a 了/u （老师）

写/v 秃/a 了/u （毛笔）

汉语语境分析

小张/n 打针/v 去/v 了/u （小张是护士？病人？）

其他：长句与句号、逗号

新一届测绘学名词审定委员会的主要特点是年青化，吸收了一些工作在教学、科研前沿的青年专家学者，充分发挥他们接触新知识多，对名词工作热情高、活力大的特长，同中老年专家共同做好新一届委员会的名词审定工作。

（长句子，一逗到底）

其他：长句与句号、逗号

1. 你得藏在一个你看得见他，可是他看不见你的地方。（逗号断开了结构）
2. 车臣武装分子和世界其他地区的恐怖分子是一丘之貉，应该合力打击他们。（分句主宾语不明确）

其他：指代与省略

小明要求他爸爸给他弟弟买一件
他喜欢的衣服，他同意了。

(4个“他”，各指谁？)

重庆队得**88**分，客场负于台湾队**2**分。

(**CBA**, 台湾队和重庆队各得多少分？比赛地点？)

其他：时态、语态、语气

我在家里。(be)

我在家里看书。(in)

我在看书。(-ing)

你在干什么？——看书。

你喜欢干什么？——看书。

如果我是你，我就去了。

如果我有时间，我就去。

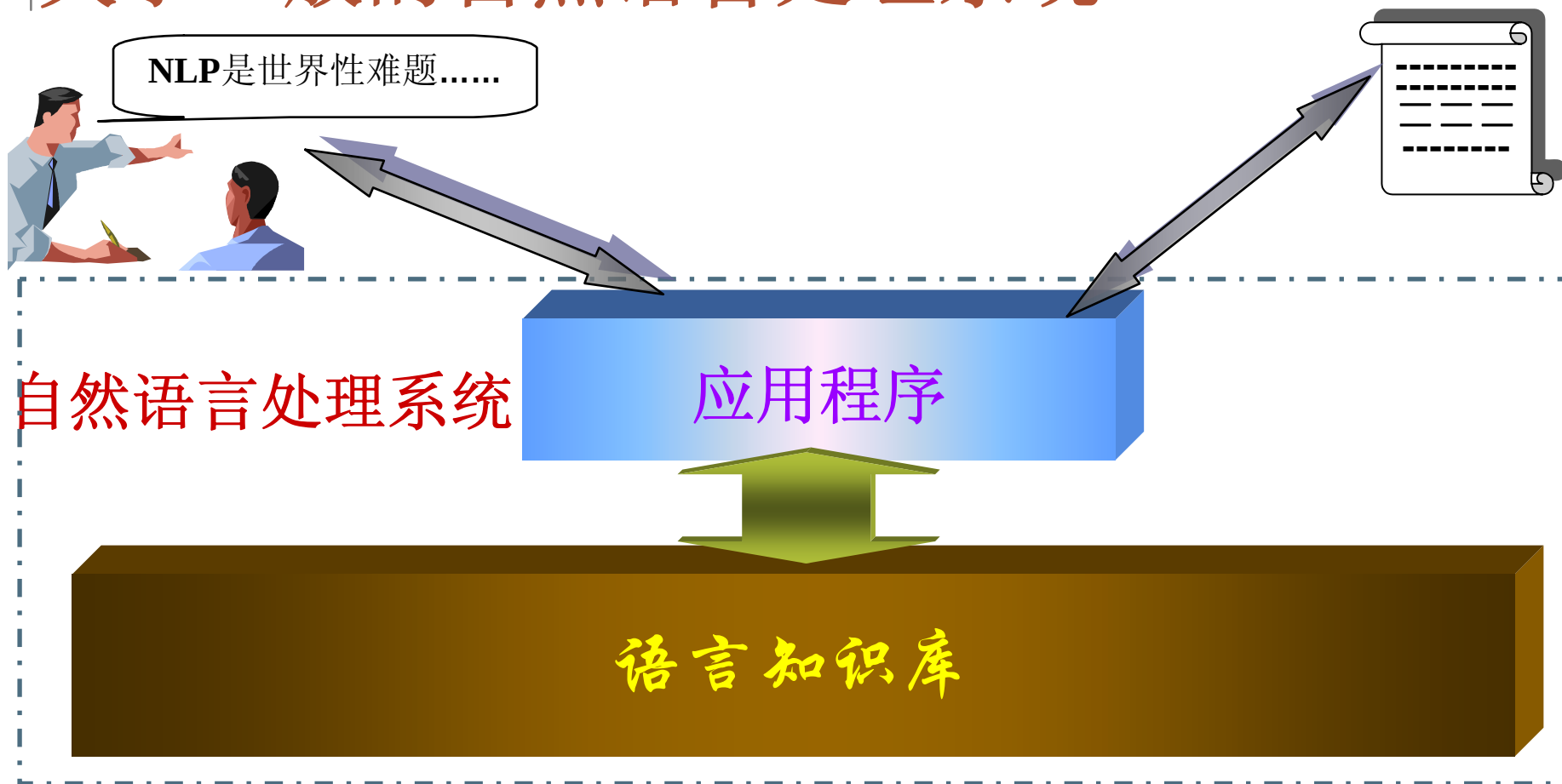
书面汉语特点及其对信息处理的影响

吕叔湘：“有了形态变化，语法分析就比较容易进行。没有严格的形态变化，在语法分析上就比较容易引起问题。”

