# 论文

## 细读

### 论文：委婉语检测

英文名：Self-Supervised Euphemism Detection and Identification for Content Moderation

**级别及代码**：a类，https://github.com/WanzhengZhu/Euphemism

**任务**：1.委婉语检测：了解哪些词被作为关键字的委婉语 2.委婉语鉴别：识别委婉语的含义

**角度**：委婉语与相应得禁用词上下文的相似性

**难点**：没有公开合理数据集，委婉语不一定使用了委婉语义

检测难点：语料不均衡，作为正常使用数目要比非正常使用数目多。

用bert的话基于相似度方法不再合理，对某些句子而言，mask在句子中含信息量不大，可替代性较高。

消歧与本任务：本任务更多倾向于发现新的意义而非在某个特定句子下的语义

鉴别难点 :意义关联小、Code words 的多义词问题 、无数据集（犯罪类型的code words 类型有限）

**实现**：

1.委婉语检测：

从含target的语料a中提取掩码句s，用bert在某类型语料b上进行微调，用bert来判断相关性从而筛选掉通用的掩码句子。再使用一系列可能委婉词在该bert上计算mlm概率，根据rank list 来确定委婉语e\_word。

过滤过程：使用掩码句s，在某类掩码句子（例如药品）上微调mlm模型，根据候选列表来的top k 中是否有target 来筛选掩码句子。将全部词表作为候选。

代码细节：

2.含义鉴别:

使用一个粗分类器，将语料c中所有包含e\_word句子选取，使用粗分类器判断是否与关键字相关(是否使用了委婉语义，解决难点)，再通过多分类器来进行判断委婉语义，

粗分类器：使用关键字筛选的掩码句s和负采样句子1：1进行训练。（即关键字使用的上下文和e\_word如果使用了委婉语义的上下文应该是类似的，这样的一个分类器也就能够判断相应结构的上下文）

细分类器：使用掩码语句s和关键字作为标签进行mlm训练，同理这样的一个分类器应该能够去判断带e\_word的句子上下文符合哪一类关键字的上下文。

**实验**：

指标 p@k 用于衡量接受与结果查询的有效性，不用recall 不好确定（如何设计好的recall 指标）

Baseline：word2vec、tf-idf+word2vec 、cantreader、senteuph、eigeneuph、grapheuph、MLM-no-filtering

Tf-idf用词频进行一次排列，再用word2vec 实验证明有效。MLM-no-filtering 没有使用粗分类器的版本。

鉴别实验：

指标：根据标签是否在输出的列表前1、2、3确定准确率，列表由概率分布排序。

Baseline：word2vec、clustering+word2vec、binary+word2vec、fine-grained-only

Clustering 聚类方法 效果一般，经常根据词出现来进行分类。Binary效果较好。

鉴别中使用的粗分类器包括 lrt、lte、rnn、lstm、lstm-attention、cnn、rcnn、self-attention

其中lrt、lte都是逻辑回归分别在one-hot 和词嵌入向量上训练。以上所有词嵌入使用都是glove-embedding。粗分类器上 lstm-attention效果最好，在细分类器上效果都差不多使用lrt。

**不足**：没有短语、没有更合理的筛选。只能处理文本、在其他领域比如情感分析效果不好，可能与委婉语与目标词指代关系模糊有关。数据来自各个论坛，并非完全为了逃避检测的文本，没做对抗性测试。没在实际中测试，但是在各个平台的毒性检测上证明委婉语有逃避检测效果。

**扩展**：

委婉短语检测：针对委婉短语进行的检测，添加了一个筛选候选词的过程。使用autopharse来筛选候选委婉短语。具体使用w2v 结合余弦距离来进行筛选。（word2vec是静态的，可能存在问题）。其余与上一篇类似，区别是使用spanbert从短语层面进行遮掩。

