常识元基础构建的比喻系统对自然语言生成的可行性研究

南京大学 计算机科学与技术系 崔博涵

摘要

时下自然语言为人工智能领域发展的主要方向,常识元分析作为对于语义分析的有力工具，自然可以作用于自然语言生成。本文基于对若干概念的常识元分析的背景研究，构建出一个小型的比喻系统作为样本，并尝试给出该小型比喻系统用于自然语言生成的可行性分析。

关键词：文化语义学；自然语言生成；人工智能；比喻

一、 研究背景

日前，人工智能模型 ChatGPT 引发关注，以深度学习与神经网络算法为核心的自然语言（包括自然语言处理、自然语言生成等）人工智能算法是当下研究热门。作为自然语言生成算法，其重要特质是输出语句需满足人类语言的正常语法和语义，而作为一种语义分析的方法，常识元分析有助于我们深入理解一个“合理语句”[1] 需要哪些必备的常识要件，因而可能通过常识元分析作为一部分构建出一个自然语言输出系统。修辞是自然语言中一个重要组成，因此基于在常识元分析中得出的关于比喻中的常识元规律，本文尝试构建出一个小型的比喻系统。

二、 常识元分析的背景学习

（一）常识元分析的基本方法

在个人实践中，采用了以下程序：

1. 预处理：采样+统一剔除+人工快速复核

（1） 采用系统抽样法抽样，并分析其中的无效条目

（2） 按照其上下文特征在全部样本中删除相应的语料

（3） 人工快速复核

2. 筛选与数据处理

（1） 寻找基本特征

[1] 合理语句，即被认为是合乎人类语言规律的语句，更严格的说是可以通过一次图灵测试的语句。
图灵测试，英国数学家阿兰·图灵提出的测试方法，即：人类提出问题，如果不能仅凭回答内容不能判断回答者为机器/人,即认为通过了图灵测试。

（2） 筛选

（3） 在筛选结果中再寻找具有符合常识元的再筛选）

（4） 重复上述步骤

3. 统计，寻找规律

（二）常识元分析实践——以“歌”为例

结果如下：

常识元 1-1：歌可以被演唱。
常识元 1-2：歌可以具有某些特征。
常识元 1-2-1：歌可以表达情感/态度，带给人某种感受。
常识元 1-2-2：歌可以具有某种主题。
常识元 1-2-3：歌可以具有归属性（来源或演唱者）和辨识度
常识元 1-2-4：歌可以具有可保存/流传性。
常识元 1-2-5：歌有对应的语言。
常识元 1-2-6：歌可以持续任意时间，具有连续性。
常识元 1-2-7：歌声音、气势有大小之分。
常识元 1-2-8：歌有音色之分。
常识元 1-3：歌可以被计数。
常识元 1-4：歌可以被听。
常识元 1-5：歌声会传播。

（三）关于比喻的规律发现

“歌”作为本体的喻体统计：

风	5
鸟叫	4
水流	3
雷	2
雨	2
图画	2
鱼	2
烟雾	2
雪崩	1
巨浪	1
蜜	1
搔痒	1
锤子	1
尖叫	1
老朋友	1
酒	1
光亮	1
天气	1
银铃	1
钟声	1
钥匙	1
泪	1
铜号	1
喊叫	1
麻醉剂	1
树木	1
舞	1
精灵	1

与比喻相关的两条规律如下：

1. 典型喻体与非典型喻体

   典型喻体（出现频次高的喻体）往往体现多个常识元，而非典型喻体往往只体现一个常识元。

2. 环境性

   典型喻体的常识元往往与其所处的语言环境有关。

   eg. 迟子建/逝川：她家的木屋顶晾制干菜时要唱，在傍晚给家禽喂食时也要唱。吉喜的【歌】声【像】炊烟一样在阿甲渔村四处弥漫，男人们听到她的歌声就像是听到了泪鱼……

   eg. 莫言/黑沙滩：他自称“老兵”，实际上只比我们早入伍一年，一副浪荡样子。【歌】声【像】泥鳅般地从他嘴里滑出来：黑沙滩云满天黑沙滩的大兵好心酸黑沙滩的……

值得注意的是，在“环境性”一条中，“所处的语言环境”可以由所在句中的一些关键词具体体现；所谓“有关”，具体是指该典型喻体所体现的常识元与**上下文中出现关键词**或**本体**对应概念的常识元相近或相同。因此，该喻体用于这个比喻句才符合语言环境的需要。

