中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Identification of jargons

2024.12.27 Fanglong Dong

Warning: this seminar includes examples that may be offensive or harmful

口 暗语（jargons）

✓ **定义**

对敏感词汇进行**特殊替代**或**歪曲**的表达方式

✓ **类别**

Dark jargons和Euphemistic Terms

✓ **特点**

高度的隐蔽性、专业性和创造性

口 暗语检测

✓ **数据层面**

获取高质量的暗语语料库

✓ **特征挖掘**

捕捉隐晦的语义表达特征

✓ **暗语识别与评估**

高效暗语识别框架与评估体系

口 新词发现

✓ 问题

① 大量的黑话/黑词对于下游任务非常有效，但却不在通用的词典中，导致**分词器无法准确切分出对应的词**。

② 如果采用现有的分词工具（如Jieba）进行分词会导致很多词被拆分，无法得到词的embedding vector。

✓ 解决办法

先发现新词，添加到分词词典工具中（如Jieba），再tokenize

✓ 需求

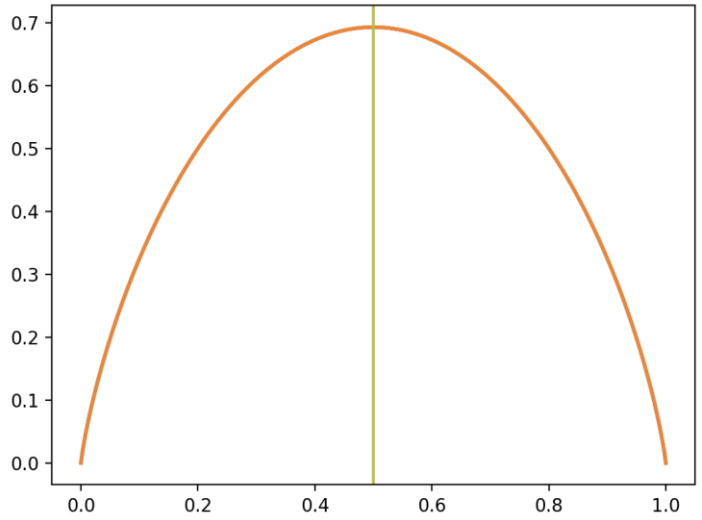
怎样才算一个“词”？如何发现新词?

口 什么样的字符组合可以称为一个“词”

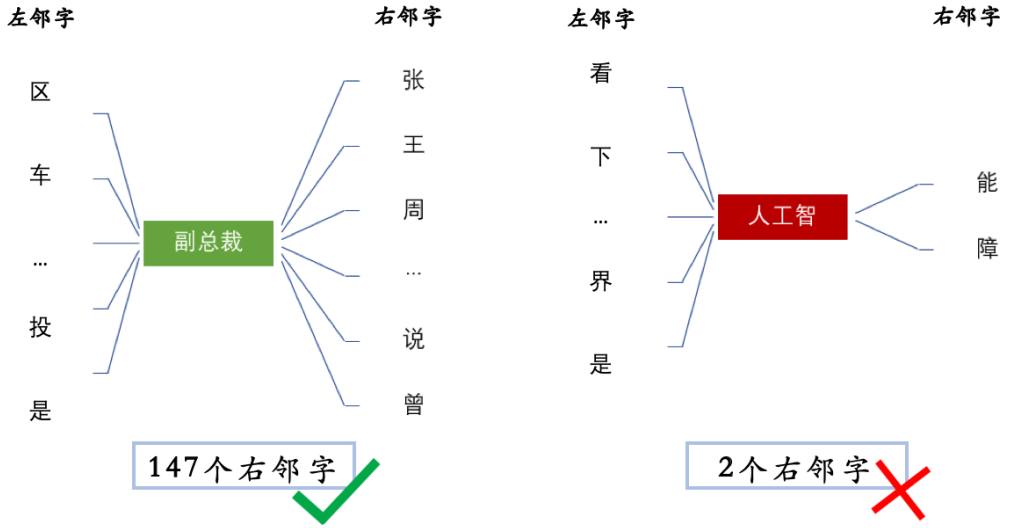
✓ 词语的**左右邻字**（上下文）要足够丰富

信息熵是对信息量多少的度量，信息熵越高，表示信息量越丰富、不确定性越大。通过计算一个候选词左边和右边的信息熵来反映了一个词是否有丰富的左右搭配，如果达到一定阈值则我们认为两个片段可以成为一个新词。

$$H(w) = -\Sigma_x p(x) \log p(x)$$



$$Entropy(w) = -\sum_{w_n \in W_{Neighbor}} P(w_n|w) \log_2 P(w_n|w) ,$$



“珠穆朗”的右邻字：{玛}，1个字，右信息熵小，确定性大

“副总裁”的右邻字：{张, 王, 说, ...}，共147个字，右信息熵大，确定性越小；

“人工智”的右邻字：{能, 障}，2个字，右信息熵较小，确定性越大；

“沂”的左邻字：{临}，1个字，左信息熵较小，确定性越大。

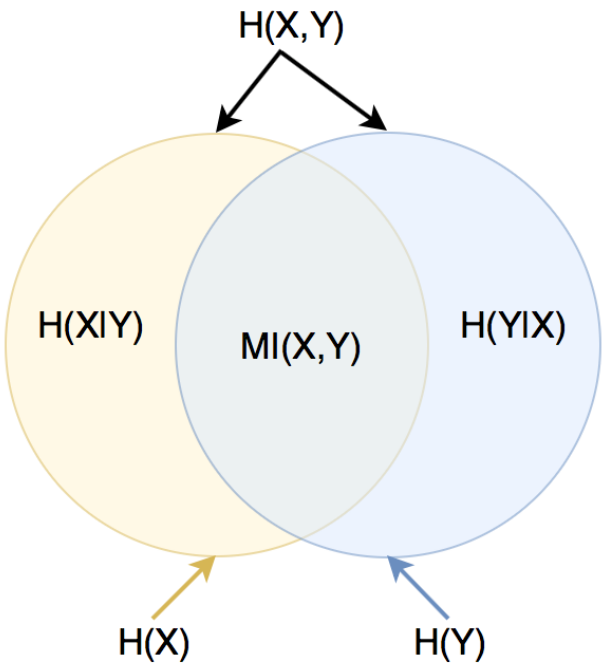
什么样的字符组合可以称为一个“词”

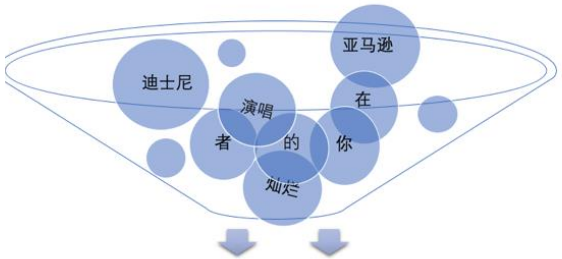
- ✓ 词语的**内部凝聚程度**要足够高

互信息（MI）表示了两个随机变量 X，Y 共享的信息量。互信息表示如果已知任意一个变量后，对另一个变量不确定性的减少。而点间互信息（PMI）计算了两个具体事件之间的互信息。

$$MI(X,Y) = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$PMI = p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$





Eg.：已知一个 music corpus
“演唱者”：117次，“的演唱”：275次
“演唱”：2502次，“者”：32272次
P(演唱)P(者)=2.24*10-7
P(的)P(演唱)=3.07*10-6
P(演唱者)=27*P(演唱)P(者)
P(的演唱)=4.7* P(的)P(演唱)

$$p(\text{演唱者}) \log_2 \frac{p(\text{演唱者})}{p(\text{演})p(\text{唱})p(\text{者})} >> p(\text{的演唱}) \log_2 \frac{p(\text{的演唱})}{p(\text{的})p(\text{演})p(\text{唱})}$$

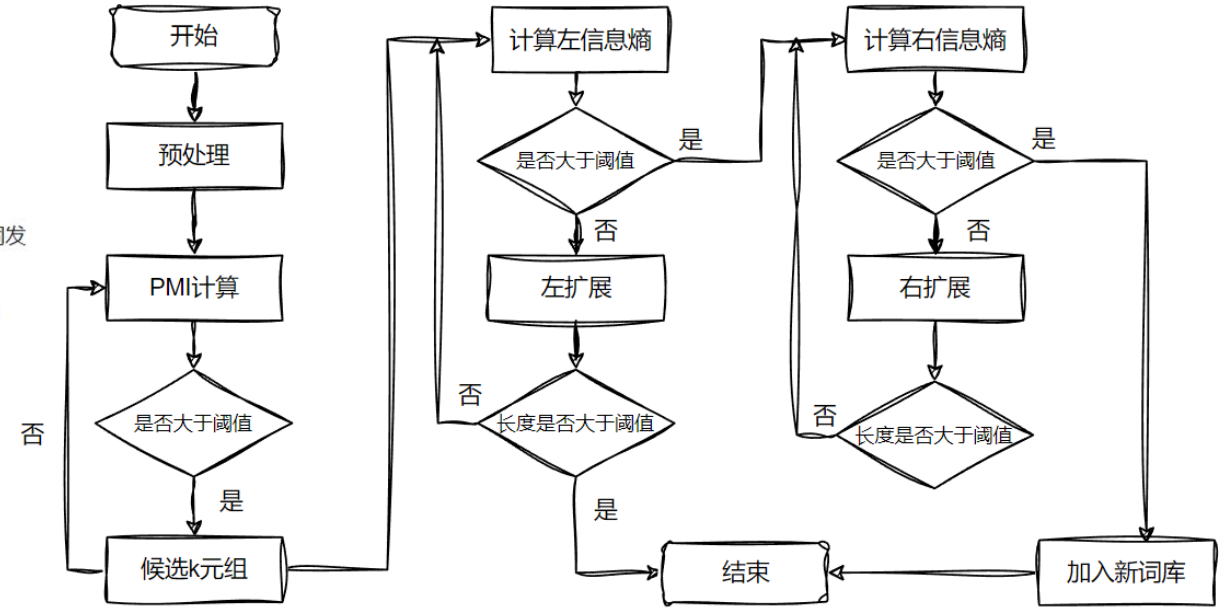
综合考虑左右邻字丰富程度和内部凝聚程度

- ✓ 只看内部凝聚程度：无法过滤“人工智”、“工智能”
- ✓ 只看左右邻字：无法过滤“xx的”、“在xx”

$$score = PMI + log\frac{LE * e^{RE} + RE * e^{LE}}{|LE - RE|}$$

■ 新词发现流程

这一步我们将文本按字分割后拼接为二元组，三元组*，...，k 元组（一般 k <= 5），如“新词发现及一词多义的解决方案”对应的二元组有：[“新词”，“词发”，“发现”，“现及”，“及一”，“一词”，“词多”，“多义”，“义的”，“的解”，“解决”，“决方”，“方案”]。



