

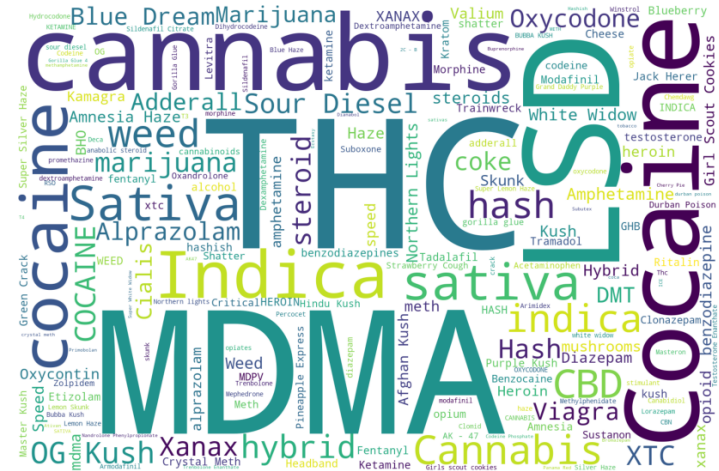
EXPLORING TRANSFER LEARNING TECHNIQUES FOR NAMED ENTITY RECOGNITION IN NOISY USER-GENERATED TEXT

Johannes Bogensperger

Supervised by:

Prof. Alan Hanbury

Dr. Gabor Recski




PROJECT COPKIT

"peruane 92 cocain very good snife and for people tghe want cooking out very good smills good frech from the block orginail good good good AAA."

– drug reseller on the DreamMarket

224g of Mac 1, Supplementary Light Greenhouse, Near Indoor Quality!



Mac 1

Category: Drugs -> Cannabis - Buds and Flowers

Price (Fiat): USD 990 (€832.59 £721.35 AUD1304.18 CAD1240.21)

Price (XMR): 2.995914661824

Measurement unit: Pound

Shipping: from: United States to: United States

Views: 5

Shipping methods:

- FREE : USD 0 (XMR 0.000000000000)

Available: In stock

Vendor: 98.80 % positive / 250 reviews Disputes: 0 won / 0 lost [400 - 410 sales]

Finalize early (FE): Listing is Escrow

Vendor last seen: Today

Imported Feedback: Empire: 99.52% / 2314 sales.

Minimum order amount: XMR 2.995914661824 (2.995914661824 for products + 0.000000000000 for shipping).

Vendor's PGP key fingerprint: [REDACTED] Show Key

Listing Description

Half Pound

224 grams

Supplementary Light Greenhouse, Near Indoor Quality!

Mac 1

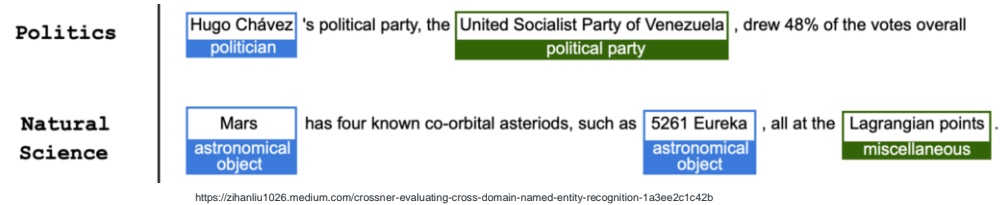
Also known as Miracle Alien Cookies, this strain has had more hype in the boutique cannabis world than almost anything aside from runtz. The indica dominated high packs a serious punch and isn't for the faint of heart. This limited batch was grown with supplementary lights. It's basically indoor that's grown in a greenhouse. The unusual look and unique nose is well-known to people who have enjoyed the Mac 1.

Real Item listing from the White House Market – accessed on the 12.04.2021

<http://voh5342e6nnxieprcsafwk5rphgfbbrq2ihftfzrjsbndmb6u5jx4id.onion/welcome>

PROBLEM DESCRIPTION

- Named Entity Recognition (NER)
- Challenges
 - Noisy user-generated data
 - Uncommon entity type
 - Data sparsity issues



10x Shroom Blotter | 200ug LSD | Special Offer

With this listing you get some of our Shroom LSD Drug Blotters amazing quality Made in Netherlands 200 ug LSD Drug We promise Fast and secure shipping delivery to all EU countrys 100% Escrow

RESEARCH QUESTIONS

1. Can we improve NER models by pre-training on different NER dataset?
2. Can Language Models for noisy text benefit from pre-training on well-structured text?
3. Can distantly supervised datasets boost the performance of a NER system with different text structure?
4. Does the aforementioned transfer learning NER model show a competitive edge against state of the art NER models?