汉语缺乏形态变化，缺乏形式标记，自动分析也就缺少可以把握的线索。汉语自动分析如果不比其他的语言更困难，至少不会比其他的语言更容易。

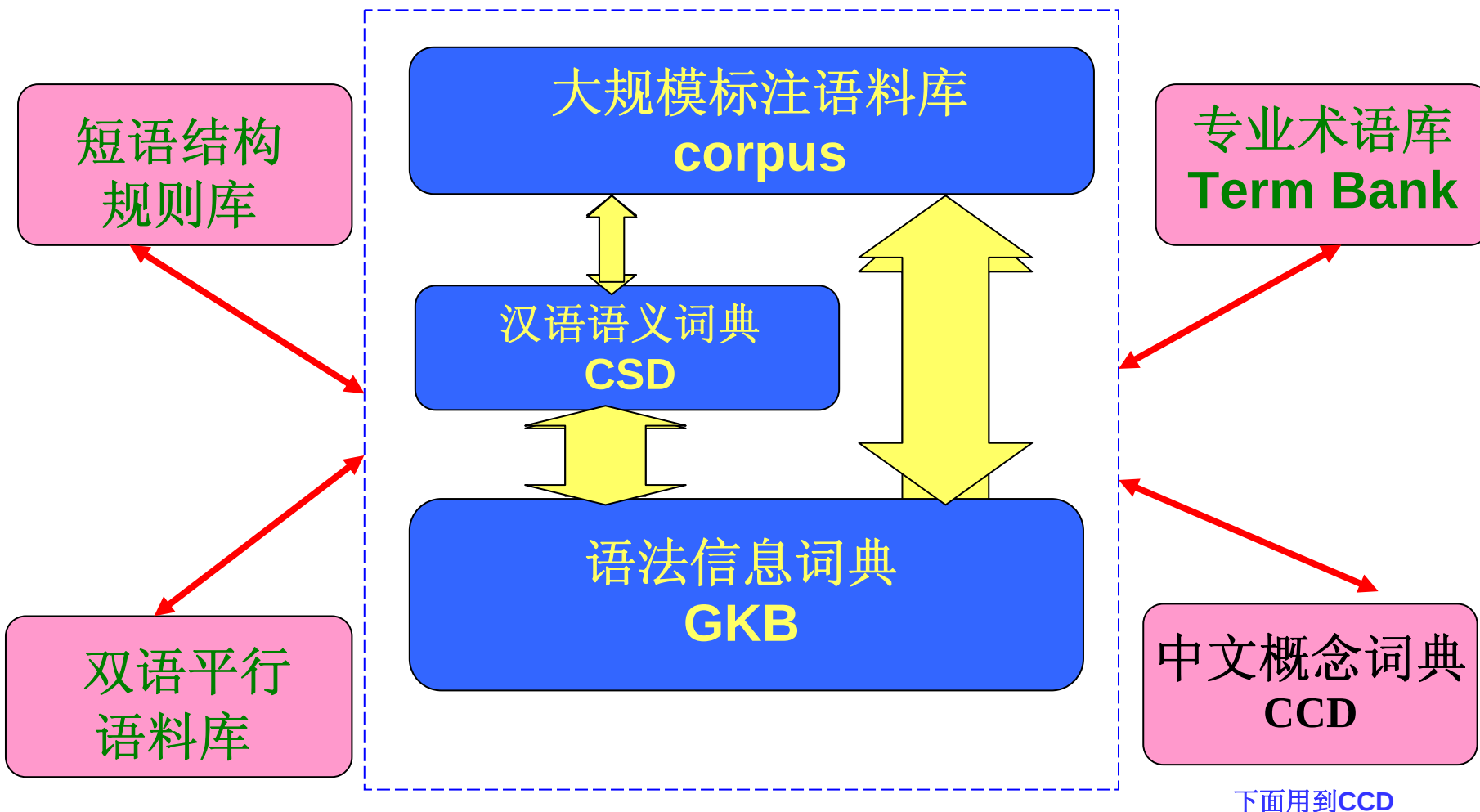
汉语信息处理尤其需要大规模的高质量的语言知识库的支持。

关于一般的自然语言处理系统



语言知识库是自然语言处理系统不可或缺的组成部分，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败。面向自然语言处理的语言知识库对语言本体研究和语言教学也有重要意义。

北大语言数据资源之概要



下面用到CCD

主要内容

- 计算语言学的研究目标
- 自然语言处理的主攻方向
- 丰富多彩的文学语言
- 隐喻的机器识别及求解初探
- 结语与致谢

影响自然语言理解的文学表现手法

“自然语言是规则系统加噪声”

噪声固然要设法剔除，但如何界定噪声？

以往实用的自然语言处理研究似乎有个约定，大家都不碰文学作品，特别是诗歌。然而，文学作品反映形象思维，表现要生动，也要含蓄，要标新立异，常常采用隐喻、影射、夸张、拟人、典故、双关 … …等表现手法，诗歌还要照顾韵律、节奏、效果等因素，由此造成的真实文本往往有不合语法、违反常识的现象，这些算不算噪声，如何识别、应对、理解？

思维分逻辑思维、形象思维、顿悟等多种类型。要达到自然语言理解的境界，不可能把形象思维排斥在视野之外。

1 隐喻与影射

“歧义消解”已使研究者疲惫不堪，但要走上自然语言理解的坦途，“隐喻”是必须逾越的路障。

隐喻和歧义的区别

歧义是同中辨异。

“那男人简直是条狼”就是隐喻说法，违反常识。

人理解隐喻，有能力异中求同，机器自然费解。隐喻符合人的认知机制，说话人利用了“男人”和“狼”的某种共同属性，比直接说“男人如何如何”的表现力要丰富得多，听话人也能理解说话人想讲什么。类似的，“董永是木头”、“郎平是铁榔头”也都是隐喻。

但在不同语境中使用这句话也会传达不同的信息。符合常识的某种语境中说的“男人都是动物”也能表达同样意思。

1 隐喻与影射

隐喻的层级划分：词汇级、语句级和篇章级。

词汇级——“山顶”、“桌子腿”、“无底洞”、“瓜分”、“露马脚”、“指桑骂槐”

（求解依赖于词汇知识库，没有特殊之处）

语句级——“铁榔头又立了新功”、“知识的海洋”、“幻想是诗人的翅膀”