① 生成候选词：将文本切割成k元组的形式（k≤5）

② 候选词得分计算：计算每个候选词的score，与阈值对比

③ 将新词加入到分词器的词典中（jieba.add_word(“xx")）

④ 计算得到每个词的 embedding vector（word2vec、bert…）

⑤ 基于预先设定的seed keywords，计算新词（或所有词）与种子词的相似度，筛选得到seed领域的新词

比如种子词选取毒品，最终发现“溜冰”这个原本看似人畜无害的词与毒品相关的种子词相似程度均很高，即可推测自己发现了一个该领域的新词。

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

- An Unsupervised Detection Framework for Chinese Jargons in the Darknet, WSDM22(ccf-b)

- A novel cross-domain adaptation framework for unsupervised criminal jargon detection via pre-trained contextual embedding of darknet corpus, Expert Systems with Applications, 2024(SCI 1区, IF=7.5)

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

Motivation

- ✓ 无论一个词的“外形”如何，**其真正的语义可以从其上下文（context）中得出**
  - 暗语的上下文语义环境往往与其隐含的原始词汇有较大的相似性（eg. Porpocon: food or drug）
  - 但相比之下，这些暗语在正常使用场景中的上下文环境却会发生显著变化
- ✓ 下位词（(hyponym)）的真正含义往往取决于上位词（hypernym）
  - **暗语（下位词）与上位词存在“is-a”关系**，确定真实含义

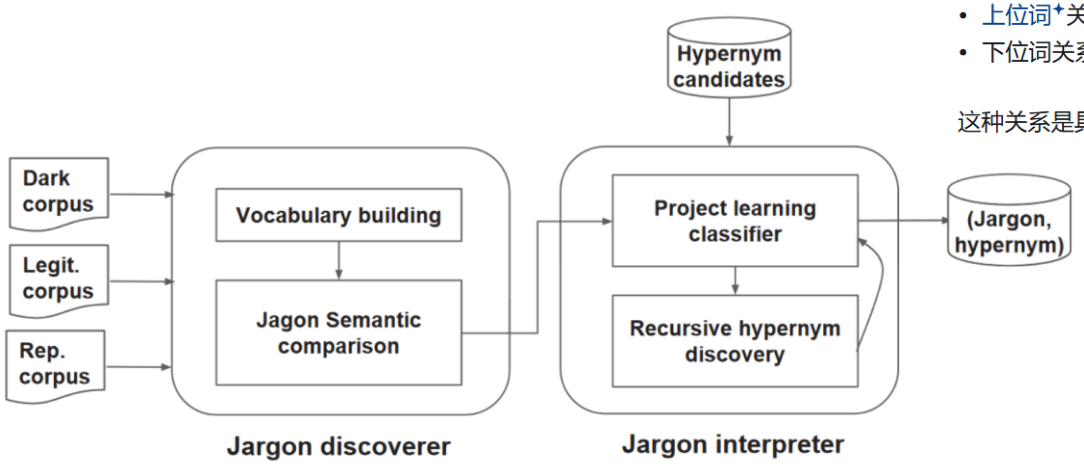


Figure 3: Overview of the Cantreader infrastructure.

3.1 上位词和下位词关系

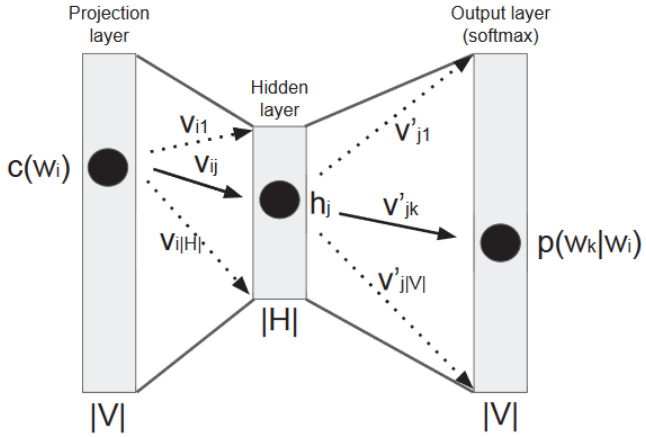
synsets之间最常见的关系是上位词和下位词关系（hypernym vs hyponymy）。

- 上位词+关系表示一个词的词义比另一个词的词义更加泛化，e.g. fruit是apple的上位词。
- 下位词关系表示一个词的词义比另一个词的词义更加具体，e.g. apple是fruit的下位词。

这种关系是具有传递性的。比如摇摇椅是一种椅子，椅子又是一种家具，那么摇摇椅是一种家具。

- Reading thieves’ cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

☐ Core innovation1：Semantic Comparison Model



✓ 核心思路：Word2Vec很好地得到不同单词在同一训练集语料中的相似度

Q：向量很可能不具有可比性：

- 两个不同的语料集简单汇总训练
- 在两个不同的语料集上分别训练词向量模型

A：在**不扩展**隐藏层和输出层的情况下，**将输入层的大小翻倍**

- 让两个不同语料库的同一个单词在训练过程中，能够基于各自的数据集从输入层到隐藏层建立**独立的权重**
- 确保该单词在两个语料库中的上下文能够结合起来，通过隐藏层共同影响神经网络的输出。
- **采用扩展的one-hot编码**，输入的vector维度变为2|V|

$$e(w) = \begin{cases} [v_{zeros}, v_{onehot}(w)] & \text{if } w \text{ is from corpus}_1 \\ [v_{onehot}(w), v_{zeros}] & \text{if } w \text{ is from corpus}_2 \end{cases} \quad (1)$$

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

□ Experiment1：Semantics Comparison Model模型的可靠性

① 两个语料均选择Text8作为模型的输入

当两个词汇的语义信息相同时，它们的余弦相似度应该近似为1，实验结果显示SCM获取的向量组的平均相似度为0.98.

② 同时在一个语料库上进行了两次独立的Word2Vec训练，结果显示，同一个词汇在来自不同的模型时，它的词向量的相似度仅有0.49.

结论：

- 来自两个模型的Word2Vec**词向量的不可比性**
- SCM模型让跨语料的单词可以进行语义相似度比较

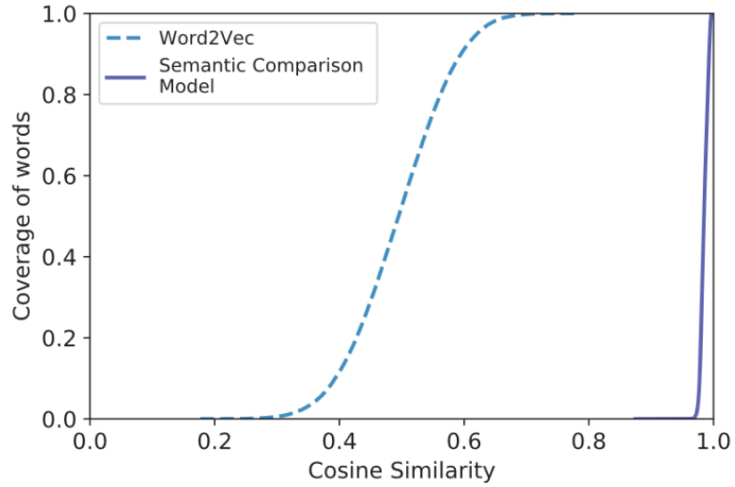


Figure 4: Results of experiment 1 in CDF.

□ Experiment2：Semantics Comparison Model模型的可靠性

- ✓ 将Text8中的某一些词汇进行了替换，形成新的语料库，并对Text8与Text8'两个语料库利用SCM建模，**替换的词相当于原词的jargons。**
- ✓ 发现被替换的词在两个语料中的相似度较低

结论：

- SCM可以捕捉同一单词跨语料的语义差异

Table 2: Results of Experiment 2

replacing word pair	similarity
(chemist → archie)	0.65
(ft → proton)	0.56
(universe → wealth)	0.67
(educational → makeup)	0.66
(nm → famicom)	0.45

□ Experiment3：Semantics Comparison Model生成word vectors的质量

- ✓ 使用了 Tomas Mikolov 提供的代码和测试集来评估单词向量的质量，测试集包含一系列句法和语义关系。
- ✓ 将与 Text8 语料库相关的 SCM 向量与在相同语料库上训练的 Word2Vec 模型生成的向量进行比较。

结论：

- SCM 向量的质量（准确率为 46%）与 Word2Vec 生成的向量质量（50%）相当。
- SCM：跨语料库语义比较带来的好处并没有以牺牲向量质量为代价

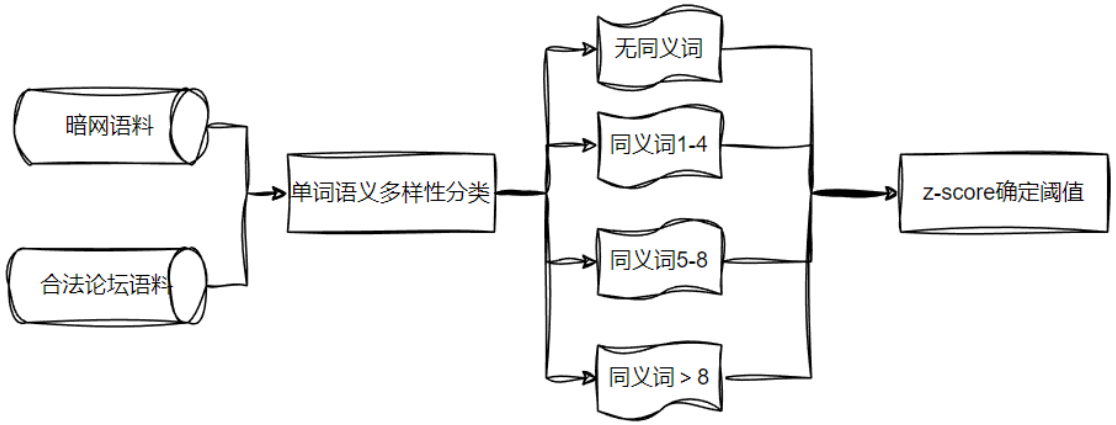
□ Core innovation2：Jargon Discovery

✓ 核心思想：单词在C_{dark}和C_{legit}的相似度（cosine similarity）低于阈值th1，C_{legit}与C_{rep}的相似度高于阈值th2，被判定为暗语。

