NLP Course 2025. NER по сводкам Министерства Обороны РФ.

Андрей Павлов

May 2025

Аннотация

Ссылка на код проекта: https://github.com/as-pavlov/NLP_2025_RMDR.

1 Introduction

В сводках Минобороны РФ о ходе проведения СВО есть статистическая информация о потерях ВСУ, но в свободной форме. Сводки публикуются Минобороны России на официальном телеграм канале https://t.me/mod_russia. Цель проекта с помощью NER (Named Entity Recognition) получить подробные структурированные данные о потерях ВСУ по направлениям и населенным пунктам.

1.1 Team

Андрей Павлов подготовил этот документ.

2 Related Work

В проекте используются новые данные, и поэтому предыдущих работ по этой теме нет.

3 Model Description

Модель обучалась на базе предобученной модели Babelscape/wikineural-multilingual-ner взятой с Huggingface. [1]. Файл в котором проводилось обучение модели NLP_2025_RMDR_NERLM_2.ipynb. Во время обучения было разморожено 11 последних слоев. В качестве Loss используется CrossEntropyLoss. [DSouza,] [HFT,]

[1] [Babelscape,] https://huggingface.co/Babelscape/wikineural-multilingual-ner

4 Dataset

Данные для обучения были получены из официального телеграм-канала Министерства обороны РФ https://t.me/mod_russia путем стандартного экспорта в виде json. В проекте путь к файлу : ChatExport_2025-04-05\result.json

В файле NLP2025_RMDR_GetData.ipynb проводится предобработка данных, а именно: выделяется "чистый"текст сообщений, удаляются лишние символы. Фильтруются только сообщения о сводках. Подготовленные тексты сохраняются в атрибут clearText объектов message. Результат сохраняется в файл RMDR.json

Файл RMDR.json размечается в Label Studio с помощью меток.

Описание меток

```
<View style="display:flex;align-items:start;gap:8px;flex-direction:row">
    <Text name="text" value="$clearText" granularity="word"/>
    <Labels name="label" toName="text" showInline="false">
        <Label value="DIR" background="#4824f9"/>
        <Label value="SLD" background="#00ff1e"/>
        <Label value="WP" background="#ff0000"/>
        <Label value="LOC" background="#57fff4"/>
        <Label value="UNIT" background="green"/>
        <Label value="COUNT" background="#000000"/>
        <Label value="FREE" background="#0008ff"/>
        <Label value="LOST" background="#ff0000"/>
        <Label value="CAPT" background="#ffbb00"/>
  </Labels>
</View>
```

- DIR - Направление или группа войск
- SLD - Солдаты, военнослужащие
- WP - Оружие
- LOC - Населенные пункты
- UNIT - Штабы, склады оружия и пр.
- COUNT - количество (солдат или оружия)
- FREE - освобожденные населенные пункты (глаголы)
- LOST - утерянные населенные пункты (глаголы)
- CAPT - пленные (глаголы)

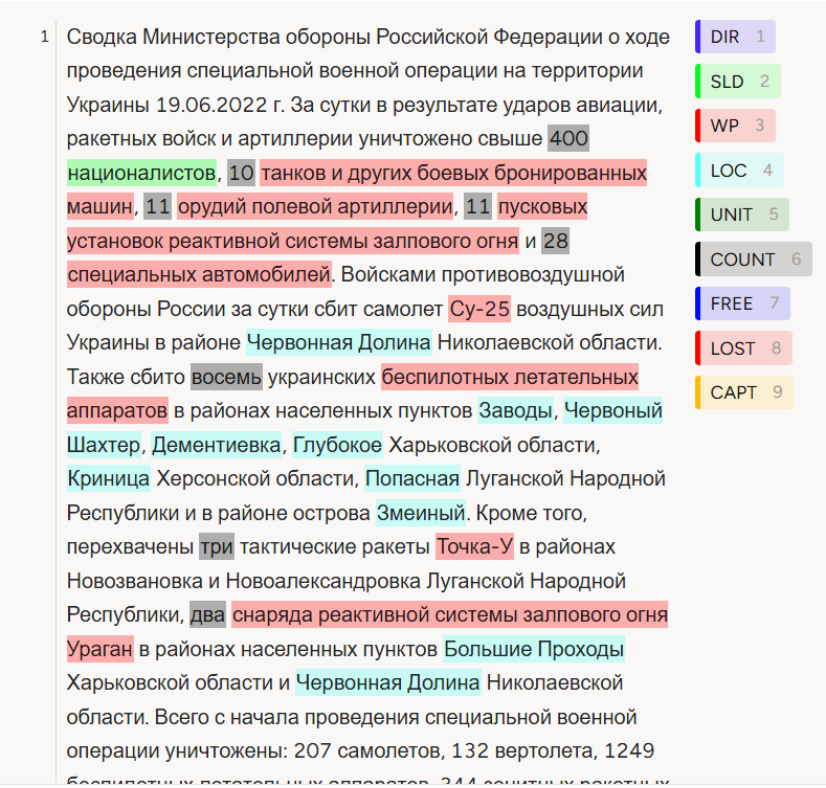

Рис. 1: Пример разметки в Label Studio

Результаты разметки в Label Studio храняться в файле RMDR_ANATATION_3_MONTH.json .[2]