这一规律是比喻系统构建的主要理论根据。

三、 比喻系统的构建

（一）方法论

由于上述的规律，我们可以看出，我们可以将一个本体通过常识元连接到喻体上，而喻体亦可作为本体，通过同样的手段连接到其他概念上。这样，我们可以对常见概念做常识元分析，以常识元作为边，概念作为节点，交错成为网状结构，就构成了一个有向连通图（其中 A 到 B 的有向线段（标签为“…”）代表“A 作为本体因常识元‘…’而被比作喻体 B”），成为一个比喻系统。

为简化模型，此处只考察本体喻体关系，不考虑上下文关键词。

（二）比喻系统构建的实践

1. 常识元分析（只撷取较为典型的用于比喻的常识元，并非这些概念的全部常识元）

水

1.1 水能反光

1.2 水（清澈的）透明

1.3 水可以很凉（冷却）

1.4 水能流动

眼睛

2.1 眼睛可以用来看

2.2 眼睛能反光，部分透明

镜子

3.1 镜子可以反光

宝石

4.1 宝石多数透明

4.2 宝石有光泽，能反光

4.3 宝石多数有鲜艳颜色

火

5.1 火是红色的

5.2 火能传播

5.3 火有流动堆叠的样态

绸缎

6.1 绸缎可堆叠，有柔软的质地

6.2 绸缎有光泽

刀

7.1 刀是冰冷的

7.2 刀可以用来切割 7.3 刀可以反光

钢铁 8.1 钢铁能反光

风 9.1 风是流动的 9.2 风可以带来寒冷感 9.3 风会给人带来触感

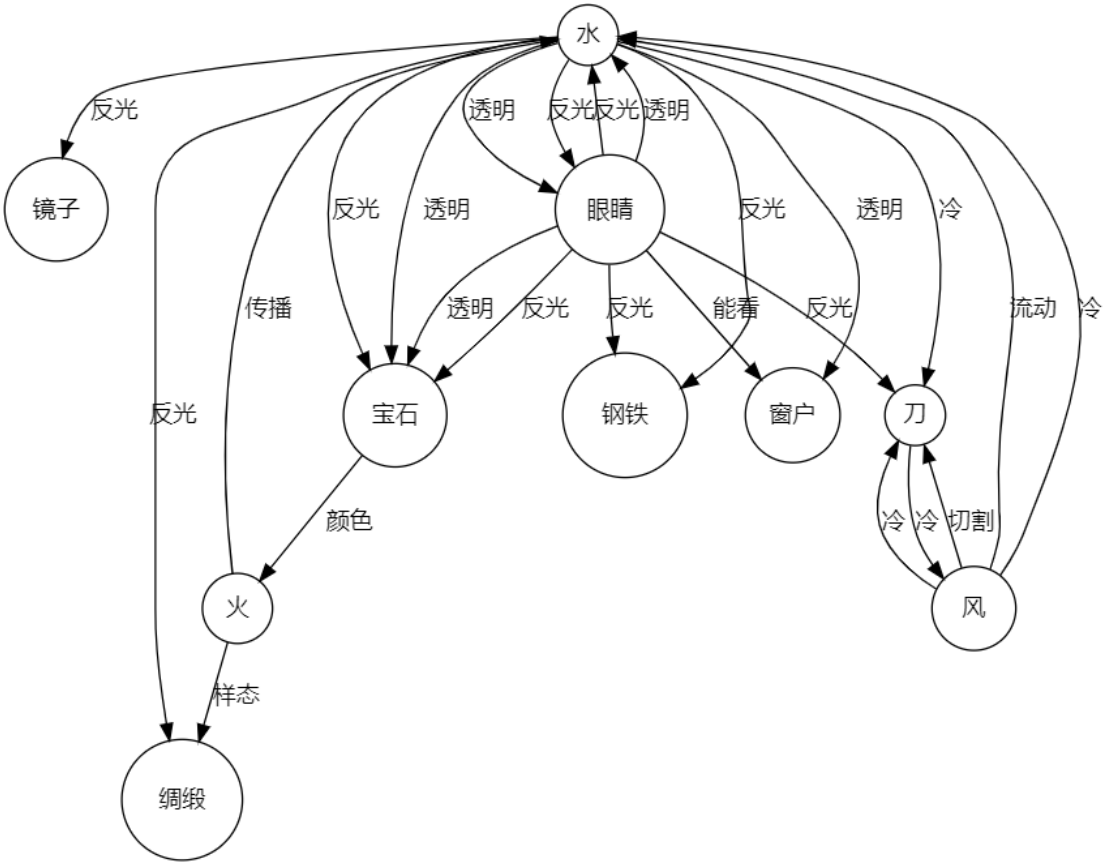
2. 按照上述规则，使用 Graphviz 构建该图

代码如下：

```
digraph G {
    //splines="FALSE";
    /* Entities */
    water [label="水", shape="circle"]
    mirror [label="镜子", shape="circle"]
    gem [label="宝石", shape="circle"]
    window [label="窗户", shape="circle"]
    eye [label="眼睛", shape="circle"]
    knife [label="刀", shape="circle"]
    iron [label="钢铁", shape="circle"]
    wind [label="风",shape="circle"]
    silk [label="绸缎",shape="circle"]
    fire [label="火",shape="circle"]
    /* Relationships */