Q1：因Word2Vec和SCM等嵌入技术依赖单词上下

Q2：存在一些正常词因在合法语料库中的上下文独

Q3：如何**确定阈值**th1与th2？



A1：窗口化上下文（windowed context）

- 给定窗口大小k，统计单词w前k个到后k个

A2：降低误报率

- 不选择C_{rep}，选择C_{legit}，避免单词在正式文档和论坛帖子中用法不同而误报
- 求解C_{legit}与C_{rep}单词的语义相似度，有效去除在C_{legit}中语义特殊但并非暗语的单词

A3：对单词根据其同义词数量进行分类，运用统计方法基于z-score确定阈值

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

☐ Core innovation3：Jargon Understanding

✓ 核心思想：生成候选上位词集合，使用二元随机森林分类器确定暗语与上位词的关系概率

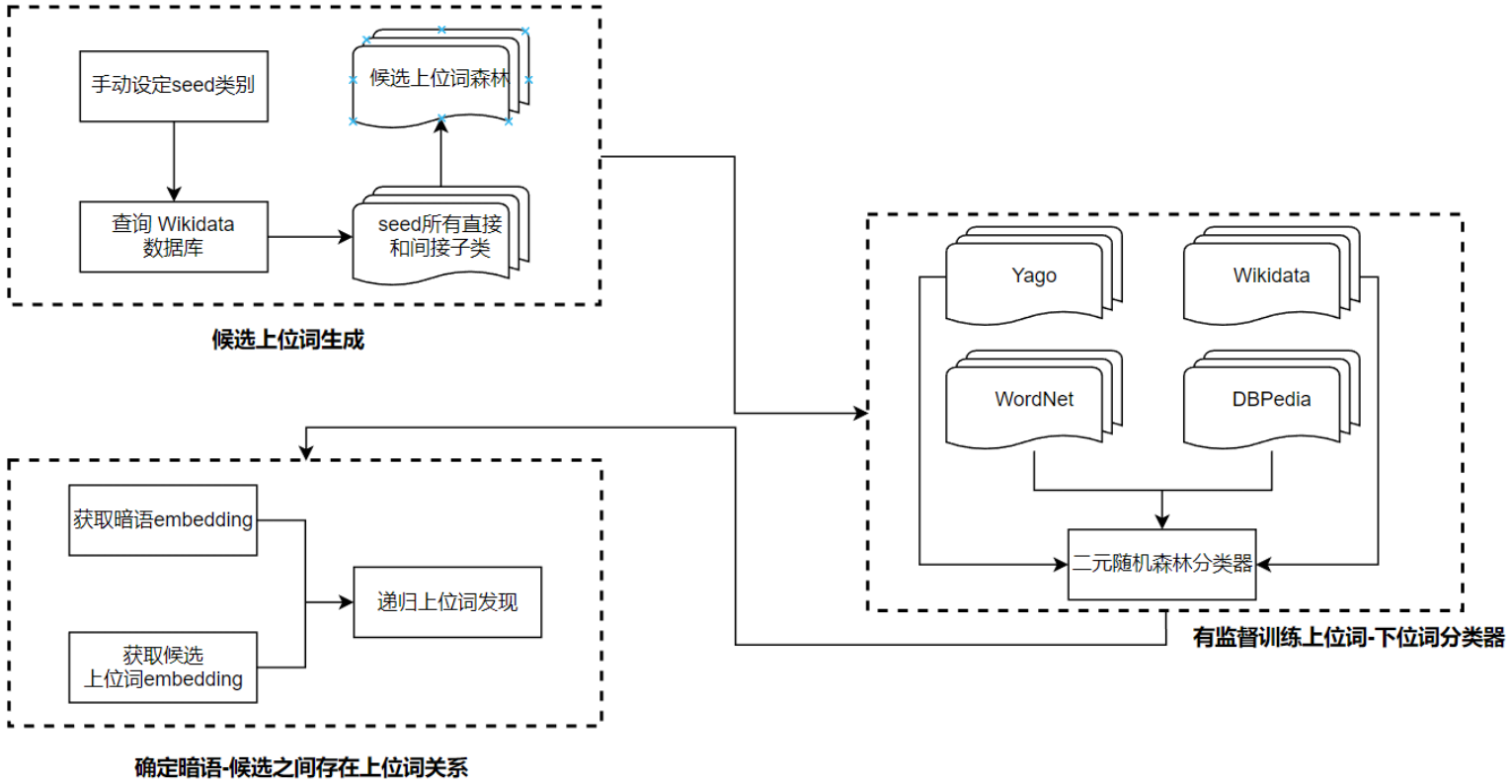


Table 6: Dark jargons in categories

category	# hyper-nyms	# jargons	# traces	examples of jargons
drug	304	736	830,270	blueberry, popcorn, mango
person	1,517	591	460,261	stormtrooper, zulu
software	300	650	512,379	athena, rat, zeus
porn	1	33	2,926	cheesepizza, hardcandy
weapon	672	80	12,055	biscuit, nine, Smith
others	-	1,372	479,789	liberty, ats, omni

```
SELECT ?x ?xLabel WHERE {
  SERVICE wikibase:label {
      bd:serviceParam wikibase:language
          "[AUTO_LANGUAGE],en".
  }
  ?x wdt:P279 wd:Q7397.
}
LIMIT 100
```

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

总结

- ✓ **英文暗语识别的sota**
- ✓ 引领暗语识别技术思想：跨语料相似度对比、向量可比性、seed keywords、抽样人工评测…
- ✓ 只能解决**word-level**的识别问题，而无法处理**phrase-level** jargon detection
- ✓ Jargon Understanding时需要预先设定种子类别，扩展性弱
- ✓ 模型的性能在很大程度上**依赖于所使用的语料库**

展望

- ✓ 利用大语言模型的知识储备进行暗语初始化（越狱攻击？）
- ✓ 借助大语言模型的语义理解优化上位词候选生成（COT/In-context-Learning）
- ✓ …

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

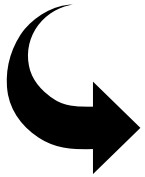
- An Unsupervised Detection Framework for Chinese Jargons in the Darknet, WSDM22(ccf-b)

- A novel cross-domain adaptation framework for unsupervised criminal jargon detection via pre-trained contextual embedding of darknet corpus, Expert Systems with Applications, 2024(SCI 1区, IF=7.5)

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

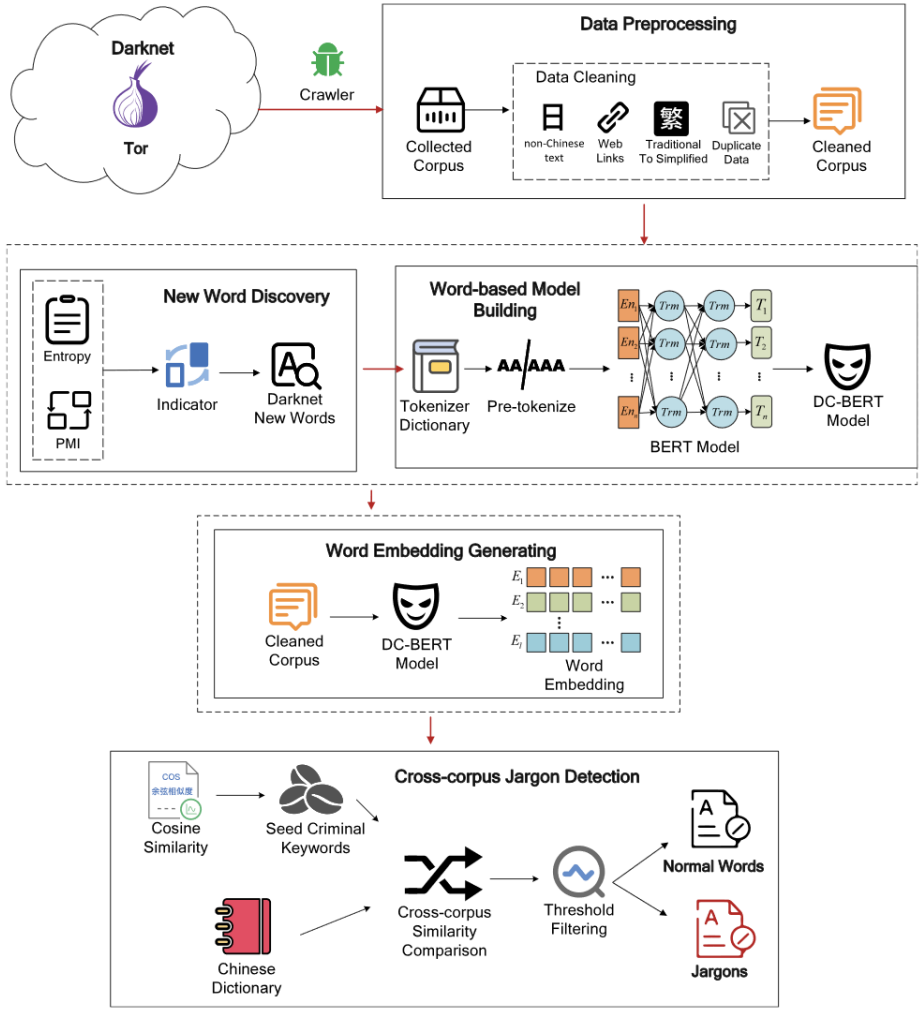
Motivation

- ✓ 中文暗网暗语数据集匮乏（均不开源）
- ✓ 中文暗网中暗语检测方法缺乏研究：大多是根据预先设定的关键词匹配
- ✓ 暗网暗语使用监督学习困难（数据标注难度大、成本高）



- ✓ **构建了中文暗网数据集Darknet Corpus dataset**
- ✓ **无监督跨域适应中文暗语检测框架（CJD - Framework）提出**
- ✓ **大量暗语的发现与模式分析，并部分开源**