（求解要有新招，如：隐喻知识库，逻辑推理，机器学习）

（“幻想是诗人的翅膀”在另一种意义上又可分层级？）

篇章级——《狂人日记》、《春怨》

（求解？与“影射”、“寓意”、“弦外之音”等的关系）

1 隐喻与影射

影射可否看作篇章隐喻的范畴？或许是隐喻的最高境界。

弦外之音与含沙射影

孟浩然（临洞庭上张丞相）

八月湖水平，涵虚混太清。气蒸云梦泽，波撼岳阳城。

欲济无舟楫，端居耻圣明。**坐观垂钓者，徒有羡鱼情。**

（一幅山水画：湖面开阔，烟霭缭绕，垂钓者怡然自得，令人生羡。自己要有一根鱼杆，也可一显身手。言外之意——“垂钓者”喻在位的人。“羡鱼情”喻当官的愿望。请张丞相给予机会。有求于人，但仍不失风度。虚伪否？）。

朱庆余（宫中词）

寂寂花开闭院门，美人相并立琼轩。

含情欲说宫中事，**鹦鹉前头不敢言。**

（想象两个宫女想讲讲贴己话，又恐鹦鹉学舌，洩露机密。或许有更深一层的含义，表现皇宫内院的压抑、沉闷、甚至恐怖，人人自危）。

1 隐喻与影射

触景生情与文化背景

王昌龄《闺怨》

闺中少妇不知愁，春日凝妆上翠楼。

忽见陌头杨柳色，悔教夫婿觅封侯。（杨柳——后悔？）

张九龄

纤纤折杨柳，持此寄情人，一枝何足贵，怜是故园春。

李瑞

赠君折杨柳，颜色岂能久，上客莫沾巾，佳人正回首。

《遠離別》 昨日盧梅塞口，整見諸人鎮守。

都護三年不歸，折盡江邊楊柳。

《送別》 杨柳东门树，青青夹御河。近来攀折苦，应为别离多。

诗经《采薇》

昔我往矣，楊柳依依，今我来思，雨雪霏霏

2 引用典故

韦庄《章台夜思》：

清瑟怨遥夜，绕弦风雨哀。

孤灯闻楚角，残月下章台。

芳草已云暮，故人殊未来。

乡书不可寄，秋雁又南回。

杜牧《赤壁》：

折戟沉沙铁未消，自将磨洗认前朝。

东风不与周郎便，铜雀春深锁二乔。

3 遣词造句的形象化

使用形象化的量词：一钩新月

雪一程，风一程，灾区雪夜千帐灯。

词性的变通使用：晓镜但愁云鬓改（李商隐）

春风又绿江南岸（王安石）

韵律影响语序，造成不合语法、语义的词序：

帘卷西风，人比黄花瘦。（李清照）

六亿神州尽舜尧，春风杨柳万千条。（毛泽东）

拟声、拟态词：无边落木萧萧下，不尽长江滚滚来

离离原上草，一岁一枯荣。野火烧不尽，春风吹又生。

远芳侵古道，晴翠接荒城。又送王孙去，萋萋满别情。

正意反说与褒贬色彩转换：

岭外音书绝，经冬复立春。近乡情更怯，不敢问来人。

4 夸张

李白 《将进酒》：

君不见黄河之水**天上来**，
奔流到海不复回。

（空间夸张）

君不见高堂明镜悲白发，
朝如青丝暮成雪。

（时间夸张）

毛泽东 《六盘山》：

惊回首，
离天三尺三。

5 双关——语义双关

曲波《林海雪原》：

可是匪徒们走上几十里的大山背，他们没想到包马脚的麻袋片全烂掉在马路上，露出了他们的马脚。

曹雪芹《红楼梦》——旁敲侧击，指桑骂槐：

贾宝玉说薛宝钗像杨贵妃。薛宝钗甚怒，又不便发作。恰有小丫环前来说宝姑娘藏了她的扇子，要讨回扇子。宝钗指她道：“你要仔细！我和你玩过，你再疑我。和你素日嬉皮笑脸的那些姑娘们跟前，你该问她们去。”

5 双关——谐音双关

“东边日头西边雨，道是无晴却有晴（情）”

许多歇后语就是利用谐音双关构成的。例如：

- 学生多四眼，勤读成进士（近视）。
- 老虎拉车——谁赶（敢）？
- 孔夫子搬家——净是书（输）。
- 窗户眼吹喇叭——鸣声（名声）在外。
- 歪嘴吹喇叭——斜气（邪气）