    water->mirror[label="反光"]
    water->gem[label="反光"]
    water->gem[label="透明"]
    water->window[label="透明"]
    water->eye[label="透明"]
    water->eye[label="反光"]
    water->knife[label="冷"]
    water->iron[label="反光"]
    eye->window[label="能看"]
    eye->iron[label="反光"]
    eye->water[label="反光"]
    eye->water[label="透明"]
    eye->gem[label="反光"]
    eye->gem[label="透明"]
    eye->knife[label="反光"]
    knife->wind[label="冷"]
    wind->knife[label="切割"]
    wind->water[label="流动"]
```

```
    wind->knife[label="冷"]
    wind->water[label="冷"]
    gem->fire[label="颜色"]
    fire->silk[label="样态"]
    fire->water[label="传播"]
    water->silk[label="反光"]
    /* Ranks */
}
```

图形如下：



注：上图示中每一条有向线段均有相应比喻句作为佐证，相关比喻句请参见附录Ⅰ。

（三）对该系统可靠性的验证

在上图中，可以看到存在“眼睛→钢铁”和“眼睛→水→钢铁”两条比喻句构成的链式结构。在眼睛比作钢铁的比喻句中，是因为眼睛的反光性质。而眼睛比作水和水比作钢铁的比喻句中，也都是处于其能够反光的特性。可以看出该系统有较好的自洽性。

另外，考察“宝石→火”，由样本，宝石可以被比作火，而根据我们的经验，火很难被比作宝石，而图中火并不直接指向宝石，若想通向宝石，需经过极远通路，符合人们语言习惯。可以看出该系统能较好抽象。

四、 可行性分析与算法讨论

那么，上述系统可能有何实际应用呢？

首先，其对于自然语言生成可以有所帮助，如需生成一个比喻句，可考察其本体和各种备用喻体在上述系统中的距离（对于每两个节点，若直接相连，其直接相连的线段越多，相应对应距离越短（d<=1），若不直接相连，按照通路的线段条数对应相应距离（d>1)），用 Floyd 或 Bellman-Fort 算法求最短路即得最佳喻体。这样，程序构建出的比喻句更可能是合理语句。