В файле размечено только 3 месяца сводок: июнь-июль 2022, июнь 2023, июнь 2024. При обучении модели ровно эти же данные получаются непосредственно из Label Studio с попощью http запроса GetTasks. Перед обучением данные приводятся к виду:

```
"ids" : torch.tensor(ids, dtype=torch.long),
"mask" : torch.tensor(mask, dtype=torch.long),
"token_type_ids" : torch.tensor(token_type_ids, dtype=torch.long),
"target_tags" : torch.tensor(target_tags, dtype=torch.long)
```

- ids - Список id токенов, на которые разбит текст
- mask - везде 0

[2]Описание формата анотаций в Label Studio можно посмотреть по этой ссылке [Staples,] https://labelstud.io/blog/understanding-the-label-studio-json-format/

- token_type_ids - везде 1
- target_tags - id метки для каждого токена. Для токенов для которых нет метки - 0

5 Experiments

5.1 Metrics

В процессе обучения использовались метрики : precision, recall, accuracy.

5.2 Experiment Setup

В процессе обучения данные были разбиты в соотношении 0.9(train) к 0.1 (val). Так же существенно улучшила результат аугментация данных. Данные разбивались по порциям в 512 токенов (модель больше не может принять), но разбивались со смещением, то есть следующая порция начиналась не с 513 токена, а с 513 - SHIFT_SIZE (200). Так же опытным путем было подобрано количество размороженных слоев, достаточное для хорошего результата.

5.3 Baselines

Так как это новые данные (разметка), и использовалась только одна модель, которая показал, хорошие результаты, то других моделей нет и не с чем сравнивать.

6 Results

В результате обучения модели удалось добиться следующих значений метрик:

- train
  - precision = 0.6249239604801263
  - recall = 0.9428508267145929
  - accuracy = 0.914273631816007
- Val
  - precision = 0.6864593295870401
  - recall = 0.9605542609089559
  - accuracy = 0.8980305989583334

6.1 Пример результата действия модели

6.1.1 Текст

Министерства обороны Российской Федерации о ходе проведения специальной военной операции по состоянию на 5 мая 2025 г. Вооруженные Силы Российской Федерации продолжают проведение специальной военной операции. Подразделениями группировки войск Север нанесено поражение скоплениям живой силы и техники механизированной, танковой, егерской бригад ВСУ и двух бригад теробороны в районах населенных пунктов Садки, Рясное, Великая Писаревка Сумской области и Гранов Харьковской области. Потери ВСУ составили до 150 военнослужащих, три танка, две боевые бронированные машины и шесть автомобилей. Уничтожен склад боеприпасов. Подразделения группировки войск Запад заняли более выгодные рубежи и позиции. Нанесли поражение формированиям двух механизированных, горно-штурмовой, штурмовой бригад ВСУ и бригады теробороны в районах населенных пунктов Пески, Купянск, Григоровка, Кутьковка Харьковской области и Карповка Донецкой Народной Республики. Противник потерял свыше 225 военнослужащих, боевую бронированную машину, шесть автомобилей и два артиллерийских орудия западного производства. Уничтожены три склада боеприпасов. Подразделения Южной группировки войск улучшили тактическое положение. Нанесли поражение живой силе и технике двух механизированных, аэромобильной бригад ВСУ и бригады теробороны в районах населенных пунктов Серебрянка, Дружковка, Северск и Заря Донецкой Народной Республики. Потери украинских вооруженных формирований составили свыше 315 военнослужащих, две боевые бронированные машины и восемь орудий полевой артиллерии. Подразделения группировки войск Центр улучшили положение по переднему краю. Нанесли поражение формированиям двух механизированных, егерской бригад ВСУ, бригады спецназначения и двух бригад нацгвардии в районах населенных пунктов Удачное, Димитров, Новопавловка, Новосергеевка и Гродовка Донецкой Народной Республики. Противник потерял до 465 военнослужащих, танк, четыре боевые бронированные машины, шесть автомобилей и четыре артиллерийских орудия.Подразделения группировки войск Восток продолжили продвижение в глубину обороны противника. Нанесли поражение живой силе и технике двух механизированных бригад ВСУ, бригады морской пехоты и бригады теробороны в районах населенных пунктов Богатырь, Федоровка, Комар и Новополь Донецкой Народной Республики. Потери противника составили до 170 военнослужащих, боевая бронированная машина, пять автомобилей и четыре орудия полевой артиллерии. Уничтожены две станции радиоэлектронной борьбы. Подразделения группировки войск Днепр нанесли поражение формированиям механизированной бригады и двух бригад береговой обороны ВСУ в райо-