DATA SOURCES

- Only a single Darknet NER dataset available from Al. Nabki et. Al 2019 [3-5] - NuToT
- DreamMarket & Grams
- Creation of a new dataset
 - Regex-based datasets
 - Distantly supervised dataset
 - Manually create dataset
 - Crowd-Sourcing



DARKNET MARKET ARCHIVES (2013-2015)

SITE

Mirrors of ~89 Tor-Bitcoin darknet markets & forums 2011-2015, and related material



CROWD SOURCING

- Annotation Guidelines are a first class citizen
- Pilot Studies are fundamental
- Static pre-processing required
- Quality Assurance

I. Project Definition

- 1a. Select NLP Problem and crowdsourcing genre
- 1b. Decompose NLP problem into tasks
- 1c. Design crowdsourcing task

III. Project Execution

- 3a. Recruit and screen contributors
- 3b. Train, profile and retain contributors
- 3c. Manage and monitor crowdsourcing tasks

II. Data Preparation

- 2a. Collect and pre-process corpus
- 2b. Build or reuse annotator and management interfaces
- 2c. Run pilot studies

IV. Data Evaluation and Aggregation

- 4a. Evaluate and aggregate annotations
- 4b. Evaluate overall corpus characteristics

DATA PREPARATION

- DreamMarket SQL Dump
- Pseudonymization
- Data Cleansing
- Default NEA Interfaces
- Pilot Studies
 - Simplicity of Language
 - Importance of Examples



426842 - 1g Jack Flash (Hybrid)

Crossed it with Super Skunk and Haze. Jack Flash offers an earthy citrus aroma , impressive yields , and the active cerebral legacy of her Jack Herer parent. This hybrid gets its name from its lightning - fast onset which may prove useful to patients needing immediate symptom relief.

Labels

■ Drug

PROJECT EXECUTION - APPEN

- Mediocre results
- Entry tests
- Continuous Quality Monitoring
- Geo Restrictions + Top Annotators
- Cumbersome Quality Assurance
- Max. row limit without license = 1000

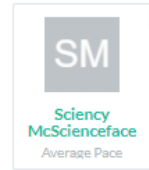
10 tablets 54 mg of ^{Drug} **Methylphenidate Hydrochloride** . Extended Release . Shipped
 in polyethylene bag suitable for food contact . Brand. Concerta Manufacturer
 Janssen .

Context




Ritalin 54 mg XR 10 tablets - Concerta -

PROJECT EXECUTION - AMAZON MTURK

- Good results
- „Master“-Annotators
- Reputation is important
 - Entry tests
 - Rejections/Blocks
- Self-Service



Drug Named Entity Annotation - \$0.11

 Fair	 Good	 Approved	\$13.20 / hour 00:00:30 / completion time
--	--	--	--

Pros

These HITs are interesting and the instructions are exceptionally detailed. The Requester has been exceptionally fair with me after a portion of the HITs that I submitted appeared to contain no data when he received them.

After initially rejecting the all HITs, the Requester promptly sent me a

[Show more](#)

Cons

EDIT: We've confirmed that the "MTurk Submit Orange Buttons with Hotkey (")" script somehow bypasses all completion checks and it erases all data upon HIT submission.

Therefore, DO NOT use this script to submit these HITs!

Advice to Requester

I wish there were better guidelines provided by Amazon for new Requesters running relatively small projects. Thank you for being receptive to feedback and best of luck with your Thesis project!

Mar 12, 2021 | 2 workers found this helpful.

[Helpful](#) 

<https://turkerview.com/requesters/ARC1S630YUZZE/reviews>

AGGREGATION AND QUALITY ASSURANCE

- Majority Voting (+Weights)
- IAA and data conversion scripts
- Review via Labelstudio

Task #1548

↶
↷
↻

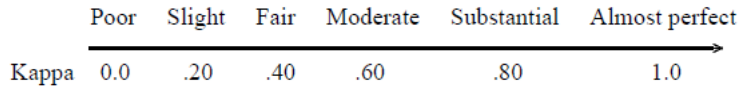
⊗ Skip
↻ Update

Drug^[1]
None^[2]