同时，上述系统具有可拓展性，只需加入新的语言材料，即可在原系统的基础上构建更大的系统，毋需重新构建。

以下是对上述系统可能的两种改进：

1. 距离加权：对于两个概念之间的距离的描述，可以引入更全面的评估体系，取加权距离，可以更准确的刻画日常语言的使用。
2. 环境词：该系统只考察了“环境性”规律中本体喻体关系，未考察上下文中特征词对于喻体选择的影响，如果将特征词加入系统的构建，求“综合距离”，可以改善这种情况。

   Eg. 清明阴郁的雨压在大地上，积着雨水的云伸手触摸着树冠，荒郊的石头路，好像破碎的_____。

   此处如果使用“石碑”，较合适，如果使用“砚台”显然就不甚合适。但二者跟石头显然都关系密切，区别在于与语境的联系。

   因此如可改进该系统，可以考察“石碑”和“砚台”二词到“石头（碎石)”以及环境特征词（“清明”,“阴天”,“荒郊”等）的“综合距离”，即可选出合适的喻体。

五、 总结与反思

人工智能是目前人类计算机技术的前沿，而自然语言方向是热门方向。一方面，语言学的研究可以有助于自然语言生成的合理性，另一方面，基于计算机技术带来的强大算力又可以帮助语言学发现一些规律。可以说，交叉学科数字人文等前景光明。

在本学期研讨课的过程中，笔者也体会到了研讨交流的重要性，如本文中需要

分析大量概念的常识元，如“刀”和“水”的常识元分析结果援引自韦尔奇同学的研究，众多同学的研究结果汇集到一起，可以形成一个更大规模的数据库，有利于研究的进一步开展。

参考文献

[1] 荀恩东，饶高琦，肖晓悦，臧娇娇. 大数据背景下 BCC 语料库的研制[J]，语料库语言学，2016(1).
[2] 韦尔奇.对常识元是否有等级和数量的研究[亦为本课程期末论文]

附录 I:

1. 狐女若若/卫小游：湖水像镜子似的映照出她的相貌，她看得出神；突然，湖面浮现一张她再熟悉不过的脸庞，她的心跳倏地乱了一拍。
2. 莫言/红高粱家族：她飞着，飞着，然而离井口总是那么远，她飞得越快井筒延伸得也越快。半夜时她有过一次短暂的清醒，她触到了弟弟冰冷的身体，她不敢想弟弟已经死去了，她想一定是自己发烧了。一帘折射进井底的月光，照亮了那汪绿水，癞蛤蟆像个宝物一样，眼睛和皮肤都放出宝玉光泽，那汪水也像翡翠一样绿得可爱。
3. 托尔金/魔戒全集(指环王 1、2、3)：他听见一阵微弱的嘶嘶声，一种诡异的味道飘出来，光芒开始闪烁，四处飘汤。一瞬间，他眼前的池水变得像是某种窗户，让他可以看到另外一个世界。
4. 莱蒙特/农民们（下）：篱外的田野灰雾弥漫，只有最高的树梢依稀浮出来，像一股浓密的黑烟。水塘像 一只看不见的巨眼，在暗夜中发光，
5. 魏巍/东方：到了中流，江水已经齐了人们的腰部。激流卷起的波浪，溅到人们的脖子里，棉裤成了千斤重的水袋，坠得迈不开脚步。冰冷的江水 就像 刀割一般。但是战士们高高地举着枪支，互相搀扶着，顽强地向对岸前进。
6. 俗语：眼睛是心灵的窗口。
7. 温赛特/克丽丝汀的一生（下）：我发现你对我从无恶意——尔郎，我心思不如你想象中那么高洁——不如你高洁——我若对不起人家，对他难免有芥蒂他激动得双颊布满浅斑，眼睛凝视尔郎的双眼。尔郎半开着嘴巴听他说。他骇然耳语说：“我做梦都没想到这一点！西蒙，你恨我吗？”“你不觉得我有理由两个人不知不觉拉住马儿。他们凝视对方的面孔：西蒙的小眼睛亮得像钢铁。
8. 路遥/平凡的世界：她好像一下子老成了。那双春波荡漾的眼睛一夜间变得