Framework



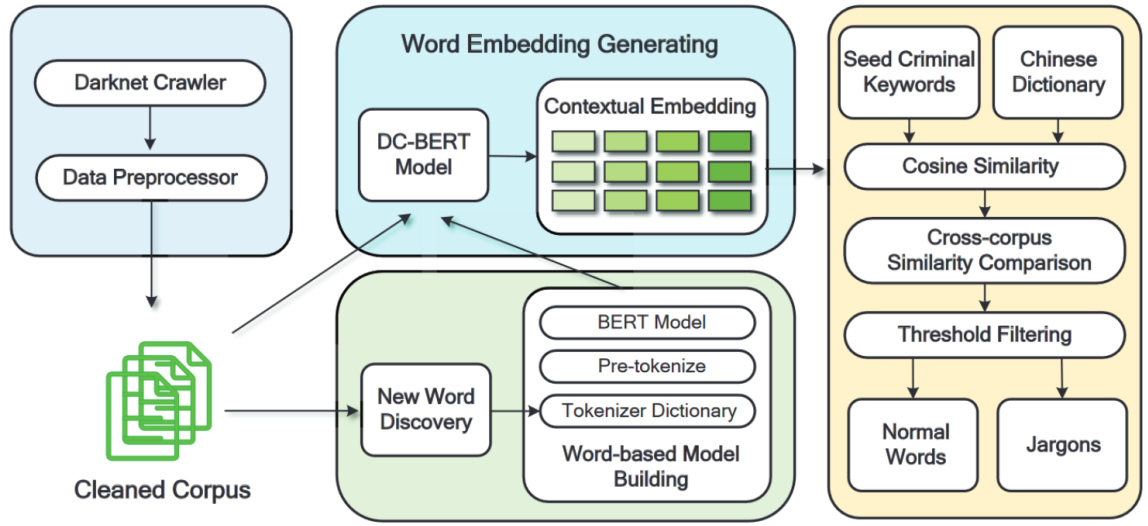


Figure 1: The Framework of Chinese Jargon Detection

① 爬取6个热门中文暗网交易网站，进行数据预处理。

② 发现新词并添加到分词器字典。

③ 选择RoBERTa-wwm-ext作为基础模型，修改tokenizer，在暗网预料上利用MLM任务进行预训练（没有NSP）。

④ 生成jargons的word embeddings

⑤ 跨中文词典&暗网语料求解word余弦相似度，检测jargons

New word discovery（提取未登录词）

- ✓ 暗网语料库中的词很多不在常规词典中（如Jieba）
- ✓ 黑话/行话的搭配需要挖掘（如“轨道料”、“埋包”...）

① 信息熵衡量左右相邻词的丰富度

$$Entropy(w) = -\sum_{w_n \in W_{Neighbor}} P(w_n|w)\log_2 P(w_n|w)$$

② PMI衡量字符组合的内聚度

$$PMI = p\left(c_i, c_j\right) log_2 \frac{p\left(c_i, c_j\right)}{p\left(c_i\right) p\left(c_j\right)}$$

③ 全面判断一个字符组合是否为有效的暗语

$$T_{Score} = PMI + \log \frac{LE * e^{RE} + RE * e^{LE}}{|LE - RE|}$$

An Unsupervised Detection Framework for Chinese Jargons in the Darknet, WSDM22(ccf-b)

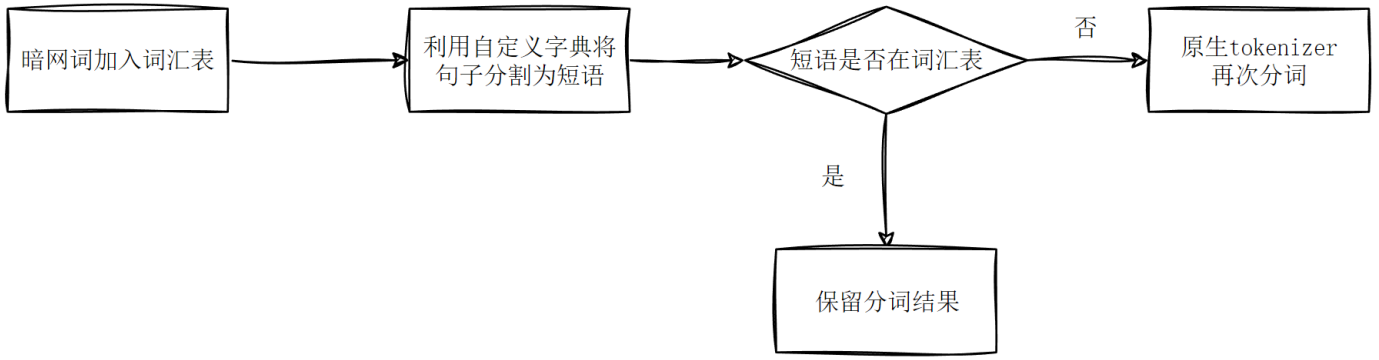
Tokenizer

✓ “预分词”操作实现char-based到word-based

Algorithm 1: DC-BERT Tokenizer Algorithm

Input: Chinese Sentence S, Vocabulary V
Output: Tokenize Result T

```
Load Model Vocabulary V;
W = pre_tokenize(S);
foreach w ∈ W do
    if w ∈ V then
        T.append(w);
    else
        t = tokenize(w);
        T.append(t);
    end
end
```



	Chinese	English
Original Sentence	出少量菠菜数据加TG	Sell small amount spinach data add TG
Segmented Sentence	出/少量/菠菜/数据/加/TG	Sell/small amount/spinach/add/TG
BERT Tokenizer	出/少/量/[M]/菜/数/据/加/[M]/G	Sell/small amount/[M]pinach/add/[M]G
WWM-BERT	出/少/量/[M][M]/数/据/加/[M][M]	Sell/small amount/[M][M]/add/[M][M]
SPAN-BERT	出/少/[M][M][M]/数/据/加/TG	Sell/small [M]/[M]/add/TG
ERNIE	出/[M][M]/[M][M]/数/据/加/[M][M]	Sell/[M][M]/[M]/add/[M]
DC-BERT	出/少量/[M]/数据/加/[M]	Sell/small amount/[M]/add/[M]