Actavis Promethazine Codeine Cough Syrup

Actavis Promethazine Codeine Cough Syrup DETAILS, Actavis promethazine with codeine purple cough syrup Tussionex cough syrup for sale Rexcoff codeine cough syrup for sale buy actavis online, actavis for sale actavis promethazine cough syrup online, actavis online, promethazine cough syrup online, actavis online, promethazine cough syrup with codeine online, buy actavis promethazine cough syrup online, buy actavis online. We sell top quality Actavis Promethazine with codeine purple cough syrup in large quantities at very affordable prices. Place your order and contact now. Actavis promethazine with codeine purple cough syrup, OXYCOTIN 80MG, XANAX 2 MG, morphine 30 MG, WATSON 853 10 MG, EPHEDRINE 30 MG, Roxycodone, Zyrtec, buy actavis online, Celebrex, Oxycodone, Tenormin, Testosterone, Tramadol, Trazodone, Ultram, Valium, Viagra, Vicodin, Zyban, Zylprim, Zyprexa, Zyrtec, Allegra, Alprazolam, Ambien, Amitriptyline, Aristocort, Atenolol, Ativan, Buspar, Carisoprodol, Celebrex, Cialis, Cipro, Claritin, Clonazepam, Codeine, Cozaar, Darvon, Deltasone, Desyrel, Diazepam, Effexor, Famvir, Glucophage, Hydrocodone, Klonopin, Lasix, Levitra, Lipitor, Lora Actavis promethazine online specifications. Each 5 mL contains Promethazine hydrochloride 625 mg codeine phosphate 10 mg Alcohol 7. Indication Cough Suppressant Dosage Form Syrup Validity 2yrs Strength 200mg Drug 1622-62-44 8 oz, 16 oz, 32 oz, 4oz 250 32oz / bottle 150 16oz / bottle Minimum order 5 bottles. Faster Communication. WICKR ID orlando50 KIK orlando5050.

ANNOTATION RESULTS



IAA	Appen	MTurk
Cohen's Kappa	0,43	0,76
Makro F1	0,60	77,55
Micro F1	0,55	0,79

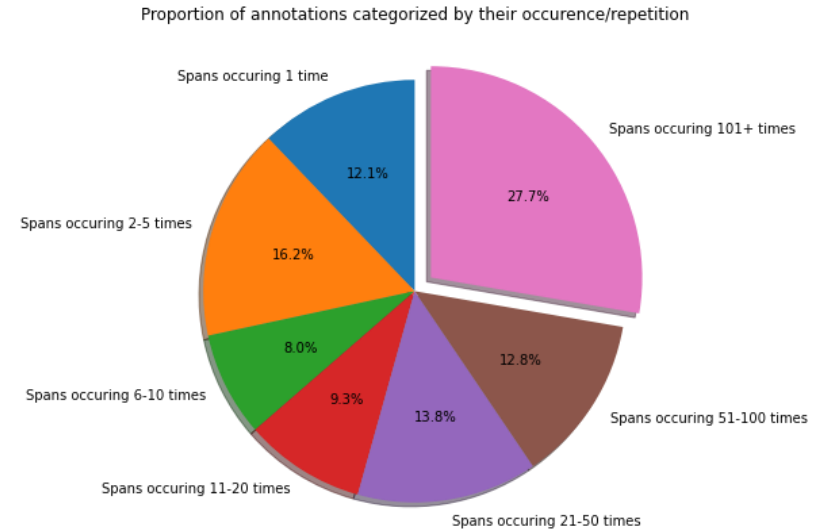
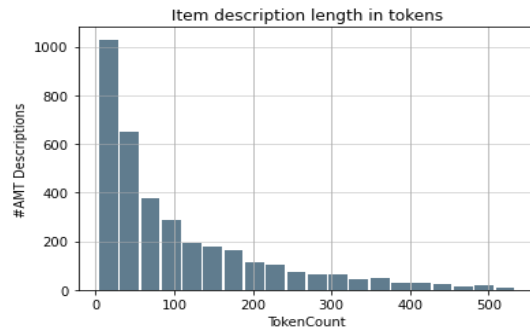
Final Annoation..	MTurk	Appen
% of chars added by reviewer	8,56	25,81
% of chars deleted	3,51	0,89
% of tags added/edited by reviewer	13,19	27,49
% of tags deleted/altered	8,65	10,45

CONCLUSIONS – CROWD SOURCING

- Annotation Guidelines! Annotation Guidelines! Annotation Guidelines!
- Be aware of entry barriers
- Volatile market
- F1-Agreement > Cohens Kappa
- More Annot. \neq higher agreement
- Custom solutions pay off