6 幽默

牙医拿着拔牙的钳子，对坐在椅子上张着只剩下最后一颗牙的病人，信誓旦旦地说：“这一次绝对不会拔错了。”

现在已经有“计算幽默”这样的研究。

7 拟人化

寓言故事

《狼和小羊》

喻指以强凌弱、要做坏事的人总是可以找到借口的。

《狐狸与葡萄》

“狐狸吃不到葡萄说葡萄酸”故事浓缩成了这句话，成为“自我解嘲”的隐喻说法。

《乌鸦与狐狸》

警示人们不可得意忘形，警惕吹捧你的人可能是另有所图。

文学表现手法并非只见于文学作品

半夜，孩子哭。爸爸恐吵醒邻居，唱摇篮曲哄之：“小燕子，穿花衣，年年春天来这里，… …”。不一会儿，传来邻居的喊声：“你还是让孩子哭吧！”

（弦外之音：唱得比哭还难听。也要联系什么年代和居住条件）

理发店一语双关的好招牌：顶上功夫

天大药业的广告词：您的健康是天大的事。

汽车后玻璃上含蓄地写着：别吻我，我怕羞（修）

主要内容

- 计算语言学的研究目标
- 自然语言处理的主攻方向
- 丰富多彩的文学语言
- 隐喻的机器识别及求解初探
- 结语与致谢

隐喻的机器识别及求解初探

隐喻计算研究的内容：

(1) 隐喻识别

“知识的海洋”

(2) 隐喻理解

知识像海洋一样丰富

(3) 隐喻生成

例如，“信息的海洋”

作为起步，首先进行隐喻识别研究。

隐喻的机器识别及求解初探

973课题“文本内容理解的数据基础”

(课题编号：2004CB318102，2004年9月起)

- 面向跨语言信息处理的多语概念词典 **MCD**
- 现代汉语词义标注语料库 **WST**
- 现代汉语词语语法属性的概率描述
- 成语知识库建设
- 汉语文本中的隐喻计算研究
- 应用研究探索

隐喻的机器识别及求解初探

2003年9月——2006年6月

博士生王治敏完成博士论文

“汉语名词短语隐喻识别研究”

同期姜柄圭博士论文也分析了科技专著中的隐喻

2006年1月——2007年6月

博士后曲维光前后衔接，深入研究

2006年9月——2010年6月

博士生贾玉祥拓展研究范围：语句层次和篇章层次的隐喻形态，动词隐喻研究，集成名词隐喻研究成果。同时探索在（深度）信息检索等相关领域的应用。

王治敏博士论文概要

——汉语名词短语隐喻识别研究

即便只定位于识别，也不宜全面铺开。

作为第一次系统的探索，只考虑如何识别由
名词n和名词n

组合而构成的短语结构中所包含的隐喻表达，也就是辨别
“nn、n的n、n是n”

形式的短语结构“n + n”是不是隐喻。

希望通过对这样一种短语结构的考察，由点到面，探索更一般的隐喻识别问题。

经常作为隐喻“源域”的部分名词例词

主线	种子	支柱	钥匙	雨露	阴云	摇篮	阳光	烟水
漩涡	序幕	星	峡谷	舞台	微波	土壤	隧道	曙色
食粮	森林	热浪	桥梁	旗帜	喷泉	纽带	泥潭	泥坑
门槛	脉络	脉搏	绿洲	路线	龙头	灵魂	浪潮	精髓
枷锁	脊梁	激流	基石	火炬	火花	火	灰尘	花
河流	河	航船	海洋	瑰宝	光线	光辉	光	痼疾

以“海洋”为例

➤ 隐喻表达

知识的海洋

信息的海洋

➤ 非隐喻表达

国家的海洋资源

一片蓝色的海洋

- 词语的隐喻表达与其共现的词语密切相关。因此，要识别一个词是否为隐喻表达必须依赖它所在的上下文。
这个问题与一个短语结构的隐喻表达实际上是一回事。

从统计的观点看

- 隐喻识别问题转化为计算在上下文的条件下含指定名词“海洋”的名词短语是隐喻的概率。需要在对大量语料进行统计实验的基础上对含“海洋”的短语是隐喻的概率分布做出合理的估计。
- 进一步把名词短语的隐喻识别问题转换为词语的隐喻义与字面义的分类问题。