□ Word Embedding Generating

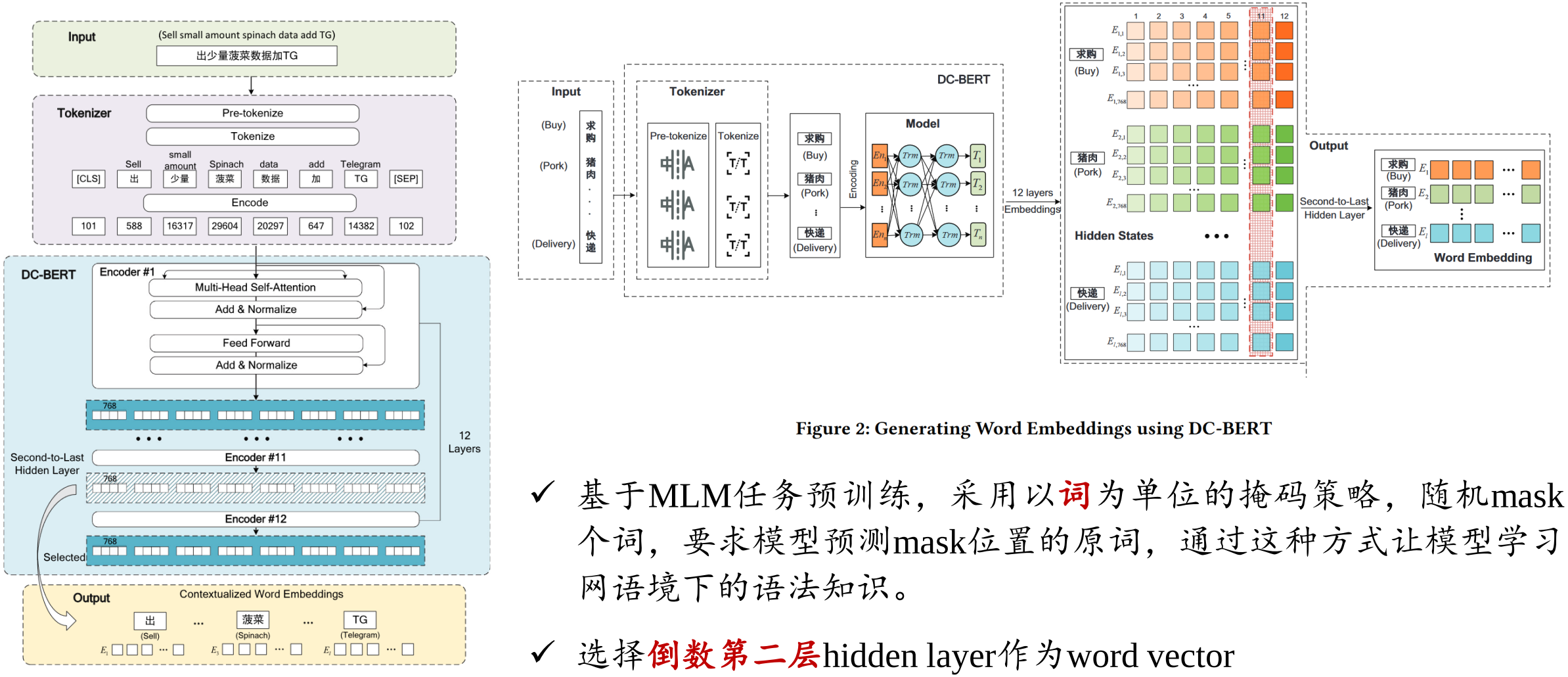


Figure 2: Generating Word Embeddings using DC-BERT

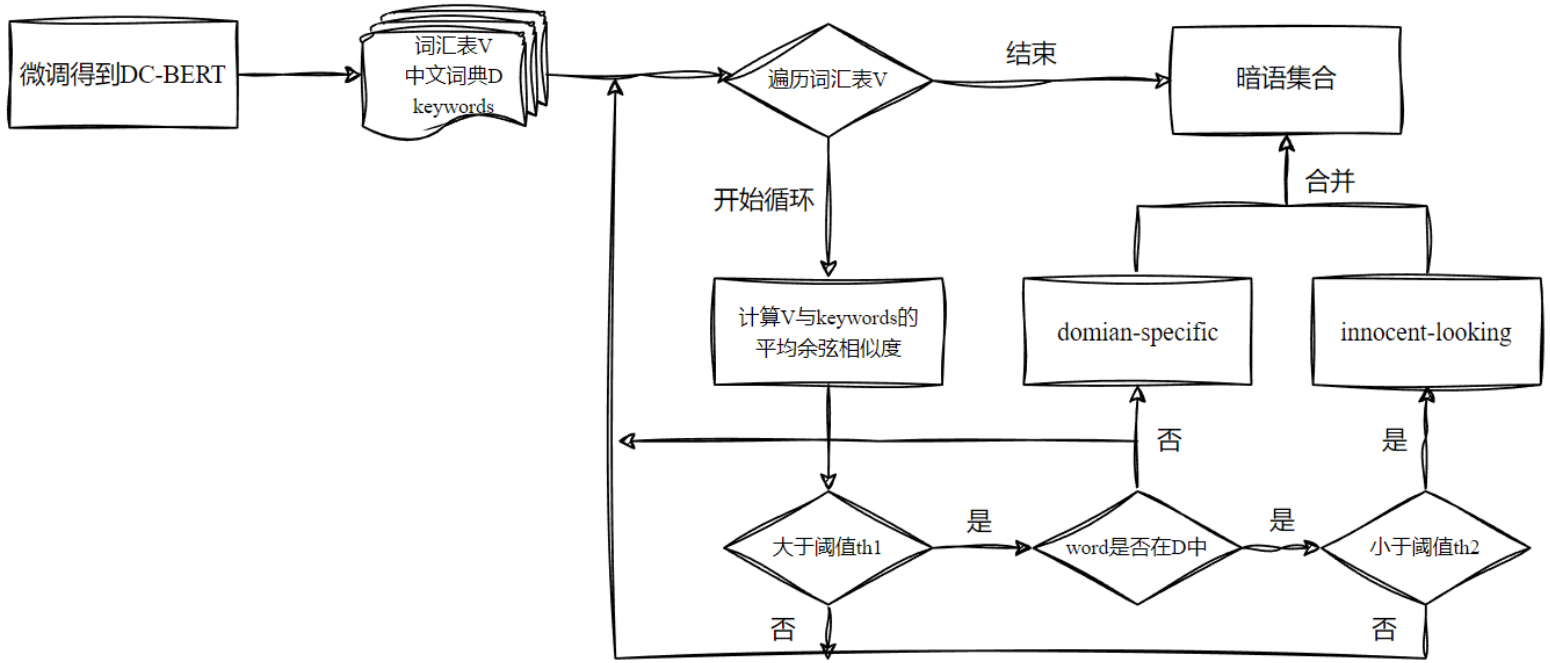
- ✓ 基于MLM任务预训练，采用以**词**为单位的掩码策略，随机mask整个词，要求模型预测mask位置的原词，通过这种方式让模型学习暗网语境下的语法知识。
- ✓ 选择**倒数第二层**hidden layer作为word vector

Cross-corpus Jargon Detection

Algorithm 1: Seed Keyword Screening Algorithm

Input: Dark Corpus C, Vocabulary V, Seed Keywords $Seed$
Output: Similarity Table Sim

```
Load Model Vocabulary V;
foreach w ∈ V do
    S = C.index(w);
    foreach s ∈ S do
        Vec_s = tfidf_mean(s);
        Sim_s = cosine_similarity(Vec_s, Seed);
        Vec_S.append(Sim_s);
    end
    l = max(Vec_S);
    Vec_w = l.index(w);
    Sim_w = cosine_similarity(Vec_w, Seed);
    Sim.append(Sim_w);
end
```



- ✓ 初始**指定部分keywords**
- ✓ TF – IDF得到语料中的sentence向量
- ✓ 计算sentence中每个word与initial keywords的余弦相似度，**筛选出seed keywords**
- ✓ 计算word与seed keywords的平均余弦相似度
- ✓ 基于阈值th1筛选出暗语候选集（**大于th1**）
- ✓ 基于阈值th2（**小于th2**）判断暗语类型（innocent - looking jargons和domain - specific jargons）

Experiment1：Evaluation of new word discovery

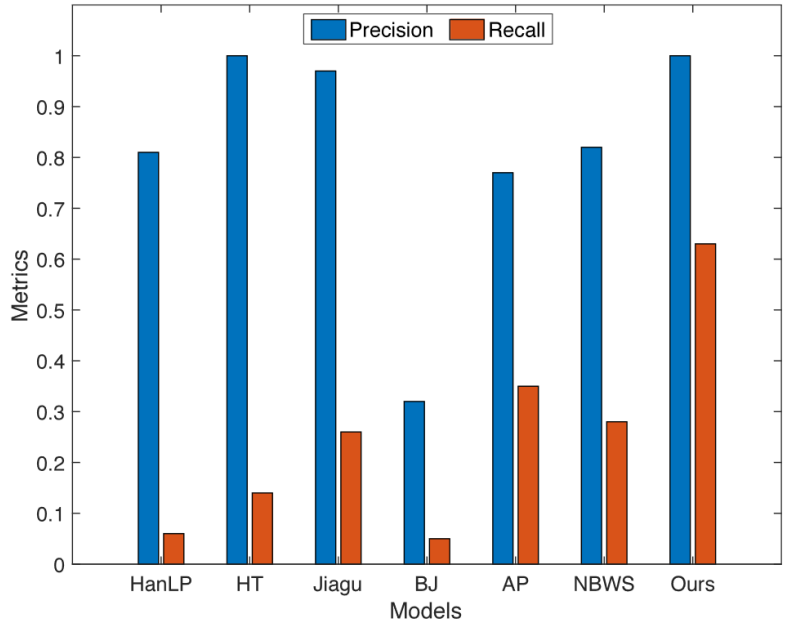


Fig. 3. Evaluation results of New Word Discovery experiment. There are seven models compared in the experiment, namely HanLP, HT, Jiagu, BJ, AP, NBWS and our model. The blue bar indicates the precision of new word discovery and the red bar indicates the recall.

$$I_p(w) = \begin{cases} 1, & w \text{ is a correctly segmented Chinese word;} \\ 0, & \text{otherwise.} \end{cases}$$

$$Precision = \frac{1}{K}\sum_{i=1}^{K} I_p(w_i)$$

$$I_r(w) = \begin{cases} 1, & w \text{ is a domain-specific jargon in a known list;} \\ 0, & \text{otherwise.} \end{cases}$$

$$Recall = \frac{1}{M}\sum_{i=1}^{M} I_r(w_i)$$

✓ 精度100%，召回率63%，基于“自由度”和“凝聚度”的新词发现对 100 个特定领域黑话的发现效果最好

An Unsupervised Detection Framework for Chinese Jargons in the Darknet, WSDM22(ccf-b)

□ Experiment2：Evaluation of word embeddings

Table 5
Performance of different word embeddings. The bold numbers are the best performance under each metric.

Models	$\overline{P@10}$	$\overline{P@20}$	$\overline{P@30}$	$\overline{P@40}$	$\overline{P@50}$	$\overline{P@60}$	$\overline{P@80}$	$\overline{P@100}$	$MAP@100$
W2V+CBOW	0.54	0.52	0.43	0.45	0.39	0.38	0.39	0.31	0.44
W2V+SG	0.51	0.50	0.41	0.37	0.40	0.36	0.32	0.27	0.41
LDA	0.37	0.33	0.23	0.19	0.21	0.18	0.15	0.13	0.25
ELMo	0.72	0.65	0.59	0.54	0.57	0.53	0.48	0.43	0.56
Wobert	0.78	0.65	0.67	0.58	0.59	0.60	0.51	0.47	0.62
DC-BERT	**0.89**	**0.75**	**0.80**	**0.77**	**0.78**	**0.78**	**0.66**	**0.61**	**0.79**

- ✓ 与Wobert相比：DC-BERT在暗网语料上的预训练，能更好地捕捉特定语义和语法
- ✓ 与ELMo相比：transformer完胜LSTM
- ✓ 与LDA、Word2Vec相比：动态词向量优于静态词向量

$$I_{w_0}(w) = \begin{cases} 1, & w \text{ is semantically related to } w_0; \\ 0, & \text{otherwise.} \end{cases}$$

$$P@K\left(w_g\right) = \frac{\sum_{i=1}^{K} I_{w_g}\left(w_i\right)}{K}$$

$$AP@K\left(w_g\right) = \frac{\sum_{k=1}^{K} P@k\left(w_g\right) \times I_{w_g}\left(w_k\right)}{\sum_{k=1}^{K} I_{w_g}\left(w_k\right)}$$

$$\text{performance} = \overline{P@K} = \frac{\sum_{i=1}^{30} P@K\left(w_{g_i}\right)}{30}$$

$$MAP@K = \frac{\sum_{i=1}^{30} AP@K\left(w_{g_i}\right)}{30}$$

□ Experiment3：Evaluation of jargon candidates finding ability

Table 6
Performance of finding jargon candidates. The bold numbers are the best performance under each metric.