像秋水一般深沉。

9. 伊凡·谢尔盖耶维奇·屠格涅夫/猎人笔记：苏乔克以从年轻时就侍候老爷的人的那种眼光瞧着我们，不时地喊道：8n“那边，那边还有一只鸭子!”他常常在背上搔痒痒响着芦苇的沙沙声：太阳照耀下的水塘处处像钢铁似的闪着亮。我们已准备返回村子，霎时问发生了一件大杀风景的事。
10. 莫言/红高粱家族：狐狸的皮毛灿烂极了，狐狸的略微有点斜视的眼睛像两颗绿色的宝石。后来他感到了狐狸的温暖的皮毛凑近了自己的身体，他等待着它的尖利牙齿的撕咬。
11. 莫言/丰乳肥臀：时候到了，我手中的刀磨得比北风还要快，还要凉，我的刀像北风一样凉快，我要让他知道杀人者必得偿命的道理。
12. 余华/祖先：我母亲的眼中越来越显示出了疑惑不解。前面浓密的树林逐渐失去阳光的闪耀，仿佛来到了记忆中最后的情景，树林在风中像沉默的波涛在涌动。
13. 海泽/诺贝尔文学奖文集（倔犟的姑娘 葡萄园看守 耶恩森短篇小说集）：火焰劈劈啪啪地在烧着，当我们靠近农家时，你简直不会想到温度有多高了。风也像种痘的小刀般地刮着面颊。嘴唇干渴得连话都说不出来。
14. 亨利·米勒/情欲之网：“他们是疯子、白痴!”我自言自语。突然，我问自己为什么会趴在如此冰冷的脏水里。荷比轻轻地叫我：“上来吧，河岸边还亮，我们可以在这儿稍呆几分钟。”当我扶着梯子往上爬时，风冷得像刀刮一样。
15. 王小波/2015：碱场有好几台拖拉机，冒着黑烟在荒原上跑来跑去，就像十九世纪的火轮船。那个地方天蓝得发紫，风冷得像水，碱又白又亮，空气乾燥得使皮肤发涩。我舅舅闭上了眼睛，想要在太阳底下做个梦。
16. 亨利·米勒/情欲之网：现在我正走近干草市场。突然一个名字从广告栏中跳出来将我的眼睛 削得像刀刃一样闪亮。我正好路过一座早已认为毁掉了的剧院。
17. 莫言/丰乳肥臀：母亲的大姑姑讲起话来嘎巴脆，像快刀切萝卜。炉中的火焰失去了风箱的鼓动软弱得很像黄色的绸子。火苗上摇曳着焦香的煤烟。
18. 托尔金/魔戒全集(指环王 1、2、3)：当他转过身，走向他们时，佛罗多才发现这是甘道夫，他的手上戴着第三戒——纳雅，上面的宝石红得像火一样。要离开这里的人们都觉得很高兴，因为甘道夫将会跟众人一起出发。
19. 林徽因/记忆：断续的曲子，最美或最温柔的夜，带着一天的星。记忆的梗上，谁不有两三朵娉婷，披着情绪的花无名的展开野荷的香馥，每一瓣静处的月明。湖上风吹过，额发乱了，或是水面皱起像鱼鳞的锦。四面里的辽阔，如同梦荡漾着中心彷徨的过往不着痕迹，谁都认识那图画，沉在水底记忆的倒影!
20. 亨利克·显克维支/你去什么地方（你到何处去）：无可置疑有些罪恶的手在四处放火，因为时刻都有新的火场在远离大火中心的地方爆发出来。从罗马市所在的高地，火焰像海浪一般，朝那布满密密层层房屋的洼地灌注，房屋都有五六层，全是店铺、摊头以及为了便于各种表演的木造流动圆戏场