### 论文：地下论坛黑话识别

英文名：Reading Thieves’ Cant: Automatically Identifying and Understanding Dark

Jargons from Cybercrime Marketplaces

**级别及代码**：a类，https://sites.google.com/view/cantreader

**任务**：1.委婉语检测：了解哪些词被作为关键字的委婉语 2.委婉语鉴别：识别委婉语的含义

**角度**：委婉语与相应得禁用词上下文的相似性

**难点**：1.如何对两个不同语料之间的差异进行建模 2.如何处理多义问题，在合法语料中 3.如何理解黑话

**实现**：

1.委婉语检测：

同时处理两个语料库的句子，输出两个向量，使单词上下文之间的语义差距可测量。对抗多义带来的假阳性问题，使用新的模型来比较在合法语料中意义的偏移。

流程：使用合法与非合法的语义进行相似度比较，再利用合法与专业解释进行相似度比较。当第一个相似度低第二个相似度高时候确认为黑话。

模型实现：增大输入，然后使用w2v方法来训练，根据语料来源确定输入onehot在输入向量位置，其余用0填充。将隐藏层分解作为词嵌入矩阵。

模型使用原因：word2vec混合训练不同语料库的不行，但是单独训练由于word2vec有随机性，意义不大。相当于训练出的向量在不同的坐标下，没法直接比较。

验证：比较了使用word2vec与scm 的向量质量，和同义下的输出向量比较。以及跨语料的捕捉能力。（只考虑到了合法语义的偏移带来的假阳性问题，）

预处理：删了一些停止词和不够多样化词

阀值选择：根据语义多少确定不同阀值

2.含义鉴别:

使用一个模型来确定上位词 是否是is-a 关系来确定黑化含义。使用一组自定义种子结合wikidata找到上下级关系，构建一棵树。然后使用SCM的词向量来进行二分判断，是否是一个黑化。如果是就继续递归往下查，根据分类器结果判断选择哪棵子树。（缺失理论验证）

**不足**：可以增多输入来进行验证词嵌入训练的正确性。受到语料质量影响，没有做短语级别的。

### 论文：地下市场分析工具

英文名：Tools for Automated Analysis of Cybercriminal Markets

**级别及代码**：a类，2017，[https://evidencebasedsecurity.org/forums/#data](https://evidencebasedsecurity.org/forums/" \l "data)

<https://github.com/ccied/ugforum-analysis>

**任务**：提取论坛产品信息的工具

**难点**：1.如何提取论坛信息 2.保证工具在跨论坛情况下的性能

子任务：1.post type 确认帖子类型，买还是卖还是交换货币或无关 2.产品：产品名词提取

3.价格和交易信息

**实现**：

1.部分论坛本身就有类型信息，其余进行人工标注。模型选择MFL（最大频率单词）和使用1~6grams 的SVM进行。跨领域效果不差。Svm好于MFL，但在加入额外语法特征时效果没有显著提升，可能时由于（1）语法提取系统一般都是处理规范文本。（2）回帖往往偏离帖子主题，导致使用更多的帖子效果反而不好。

2.标注时标注名词，在名词、基本短语、复杂短语中标注公共词，避免关于名词边界的分歧。模型：名词上下文特征的svm分类器和目标为选取最有可能为单词的潜在SVM（以要提取的产品名词作为训练时的潜在变量）。系统在跨论坛时效果下降，同时一个帖子含多个产品情况少见。

3.价格和交易信息。其中价格使用正则或使用了词特征和Brown聚类的SVM。SVM好于正则，但同样在跨领域时效果下降。交易信息：提取货币类型、数量等信息，使用了四种方式。正则：根据正则来进行匹配，第一个识别的为出价 第二个为请求价。模式匹配：进行预处理后用模板进行匹配。词级别：从词级别角度进行分类 ，全级别：考虑帖子中全部词来进行分类。