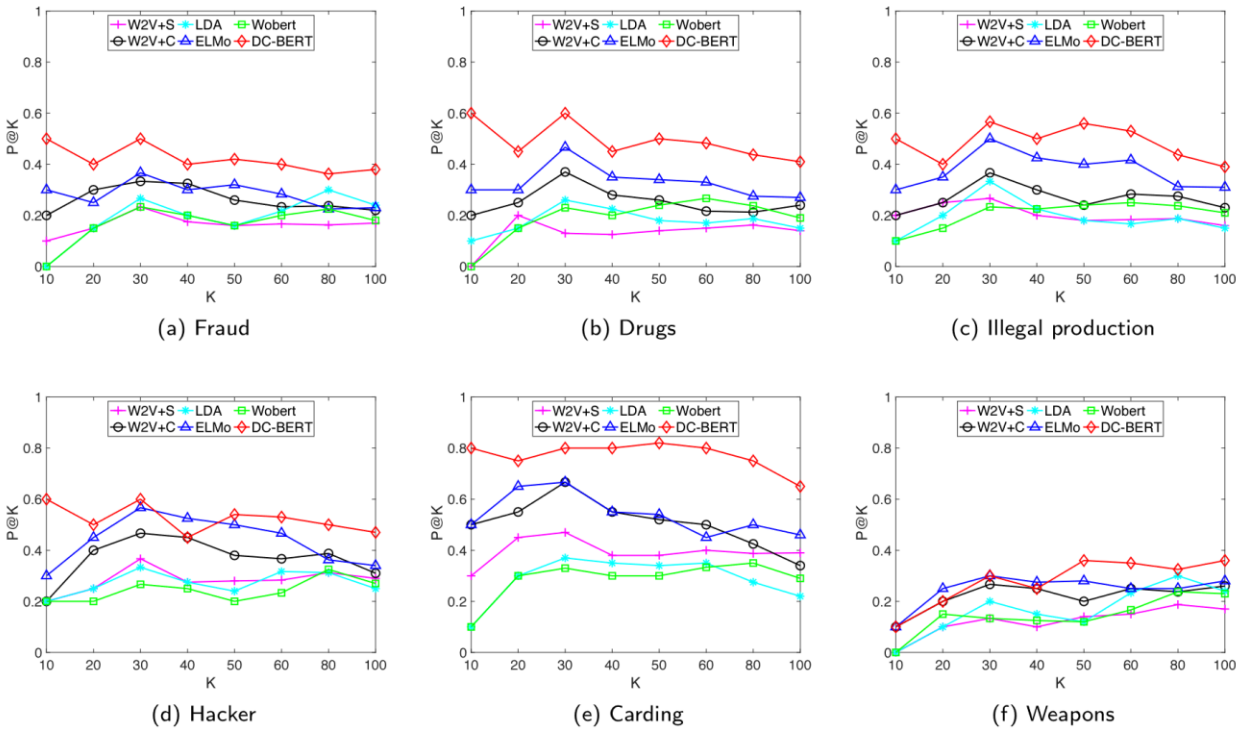
Models	$\overline{P@10}$	$\overline{P@20}$	$\overline{P@30}$	$\overline{P@40}$	$\overline{P@50}$	$\overline{P@60}$	$\overline{P@80}$	$\overline{P@100}$	$MAP@100$
W2V+CBOW	0.29	0.38	0.47	0.42	0.37	0.35	0.34	0.32	0.38
W2V+SG	0.17	0.28	0.31	0.25	0.26	0.27	0.27	0.26	0.29
LDA	0.13	0.23	0.29	0.25	0.22	0.27	0.30	0.24	0.26
ELMo	0.33	0.42	0.51	0.41	0.43	0.38	0.36	0.35	0.41
Wobert	0.12	0.22	0.27	0.24	0.23	0.25	0.29	0.25	0.24
DC-BERT	**0.58**	**0.55**	**0.63**	**0.52**	**0.56**	**0.55**	**0.49**	**0.46**	**0.57**

- ✓ Input：给定10个seed keywords
- ✓ Output：每个keyword余弦相似度的top k words
- ✓ Evaluate：人工判断是否是jargons

$$I_{w_0}(w) = \begin{cases} 1, & w \text{ is a jargon;} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{performance} = \overline{P@K} = \frac{\sum_{i=1}^{10} P@K\left(w_{g_i}\right)}{10}$$

$$MAP@K = \frac{\sum_{i=1}^{10} AP@K\left(w_{g_i}\right)}{10}$$



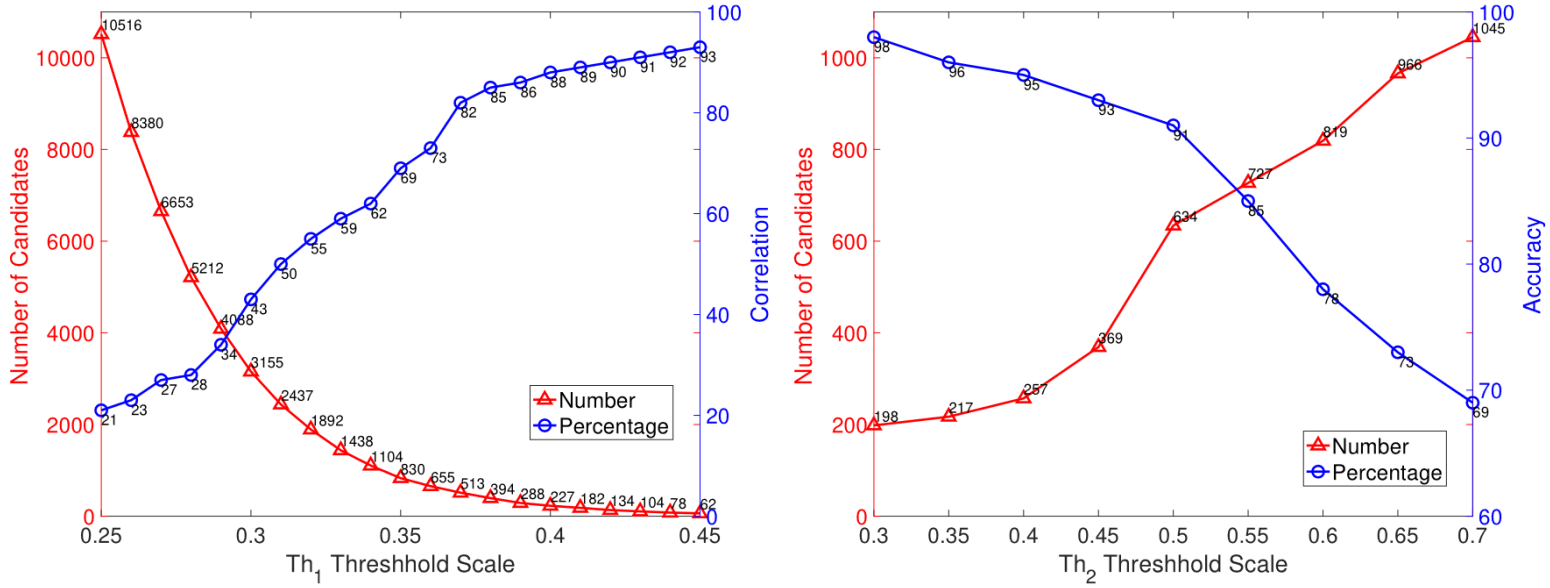
(a) Fraud (b) Drugs (c) Illegal production

(d) Hacker (e) Carding (f) Weapons

□ Experiment4：Evaluation of Chinese jargon detection framework

- ✓ 没有ground-truth dataset → **随机采样200个jargons + 人工判断**
- ✓ Jargons认定的权威性 → 三个专家众包并进行一致性检验
- ✓ 整体效果：183/200，**Accuracy=91.5%**
- ✓ domain-specific：94%，受限于分词错误；innocent-looking：89%，受限于跨语料对比

- ✓ Th1：过滤掉与keywords不相关的词
- ✓ Th2：确定innocent-looking jargons



□ Experiment5：Evaluation of domain adaptation

✓ 在一个domain dataset上pre-train（如色情），在另一个domain dataset上test（如黑客）

Table 7
Statistics of different domains in the DC-dataset.

	Pornography Industry (PI)	Data Trading (DT)	Hacking Technology (HT)	Black Industry (BI)
# traces	10,411	16,925	6,583	5,665
# words	1,024,804	1,209,134	621,279	592,123

Table 8
Experimental results of cross-domain jargon detection. The *PI → DT* means training the model on Pornography Industry (PI) dataset and testing it on Data Trading (DT) dataset. The bold numbers are the best performance under each metric.

Models	*PI → DT*			*DT → PI*			*HT → BI*			*BI → HT*			*DT → BI*			*BI → DT*		
	Pre	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*
SCM	0.12	0.08	0.10	0.13	0.15	0.14	0.06	0.07	0.06	0.13	0.16	0.14	0.17	0.15	0.16	0.06	0.04	0.05
SelfEuph	0.39	0.38	0.38	0.37	0.32	0.34	0.37	0.35	0.36	0.32	0.34	0.33	0.31	0.33	0.32	0.24	0.32	0.27
SentEuph	0.13	0.13	0.13	0.19	0.15	0.17	0.22	0.20	0.21	0.17	0.19	0.18	0.17	0.23	0.20	0.14	0.12	0.13
PET	0.23	0.22	0.22	0.21	0.21	0.21	0.23	0.28	0.25	0.26	0.25	0.25	0.27	0.31	0.29	0.14	0.22	0.17
CJD	**0.45**	**0.43**	**0.44**	**0.55**	**0.45**	**0.50**	**0.66**	**0.57**	**0.61**	**0.63**	**0.54**	**0.58**	**0.69**	**0.71**	**0.70**	**0.51**	**0.40**	**0.45**

✓ 跨域暗语检测任务，**所有模型领域适应表现不佳** → **domain-specific jargons具有自身独特的语言风格**

□ Experiment6：Evaluation of jargon evolution

✓ 根据时间戳划分数据集，在previous time上pre-train，在later time上test

Table 9
Statistics of DC-dataset by year.

Category \ Year	2018	2019	2020
# traces	2,918	4,446	27,426
# words	232,855	375,072	2,412,636

Table 10
Experimental results of detecting jargons over time. The 2018 → 2019 means training the model on the dataset of the year 2018 and testing it on the dataset of the year 2019, while 2018+19 denotes the dataset of both the year 2018 and 2019. The bold numbers are the best performance under each metric.

Models	2018 → 2019			2018 → 2020			2019 → 2020			2018 → 2019 + 20			2018 + 19 → 2020		
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1	*Pre*	*Rec*	*F*1	*Pre*	*Rec*	*F*1	*Pre*	*Rec*	*F*1
SCM	0.29	0.35	0.32	0.31	0.32	0.31	0.33	0.27	0.30	0.27	0.36	0.31	0.34	0.38	0.36
SelfEuph	0.62	0.68	0.65	0.67	0.62	0.64	0.57	0.63	0.60	0.54	0.52	0.53	0.64	0.68	0.66
SentEuph	0.38	0.28	0.32	0.32	0.35	0.33	0.32	0.37	0.34	0.29	0.24	0.26	0.38	0.41	0.39
PET	0.47	0.52	0.49	0.51	0.53	0.52	0.53	0.58	0.55	0.46	0.55	0.50	0.59	0.57	0.58
CJD	**0.77**	**0.80**	**0.78**	**0.75**	**0.82**	**0.78**	**0.78**	**0.76**	**0.77**	**0.71**	**0.73**	**0.72**	**0.81**	**0.85**	**0.83**

✓ 模型对暗语演变有**一定的鲁棒性**

□ Experiment7：Evaluation of platform transferring

- ✓ 不同暗网平台有其特有的“业务范围”、语言风格等
- ✓ 测试模型对未知平台数据的鲁棒性：在其中一个平台pre-train，在另一个平台test

Table 11
Experimental results of cross-platform jargon detection. *EC → Ali* means training the model on the dataset of “Exchange Trading Market” and testing it on the dataset of “Alibaba”. The bold numbers are the best performance under each metric.