FINAL DATASET

- 3507 item listings with 364.003 words
 - Train 2244 / Dev 561 / Test 702
- Over 3k unique drugs and 15k total occurrences



MODELLING APPROACH

- Transformer architecture for word embeddings
- LM Fine-tuning on various text corpora
- Rigid Regularization due to overfitting
- Noisy source domains for NER task such as W-NUT 2017 or Broad Twitter Corpus
- Additional Few-shot approach for domain adaption



EXPERIMENTAL RESULTS – HYPER-PARAMETER TUNING

Model - FewShot	DAPT**	Dropout	F1-Score*	Precision	Recall
Best FLAIR model	None	0,0	59,37		
BERT Baseline	None	0,0	66,69	68,03	65,41
Best BERT model	ALL	0,5	71,37	73,48	69,38

Model - Full Trainingset	DAPT**	Dropout	F1-Score*	Precision	Recall
Best FLAIR model	None	0,0	72,84		
RoBERTa Baseline	None	0,0	80,49	83,26	77,89
Best RoBERTa model	ALL	0,5	82,79	83,94	81,67

*F1-Score according to CoNLL2003 NER evaluation metric

** DAPT- Domain Adaptive Pre-Training Text is the text corpora used for adapting the Language model to the Target domain

RESULTS – TASK ADAPTATION

FULL TRAINING DATASET

Pre-Training Dataset	LM	DAPT**	Drop.	F1-Score*	Precision	Recall
<u>NO Pre-Training</u>	<u>BERT</u>	<u>All</u>	<u>0.5</u>	<u>82,17</u>	<u>85,01</u>	<u>79,52</u>
CoNLL	BERT	All	0.5	79,98	85,89	74,83
BTC	BERT	All	0.5	81,76	85,47	78,36
W-NUT	BERT	All	0.5	80,48	84,35	76,95
<u>NuToT</u>	<u>BERT</u>	<u>All</u>	<u>0.5</u>	<u>82,58</u>	<u>86,76</u>	<u>78,79</u>
Wikipedia (distantly supervised)	BERT	All	0.5	78,17	79,89	76,52

RESULTS – TASK ADAPTATION FEWSHOT

Pre-Training Dataset	LM	DAPT**	Drop.	F1-Score*	Precision	Recall
<u>NO Pre-Training</u>	<u>BERT</u>	<u>All</u>	<u>0.5</u>	<u>71,37</u>	<u>73,48</u>	<u>69,38</u>
CoNLL	BERT	All	0.5	66,73	72,44	61,86
BTC	BERT	All	0.5	69,97	77,06	64,07
W-NUT	BERT	All	0.5	67,6	71,69	63,94
<u>NuToT</u>	<u>BERT</u>	<u>All</u>	<u>0.5</u>	<u>70,28</u>	<u>73,94</u>	<u>66,97</u>
Wikipedia (distantly supervised)	BERT	All	0.5	61,03	65,2	57,36

QUALITATIVE EVALUATION - STREAMLIT

Which Types of Errors would you like to explore?

FN

DRUG Detector

FP 1471 FN 710 TP 6322 TN 93491 PredTooEarly 2078 PredTooLate 216

Please Select Item to display from Error Category



Acetylfentanyl **B/B/I/I** 99 purity FE REQUIRED . Extremely potent . If you are not aware of this product you should refrain from it until you do your research it is many more times stronger than Heroin **B/O** you have been warned .

Acetylfentanyl **B-DRUG** 99 **I-DRUG** purity FE REQUIRED Extremely potent If you are not aware of this product you should refrain from it until you do your research it is many more times stronger than Heroin **B-DRUG** you have been warned

CONCLUSIONS – RESEARCH QUESTIONS

1. No significant improvement - Effective pre-training for NER task requires datasets with high quality and vocabulary / entity type overlap, rather than textual similarity.
2. Well structured pre-training corpora are a valuable resource for fine-tuning a Language Model for noisy user generated texts.
3. Distant supervision using DBPedia Spotlight API, did not achieve the necessary annotation quality in our setting
4. Our models outperformed the „off-the-shelf“ model by 10-12 points in terms of F1-Score. This performance improvement was worth the extra effort.

CONTRIBUTION

- Dataset for illicit drug recognition in darknet markets
- Pre-training text corpora for darknet markets
- Insight onto performance factors for pre-training on NER

THANK YOU!

Johannes Bogensperger 01427678