**扩展**：主要是用于跨领域的方法：猜想：跨论坛产品差异大；或不仅产品，行文风格差异也大。1.使用brown聚类在未标记数据上训练，然后作为特征（效果一般）

2.使用部分标注数据做成产品列表，以出现与否作为特征3.使用部分目标论坛数据，并添加论坛信息标记，放大五倍目标领域数据的重要性。效果有，但有过拟合的可能。

### 论文：

英文名：Reading Thieves’ Cant: Automatically Identifying and Understanding Dark

Jargons from Cybercrime Marketplaces

**级别及代码**：a类，https://sites.google.com/view/cantreader

**任务**：1.委婉语检测：了解哪些词被作为关键字的委婉语 2.委婉语鉴别：识别委婉语的含义

**角度**：委婉语与相应得禁用词上下文的相似性

**难点**：1.如何对两个不同语料之间的差异进行建模 2.如何处理多义问题，在合法语料中 3.如何理解黑话

## 浅读

### 论文：情绪分析来寻找委婉语

英文名：Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis

**级别及代码**：a类workshop，

**任务**：1.委婉语检测：了解哪些词被作为关键字的委婉语 2.委婉语鉴别：识别委婉语的含义

**角度**：委婉语使用与潜在主题相关，往往呈现出一种情感特征。与文化不合或负面话题相关，以及一些敏感但不消极话题上。委婉语一般会降低敏感性。但委婉语仍可能有负面性。

**解决**：1.基于字典方法探索在情感极性、内涵、强度、唤醒和支配方面值。2.收集上下文根据极性来进行分类。

1.字典方法：用basilisk算法根据上下文寻找同义词，由确定的三个种子类别扩展，找到一系列动词短语和名词短语。使用词汇-句法模式来构建上下文的模式作为上下文。人进行标记得到一个数据集。使用选定的指标来作为判断委婉句的方法，其中以内涵指标效果最好，混合指标的方法效果能有更好的提升。但该方法只是单词级别的衡量。

假设围绕委婉语和语言障碍的语境在情感上是不同的

使用1中的最好用的词典来标记句子的单词，一句话中被标记单词的极性投票确认句子极性。只有当短语的情感极性超过句子一定程度时，根据积极消极标记为委婉语非委婉语（还有中性）。否则都为中性。

另一种标记短语方法是通过nrc工具标记句子，然后根据包含短语句子和所有句子极性的差异值来确定短语极性（差异最大的作为短语极性）

扩展：Determining Code Words in Euphemistic Hate Speech Using Word

Embedding Networks

### 论文：使用词嵌入网络确定仇恨发言中的编码词

英文名：

**级别及代码**：，

**任务**：检测在仇恨发言中的编码词

**角度**：词节点的中心性间接为上下文提供了重要提示

**数据集**：选择一个起始code word 来筛选数据集（确保是仇恨言论相关数据）

**解决**：由于暗语本身被选择出来，所以与仇恨语义上下文单词相关性较低。选择word2vec作为向量构建图，余弦距离表示边，暗语更有可能远离其他词。暗语很可能共现出现，将特征向量中心度作为排名依据。寻找完全图，根据特征向量中心选择一个完全图社区作为候选词。

**扩展**：寻找仇恨言论社区：通过极端社区找到相应推特账号，根据follow 和follower 构建图。根据经过某点的单源最短路径次数确定其重要性。候选词：用相似性构建边权值，选择已知的仇恨词汇，根据pagerank增加相关点的权值。根据词义丰富程度再进行筛选。