Models	*EC → Ali*			*Ali → TH*			*TH → DC*			*DC → DT*			*DT → LC*			*LC → EC*		
	Pre	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*
SCM	0.28	0.31	0.29	0.29	0.30	0.29	0.33	0.29	0.31	0.29	0.36	0.32	0.33	0.29	0.31	0.28	0.27	0.27
SelfEuph	0.67	0.71	0.69	0.71	0.73	0.72	0.63	0.73	0.68	0.64	0.69	0.66	0.73	0.68	0.70	0.62	0.66	0.64
SentEuph	0.29	0.35	0.32	0.34	0.32	0.33	0.28	0.34	0.31	0.32	0.25	0.28	0.36	0.33	0.34	0.27	0.22	0.24
PET	0.38	0.42	0.40	0.45	0.39	0.42	0.40	0.41	0.40	0.43	0.46	0.44	0.47	0.44	0.45	0.36	0.32	0.34
CJD	**0.82**	**0.73**	**0.77**	**0.77**	**0.82**	**0.79**	**0.75**	**0.72**	**0.73**	**0.82**	**0.76**	**0.79**	**0.85**	**0.81**	**0.83**	**0.71**	**0.72**	**0.71**

- ✓ 基于情感分析的暗语识别方法跨平台表现差：jargons中性 or jargons跨平台情感属性差异大
- ✓ 模型具有一定的面对未知平台暗语检测的鲁棒性

Category	Jargons	Explanation	Example
Gamble	葡京(lisboa) 菠菜(spinach) 赌狗(gambling dog)	The name of a gambling site Pinyin homonym of "博彩"(Gamble) People who are addicted to gambling	新**葡京**脱裤43W低价出 **菠菜**后台数据5.7万多条 出售一千万条**赌狗**数据
Carding	鱼料(fish material) 小丑(joker) 手拨料 (hand dial material)	• Bank card information obtained from phishing site • The name of a site that sells bank card information • Information verified by manual dialing	求购一手US,AU**鱼料**CC CVV知名料站**小丑**螃蟹等 长期出售**手拨料**实时贷款料

- Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces, USENIX Security symposium, 2018

- An Unsupervised Detection Framework for Chinese Jargons in the Darknet, WSDM22(ccf-b)

- A novel cross-domain adaptation framework for unsupervised criminal jargon detection via pre-trained contextual embedding of darknet corpus, Expert Systems with Applications, 2024(SCI 1区, IF=7.5)

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

Motivation

- ✓ 高质量的词向量需要大规模标注语料训练，但**获取大规模暗语语料很难**
- ✓ 大多暗语检测需要**预先设定暗语类别**，扩展性差，漏掉新型类别的暗语
- ✓ 大多暗语检测**依赖ground-truth jargon list**，无法应对暗语的隐蔽性和动态性
- ✓ 对中文暗语识别研究还较少，基于英文暗语识别的检测方法无法直接应用到汉语

采用迁移学习生成暗语词向量

不依赖于ground-truth jargon list

多维度特征的暗语筛选机制

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

Framework

✓ Corpus Preparation Module

- 构建Telegram 地下市场中文语料库TUMCC（**两个人工标注**）
- 构建口语中文语料库（OCC）和解释性中文语料库（ICC）

✓ Lexical - analysis - based Features

- Telegram词典构建：统计单词出现次数与上下文计数（**窗口机制**）
- 词性标注：**保留名词和动词**，滤掉与暗语无关的词类（如介词）

✓ Vectors - based Features

- 对于 OCC 和 ICC（**百万级**数据）：**直接使用Glove**生成词向量
- 对于 TUMCC：Tencent vector初始化，VCDM迁移学习
- 采用转移矩阵方法将三类词向量投影到**共享空间**，计算彼此相似度

✓ Dictionary analysis - based Features

- 筛选TUMCC语料中每个词的Top20同义词
- 利用OpenHowNet字典计算单词与top20同义词的相似度，对比阈值

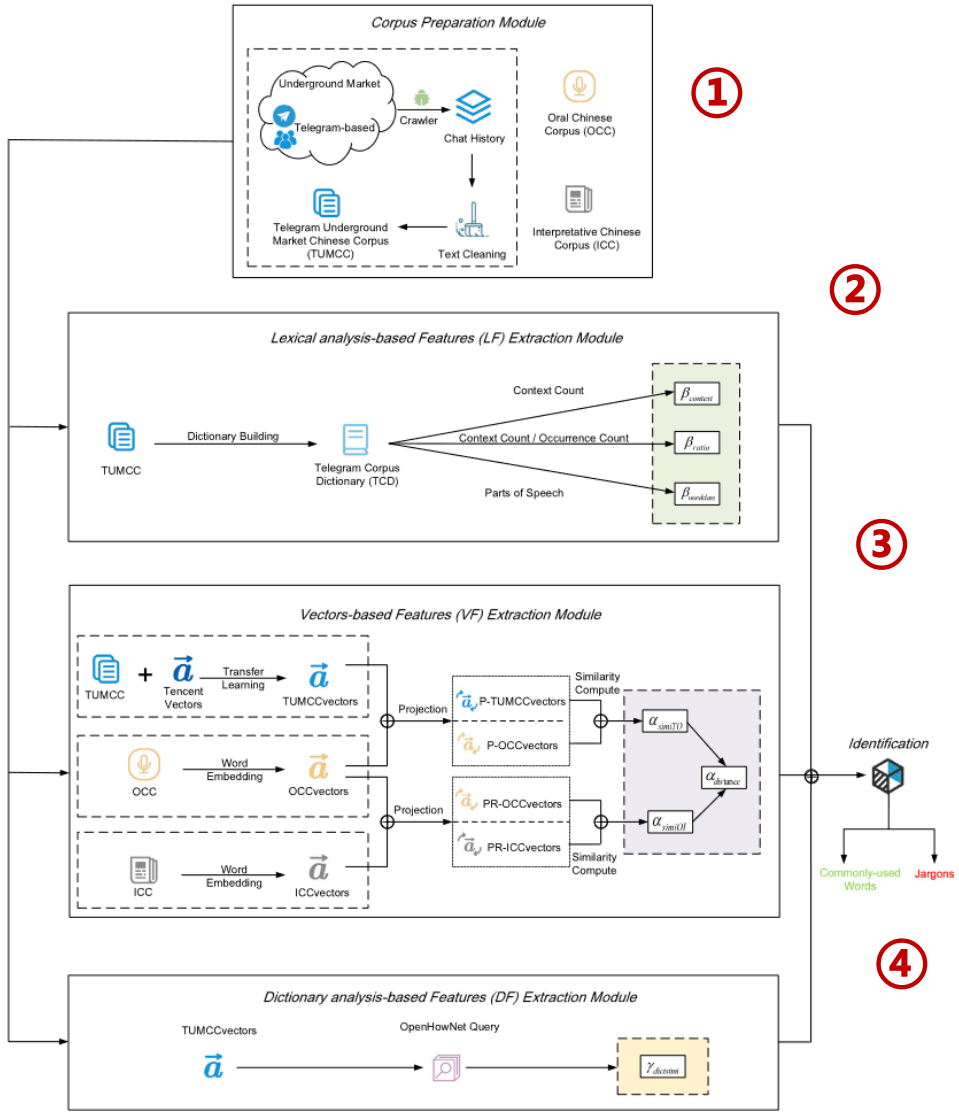
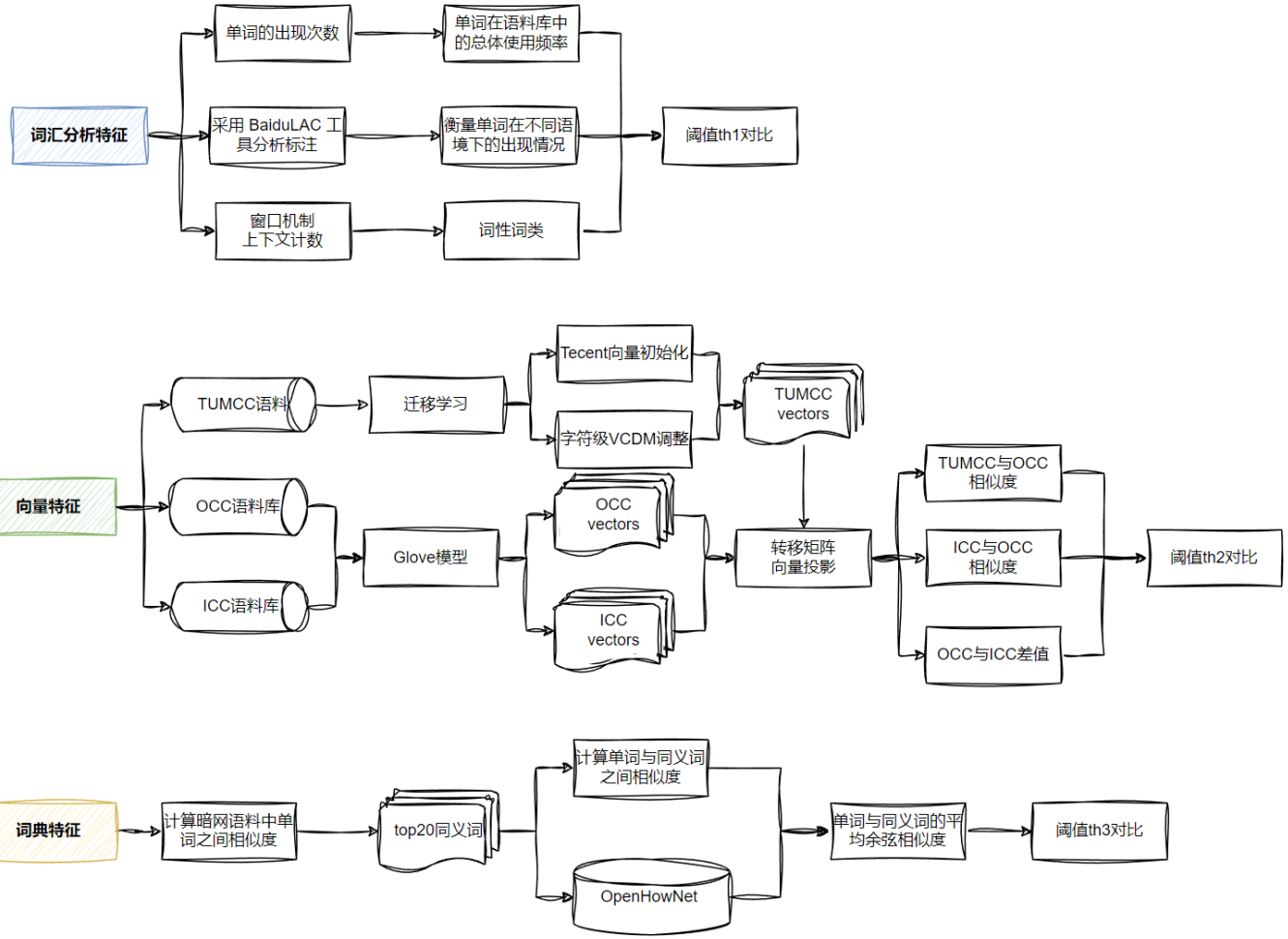