机器学习的方法（实际只是ABC的应用）

- 对一定数量的语料进行标注，把隐喻和非隐喻分开，取得第一手统计数据。
- 利用标注好的语料（有指导的学习），采用
 - 基于实例的方法
 - 最大熵方法
 - 朴素贝叶斯方法估计隐喻的条件概率分布（需要给出在什么样的上下文条件下，词语是隐喻用法的概率）

汉语名词短语隐喻标注语料样例

- 19980123-11-003-008/m 握/v 着/u 这样/r 的/u 手/n , /w 能/v 不/d 倍感/v 温暖/a 和/c 鼓舞/a ? /w 像/p 他们/r 这个/r 年龄/n 的/u 小伙子/n , /w 有/v 多少/r 还/d 依偎/v 在/p 父母/n 的/u 身旁/s , /w 享受/v 着/u 温馨/a 的/u 抚爱/vn ; /w 有/v 多少/r 正/d 坐/v 在/p 宽敞/a 明亮/a 的/u 教室/n , /w 遨游/v 于/p <知识/n 的/u 海洋/n>_ 。 /w 而/c 他们/r , /w 却/d 走/v 进/v 了/u 绿色/n 的/u 军营/n , /w 来到/v 了/u 北大荒/ns , /w 用/p 青春/n 和/c 血汗/n , /w 为/p 军队/n 为/p 国家/n 创造/v 巨大/a 的/u 财富/n 。 /w
- 19980303-05-003-003/m 与会/vn 同志/n 认真/ad 回顾/v 1997年/t 全军/n 文艺/n 工作/vn 后/f 认为/v , /w 军队/n 文艺/n 创作/vn 的/u 根本/a 任务/n 是/v 贯彻/v 党/n 的/u 十五大/j 精神/n , /w 高举/v 邓小平理论/n 伟大/a 旗帜/n , /w 站/v 在/p <时代/n 高度/n> , /w 积极/ad 弘扬/v 主旋律/n , /w 为/p 部队/n 和/c 社会/n 提供/v 更/d 多/a 更/d 好/a 的/u <精神/n 食粮/n> ; /w 加强/v 文艺/n 创作/vn 要/v 从/p 深入/v 生活/vn 抓起/v , /w 把握/v <时代/n 脉搏/n> , /w 使/v 作品/n 更/d 具有/v 时代/n <生活/vn 气息/n> ; /w 要/v 进一步/d 树立/v 精品/n 意识/n , /w 把/p 主要/b 精力/n 放/v 到/v 军事/n 题材/n 创作/vn 上/f , /w 在/p 艺术/n 上/f 进行/v 新/a 的/u 探索/vn 和/c 表现/vn , /w 开创/v 崭新/b 的/u 多样化/vn 创作/vn 局面/n ; /w 全面/ad 提高/v 文艺/n 队伍/n 的/u 整体/n 素质/n , /w 使/v 我们/r 的/u 队伍/n 成为/v 一/m 支/q 政治/n 强/a 、 /w 业务/n 精/a 、 /w 作风/n 硬/a 、 /w 特别/d 能/v 战斗/v 的/u 队伍/n 。 /w

特征的选择

- 训练模型时使用两类特征：简单特征和辅助特征。
- ✓ 简单特征就是“海洋”出现的上下文中的词语及其词性
- ✓ 辅助特征包括左右位置信息、标点、停用词等信息。
- 以左右位置信息为例：