### 论文：Transforms 进行识别+towards codewords

英文名：Towards Dark Jargon Interpretation in Underground Forums

**级别及代码**：C类

**任务**：

**角度**：上下文相似性确定含义

**数据集**：抽样标注，删除停止词后根据频率选取一个公共列表

**解决**：mlm 用clear corpus训练 用dark语料来预测。

基于单词分布，用kl散度来衡量上下文相似；

**扩展**：跨语境分析：用ccla框架来进行跨语境上得分析计算评估相似性

### 论文：微博中codewords检测：

### 英文名：

**级别及代码**：

**任务**：

**角度**：上下文相似性确定含义

**数据集**：

**解决**：从codewords在不同语料中的上下文不同出发。例如w，选出某个语料中的N个相似的词。若相似词在codewords列表中计一分，则递归考虑相似词的N/2个词的分数，递归两次。若相似词的最后分数超过一定值，给w计一分。

当w在bad corpus的分数超过一定值如A时，或者超过a（a<A）但在好坏语料中的差异超过一定值时，认为是code words。

**扩展**：“Distinguishing Japanese non-standard usages from standard ones“

### 论文：中文语境下的telegram论文：

### 英文名：

**级别及代码**：

**任务**：

**角度**：上下文相似性确定含义

**数据集**：

**解决**：收集了telegram数据，结合两个语料库。做类似 cant reader的工作，只不过没有使用scm，而是将学到的词向量投影到同一空间下。

判断的时候结合了其他字典特征等。

**扩展**：

# 相关

## 相关内容

### 索要数据

Dear Professor, Greg Durrett

I'm a student from Beijing University of Posts and Telecommunications, China.

My research interest is to use NLP technology for content moderation.

I've read your paper Identifying Products in Online Cybercrime Marketplaces: A Dataset and Fine-grained Domain Adaptation Task, which give me a lot of inspiration.

I learn a lot from your rigorous experimental demonstration in this paper.

I start to do experiment now which need your data in this paper. so i write this letter for it.

It will be very helpful for my experimet Thank you very much!

Best Regards

### 创新点由来

第一步:直接抄技术万事开头难，做研究也是，在一个课题的初始阶段，如果没有头绪，完全可以直接把觉得有价值文献里的技术，原封不动的抄过来走一遍，先让研究思维动起来，解决“有没有”的问题。在“抄”技术的过程中，当已经能够熟练掌握这套方法了，再结合自己课题的特征，通过大量复盘和跑盘，不断优化模型，调整参数，最终得到一个适合于自身的优化模型或者方法，初步的创新点就这么成型了。就比如小白做吐司面包，通过照着教程一顿努力，终于做出一个普通的奶粉吐司，那是不是可以根据自身口味，把奶粉这个参数优化替换成抹茶粉，巧克力粉，椰子粉，创新出更多口味的面包。

第二步:算法创新这个也是大多数论文常用的套路。大多基础研究，其实就是寻找世间客观存在的规律（简称找规律），也就是找实验数据和事物表象之间的一个联系，当出现什么什么参数特征时，就能推测出一个怎样怎样的事实结果。

第三步: 融合式创新读文献不局限本行业的，可以尝试把不同研究领域的结论融合起来，就好比“相机＋手机＝拍照手机，AI＋家具＝智能家具，包括最近常听到的互联网＋”等等。有时候，把跨界的研究成果拿过来用，会有意想不到的创新点。比如雷达感应系统，和扫地机器人结合叫精准保洁，和汽车结合就叫主动刹车/巡航系统/智能驾驶，和手机结合又叫高精度手势操控，等等。

第四步:横向运用型创新如果已经得到了一个运用于A领域，比较成熟的技术，完全可以再试一试拓展它的运用区间，拓展运用到A＋，或A－等横向领域，得到若干运用型创新点。

### 相关方向

* 词义发现
* 词义消歧
* 同义词
* 跨语言解决（增强语料）
* 风格识别
* Prompt

### 想法

* 委婉语与上下文的不合理性，去识别。
* 语料大小引起的数据质量差异。
* 假阳性指标缺失问题
* 候选词问题 x
* 可能存在的逃避情况
* Mlm缺点：假定预测token相互独立，以交易形式为主 ，标题也包含更加集中的信息，但是较短。
* 用关键字筛选reddit句子，得到黑暗语料。再对干净语料进行对比
* 多任务 都很难 得到一个概率分布 在语料意义上的 。二分类，且是简单的二分类，提高了下限 且对单句能快速处理。根据语料判断作者。 在此之上 把词去掉，防止 无害含义的干扰。做实验。 正负比例差异过大时 怎么处理
* 与Coleridge 比赛情况类似
* 中文语料：聊天、数目少、特点过于明显，易理解干净。英文语料：量大、不干净、帖子形式（冗余大）、有背书
* Fine-tune程度
* 后处理：基于假设进行筛查。1 名词 2.与黑暗不相干