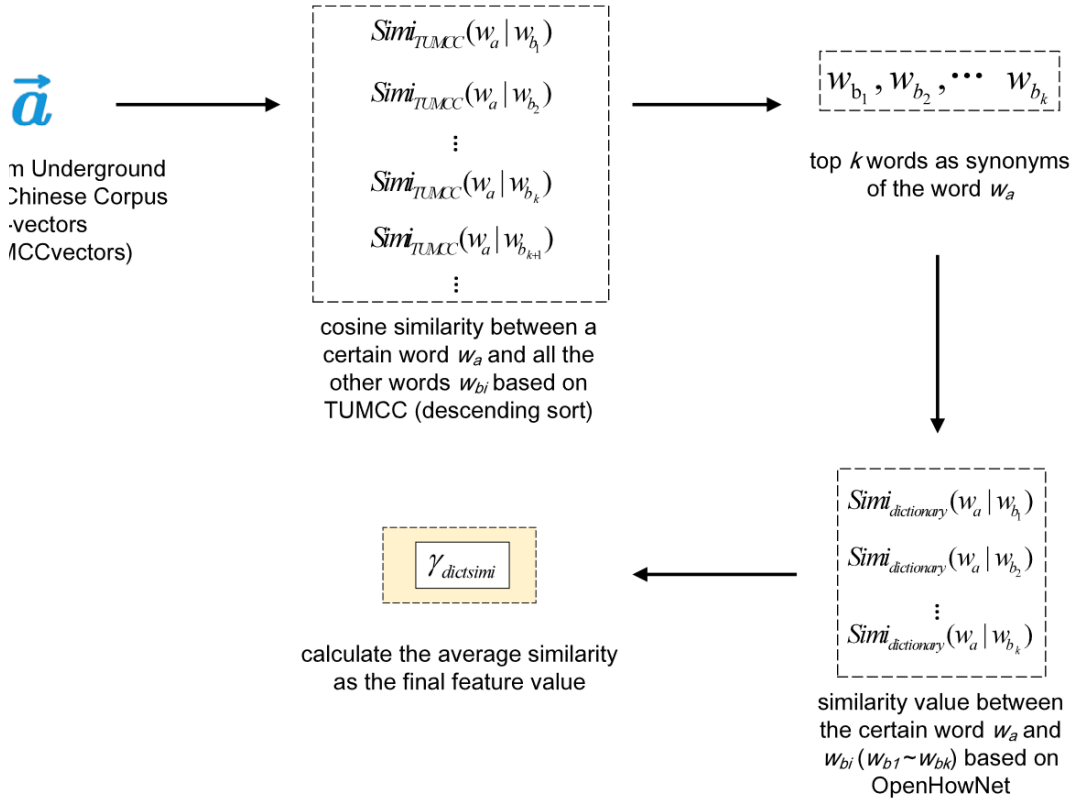
Fig. 1. The proposed Chinese Jargons Identification Framework, CJI-Framework.

□ Core innovation：多特征挖掘 & 阈值判断

✓ 只有当一个单词满足**所有特征的阈值**要求时，才会被判定为暗语。

✓ 多特征结合和严格阈值判断机制鲁棒性高，避免因单个特征偶然波动而误判。





of the feature extraction using OpenHowNet in the *Dictionary analysis-based Features (DF) Extraction Module.*

□ Experiment1：Evaluation of feature effectiveness

- ✓ **特征组合**能够有效地识别暗语，充分发挥了各自的优势，相互补充和协作，从不同角度准确地捕捉暗语的特征。
- ✓ 基于**向量的特征**在暗语识别中起到了**最为关键**的作用，词法分析特征次之，词典特征最后。

Table 1
Overview of the proposed jargons identification features.

Category	Feature description	Feature symbol
Vectors-based Features (VF)	Cosine similarity of P-TUMCCvectors and P-OCCvectors	α_{simiTO}
	Cosine similarity of PR-OCCvectors and PR-ICCvectors	α_{simiOI}
	The absolute value of the difference between α_{simiTO} and α_{simiOI}	α_{distance}
Lexical analysis-based Features (LF)	The count of context conditions (the *context count*)	β_{context}
	The ratio of the *context count* and the *occurrence count*	β_{ratio}
	Parts of speech (the *word class*)	$\beta_{\text{wordclass}}$
Dictionary analysis-based Features (DF)	The dictionary analysis results based on OpenHowNet	γ_{dictsimi}

Table 7
Jargons identification results of different feature sets.

Features sets	Categories of features included	Accuracy	Precision	Recall	F1-score
F	VF+LF+DF	87.50	92.86	86.67	89.66
F/DF	VF+LF	87.30	76.47	68.42	72.22
F/LF	VF+DF	65.08	45.45	78.95	57.69
F/VF	LF+DF	55.56	39.53	89.47	54.83

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

□ Experiment2：Evaluation of the proposed framework

✓ 中文暗网语料

- SCM不适合中文暗语检测
- 词法分析特征与词典特征模块提升暗语检测性能

✓ 英文暗网语料

- CJI框架**召回率低**：相对严格的暗语检测策略导致（三个特征均超过均值才认定为暗语）
- CJI模型框架具有跨语言适应性

Table 9
The performance of jargons identification approaches for Chinese and English.

Approach	Corpora composition	Corpus language	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SCM	Telegram Underground Market Chinese Corpus, Oral Chinese Corpus, Interpretative Chinese Corpus.	Chinese	14.12	12.64	**95.45**	22.32
SCM+LF+DF			80.79	32.35	50.00	39.28
SCM+LF			79.10	35.29	81.82	49.31
SCM+DF			79.66	36.00	81.82	50.00
CJI-Framework			**87.50**	**92.86**	86.67	**89.66**
SCM	Jargon Corpus, Oral Corpus, Interpretative Corpus.	English	62.41	61.39	**81.58**	70.06
CJI-Framework (adapted)			**70.92**	**70.11**	80.26	**74.85**

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

口 总结

- ✓ 也用到了单词上下文的跨语料差异辅助检测：
  - 引入ICC（口语）和OCC（偏书面），对比jargon在暗网和词典中的word embedding vector区别
  - 引入汉语词典OpenHowNet，对比jargon与同义词在暗网和词典中的相似度
- ✓ 针对语言本身的特点提出合适的句法和语义特征分析方法（如对汉语、英文）
- ✓ 不需要预先定义好的jargon list作为ground-truth

口 展望

- ✓ 跨语言的暗语检测技术依然缺乏 → 适应多语言的暗语检测平台
- ✓ 同时满足三种特征阈值才被认定为jargon
  - **苛刻的筛选机制会导致暗语漏检，召回率降低**
  - 阈值的设定至关重要（大模型辅助？）
  - 三种特征是否应该不同权重？ → 更合理且结果更具可解释性

- Identification of Chinese dark jargons in Telegram underground markets using context-oriented and linguistic features, Information Processing & Management, 2022(SCI 1区, IF=7.4)

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

口 研究团队

王海舟
四川大学网安学院副教授/院长助理
Citations：770
h-index：14
i10-index：22
http://www.cyber-wang.cn

陈贞翔
济南大学信息科学与工程学院院长
Citations：3483
h-index：30
i10-index：62
https://ujnloci.github.io/zh/

2. Liang Ke, Peng Xiao, Xinyu Chen, Shui Yu, Xingshu Chen, **Haizhou Wang***. A Novel Cross-domain Adaptation Framework for Unsupervised Criminal Jargon Detection via Pre-trained Contextual Embedding of Darknet Corpus [J], *Expert Systems with Applications*, 2024, 242: 122715. [Code] [Dataset] [PDF] sci一区

3. Tingke Wen, Yuanxing Xiao, Anqi Wang, **Haizhou Wang***. A Novel Hybrid Feature Fusion Model for Detecting Phishing Scam on Ethereum using Deep Neural Network [J], *Expert Systems with Applications*, 2023, 211: 118463.

4. Yifei Wang, Haochen Su, Yingao Wu, **Haizhou Wang***. A Supervised-based Identification and Classification Model for Chinese Jargons Using Feature Adapter Enhanced BERT [C], Proceeding of 19th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2022), Shanghai, China, November 10-13, 2022, pp. 297-308. [Code] [Dataset] [PDF] ccf-c

5. Yiwei Hou, Hailin Wang, **Haizhou Wang***. Identification of Chinese Dark Jargons in Telegram Underground Markets Using Context-Oriented and Linguistic Features [J], *Information Processing and Management* , 2022, 59(5): 103033. [Code] [Dataset] [PDF] sci一区

6. Liang Ke, Xinyu Chen, **Haizhou Wang***. An Unsupervised Detection Framework for Chinese Jargons in Darknet [C], Proceeding of 15th ACM International Conference on Web Search and Data Mining (WSDM 2022), Phoenix, Arizona, USA, February 21-25, 2022, pp. 458–466. [Code] [Dataset] [PDF] ccf-b

7. Wenjing Zeng, Rui Tang, **Haizhou Wang***, Xingshu Chen, Wenxian Wang. User Identification Based on Integrating Multiple User Information across Online Social Networks [J], *Security and Communication Networks*, 2021: 5533417.

8. Hailin Wang, Yiwei Hou, **Haizhou Wang***. A Novel Framework of Identifying Chinese Jargons for Telegram Underground Market [C]. Proceeding of 30th IEEE International Conference on Computer Communications and Networks (ICCCN 2021), Athens, Greece, July 19-22, 2021, pp. 1-9. [Code] [Dataset] [PDF] ccf-c

Low-Frequency Aware Unsupervised Detection of Dark Jargon Phrases on Social Platforms ccf-c

Limei Huang, Shanshan Wang, Changlin Liu, Xueyang Cao, Yadi Han, Shaolei Liu, and Zhenxiang Chen(✉)

School of Information Science and Engineering, University of Jinan, Jinan, China
{lmhuang,xueyangcao}@stu.ujn.edu.cn, {ise_wangss,zxchen}@ujn.edu.cn,
15253464198@163.com, yadi@163.com, lcl_go@163.com

Thank you !