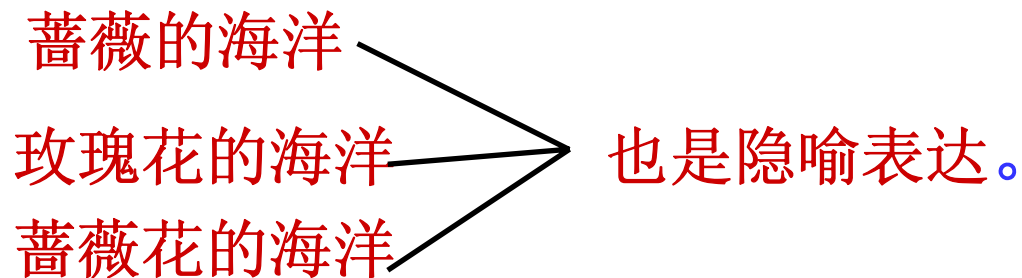
左	右
在知识的海洋中遨游	要利用海洋资源
在科学的海洋中	海洋科学探索
- 一个名词处于“海洋”左右的不同位置，形成了隐喻和非隐喻两种表达。位于“海洋”的左侧，通常为隐喻表达；位于“海洋”的右侧，通常不是隐喻表达。因此把左右位置信息也选作特征，加入训练模型。

三种模型和窗口选择的试验

- 针对训练语料，通过对3种模型采用不同大小窗口的实验比较，确定了最大熵方法是相对理想的模型。
- 然后扩大语料规模，对未加标注的语料中包含指定词（如“海洋”）的“n+n”短语隐喻进行识别。

“n+n”名词短语隐喻识别的相似推理。

假定“玫瑰的海洋”是个隐喻表达。可以根据这个原型推断出



这是因为在人的知识中，“玫瑰、蔷薇、玫瑰花、蔷薇花”都是花，具有“花”的共同特征，因而可以进行替换。

CCD词典为基于“同义或近义”的相似推理提供了基础知识

在 CCD 中概念用同义词集表示， CCD 有这样的概念节点。

- 玫瑰 [20_08920236] [20]植物
- 蔷薇 [20_08920236] [20]植物
- 玫瑰花 [20_08920236] [20]植物
- 蔷薇花 [20_08920236] [20]植物

玫瑰、蔷薇、玫瑰花、蔷薇花同属节点[20_08920236]，节点[20]植物是其上位概念。

- 同义、近义、上下位知识正好可以为计算机基于相似性处理隐喻的的扩展识别提供重要的资源。
- 在识别实验的过程中，建立了汉语隐喻知识库，该知识库又反作用于识别算法。