## 中文语境下的处理

* 同音同态出发：同音词、同态词

用一个LM模型来检测那些不流畅的句子。据此分析句子与语料之间的相似度。根据同态

同音恢复其原来形态。其中同音由codewords列表与句子窗口内的词计算拼音距离，确定后再进行替换。同态则有一个名为四角的系统来衡量字型距离。恢复之后得到行话候选语，

再使用LM舍弃低概率的句子，这些句子很少在我们的语料库中出现。

* 自监督的意义在哪里：可以发现新的词，要证明我们发现的可能是新的词，而不是仅仅依靠一个词典来证明。只是对现有词典的过拟合。我们要做的是帮助禁用词系统来达到低召回率的效果。补充：不仅仅是发现新的词，设计的分类器应该也能用到禁用词列表中的词
* 语料限制：基于禁用词列表的自监督方法无法扩展语料，由于审查的逐渐严格，没有办法再根据禁用词筛选得到新的语料，self论文中使用的也是18年以前的数据。而在中文环境下这样的语料更加缺乏。
* 半监督：从一组标签数据出发，被标签数据标注的code words 与关键字有相同词义，起到一样的作用。使用这部分数据进行mlm，捕捉新的词。（可以只使用一部分关键字的code\_words ，看在另一部分上的表现）。具体实现上可以使用新标注的结果作为伪标签，并逐渐根据反馈增加可信度。
* 筛选级别：必须从短语级别出发，telegram收集更多数据，关注方面可以放到性交易、药品上。使用ernie能完成短语级别的mlm。
* 中文语料特点是在形容词上不使用掩码（可以用于帮助识别和鉴定），
* 设想的系统流程：开发系统，A模块根据无监督模型来捕捉违禁词的上下文来帮助发现新的词，提供给禁用词列表进行更新。禁用词系统根据列表捕捉到大量新的句子，然后B模块通过分类器判断是否是黑话。最后大量句子作为伪标签反过来帮助A模块做无监督。
  + A模块使用的语料包括两类1.黑暗风格语料 2.明确含有关键字的句子

A模块使用禁用词系统提供的句子作为1、2的补充？为什么要补充？

适应新的上下文风格。怎么证明这一点呢？

黑暗风格语料中最关键的是存在未识别的委婉语的使用，这部分怎么补充？

这部分由具体某个特定社交平台数据来进行补充。（使用委婉语进行交流），或者用b模块帮助识别这个类型的语料。

* + A模块能否帮助B模块？prompt 方法来进行/语料+填充来作为训练集 lda、svd方法 ，比例偏差较大情况 如何做数据增强
  + A模块 含禁用词的句子找不到怎么办 采用一部分标注数据的上下文。自己再爬取一部分黑暗语料数据
  + 为啥不直接监督数据加伪标签形式，使用mlm做无监督好处是什么？

猜想可能标准的监督数据加伪标签的方法对模型本身的训练精度要求较高，且监督数据如果数据量较小，数据中的分布和实际分布存在差异，半监督并不能帮助修正这种差异？或者比较难

但是利用无监督加半监督的思路对于错误的容忍程度较高，可能可以接受这种差异。

* 存在的问题：
  + 是否要做鉴别
  + 既然要帮助禁用词系统，仅仅添加禁用词列表会带来大量的假阳性例子，如何处理？可能需要设计一个合理的二分类器帮助识别当code word在句子中时候，是否是一句黑话。而且速度要快。课题又变成了词义消歧。
  + 结果比较，要和有监督的比。发现新的词的能力。设定某个词典{“禁用词”：委婉语列表}。训练数据包含 某个委婉语的所有句子。测试数据 其他委婉语的词 计算准确率。

## 相关论文