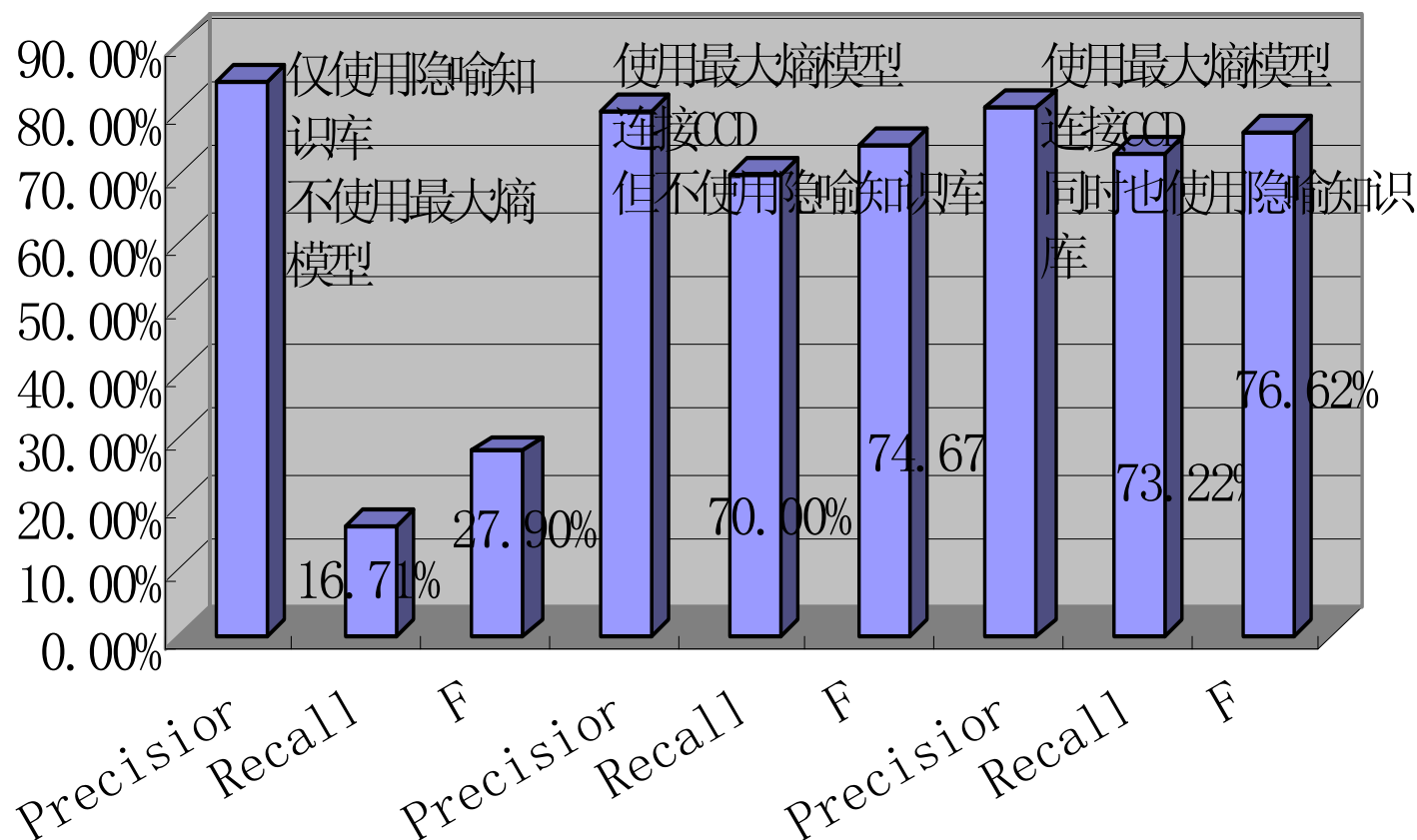
汉语名词隐喻知识库样例

词语	搭配词	目标域	标记词	隐喻类型	CCD映射	同形	评价
海洋		知识/时间/人生/歌/花/舞/笑声/灯/市场/经济/光/科学/知识/彩旗/大赛/生活	的		是		
大战		非典/洪水/挂历/羊绒衫/鳗苗/储蓄/利率/啤酒/价格/贸易/点球			是		
道路	健康	谈判/社会主义/和平/人生/革命/资本主义/致富	的		是		
堤坝		旧观念/心灵/理智/心/感情	的		是		
地狱		人间					消极

源域“海洋”的概念映射分库样例

词语	词性	频次	offset	type	csynset	cdescription	cnote
知识		16	03_00013243	心理特征 精神特征	学识 学问 知识 学力 见闻 认知 认识	通过感觉学习和推理得到的心理成果；	从书本上可以获得许多知识；
时间		2	03_00015594	抽象概念 抽象观念 抽象概括	光阴 岁月 时间 工夫 年光 时光 时日 时期 期间 辰光 蹉跎岁月 春光	事物存在或继续的期间，从过去到现在再到未来	我们应该珍惜时间；
人生		3	28_10868596	time	人生 一生 一辈 子	某人的一个特定生活阶段的一段时间	X

隐喻相似度推理+隐喻知识库测试结果



结语与致谢

“路漫漫其修远兮，吾将上下而求索”

“千里之行，始于足下”

“聚沙成塔，滴水穿石”

谢谢大家，特别感谢王珏教授、周志华教授。

ICL/PKU内外的专家、同事、朋友、学生 都给予过指教与帮助，有所贡献。以往发表的文章中已经具名致谢，这里恕不赘述。

欢迎访问北大计算语言学研究所和北大软微学院语言信息工程系

- <http://icl.pku.edu.cn/>
- <http://www.ss.pku.edu.cn/>
- <http://www.ChineseNLP.com/>