нах населенных пунктов Малая Токмачка, Павловка Запорожской области и Антоновка Херсонской области. Противник потерял до 70 военнослужащих, пять автомобилей, пусковую установку реактивной системы залпового огня HIMARS производства США и два орудия полевой артиллерии. Уничтожены две станции радиоэлектронной борьбы и склад боеприпасов. Оперативно-тактической авиацией, ударными беспилотными летательными аппаратами, ракетными войсками и артиллерией группировок войск Вооруженных Сил Российской Федерации поражены объекты инфраструктуры военного аэродрома, ремонтного предприятия авиационной промышленности, склады боеприпасов, а также пункты временной дислокации украинских вооруженных формирований и иностранных наемников в 142 районах. Средствами противовоздушной обороны уничтожены четыре управляемые авиационные бомбы JDAM и четыре реактивных снаряда системы залпового огня HIMARS производства США, а также 123 беспилотных летательных аппарата самолетного типа. Всего с начала проведения специальной военной операции уничтожены: 662 самолета, 283 вертолета, 55 452 беспилотных летательных аппарата, 605 зенитных ракетных комплексов, 23 153 танка и других боевых бронированных машин, 1 558 боевых машин реактивных систем залпового огня, 24 435 орудий полевой артиллерии и минометов, 35 005 единиц специальной военной автомобильной техники.

6.1.2 Результат модели

- Направление: Север
  - Населенные пункты: Садки, Рясное, Великая Писаревка, Гранов
  - военнослужащих - 150
  - танка - три
  - боевые бронированные машины - две
  - автомобилей - шесть
  - склад боеприпасов
- Направление: Запад
  - Населенные пункты: Пески, Купянск, Григоровка, Кутьковка, Карповка
  - военнослужащих - 225
  - боевую бронированную машину
  - автомобилей - шесть
  - артиллерийских орудия западного производства - два
  - склада боеприпасов - три

- Направление: Южной
  - Населенные пункты: Серебрянка, Дружковка, Северск, Заря
  - военнослужащих - 315
  - боевые бронированные машины - две
  - орудий полевой артиллерии - восемь
- Направление: Центр
  - Населенные пункты: Удачное, Димитров, Новопавловка, Новосергеевка, Гродовка
  - военнослужащих - 465
  - боевые бронированные машины - четыре
  - автомобилей - шесть
  - артиллерийских орудия - четыре
  - Населенные пункты: Богатырь, Федоровка, Комар, Новополь
  - военнослужащих - 170
  - боевая бронированная машина
  - автомобилей - пять
  - орудия полевой артиллерии - четыре
  - станции радиоэлектронной борьбы - две
- Направление: Днепр
  - Населенные пункты: Малая Токмачка, Павловка, Антоновка
  - военнослужащих - 70
  - автомобилей - пять
  - пусковую установку реактивной системы залпового огня HIMARS
  - орудия полевой артиллерии - два
  - станции радиоэлектронной борьбы - две
  - склад боеприпасов, объекты инфраструктуры военного аэродрома, склады боеприпасов, пункты временной дислокации районах - 142
  - управляемые авиационные бомбы JDAM - четыре
  - реактивных снаряда системы залпового огня HIMARS - четыре
  - беспилотных летательных аппарата самолетного типа, - 123

Так же в рамках проекта был реализован web service который возвращает результат выполнения модели в формате Label Studio (json). Этот web service можно подключить к Label Studio и использовать его для дальнейшей разметки данных [niklub and nik,]. Подробнее описано в README.md резозитория.

7 Conclusion

В задачи проекта входило получение структурированных данных о потерях ВСУ из сводок Минобороны РФ. Были размечены данные для модели и была обучена модель на этих данных, которая показала хорошие результаты. Также был написан web service, помогающий в дальнейшей разметке данных.

Список литературы

[HFT,] Token classification.

[Babelscape,] Babelscape. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner.

[DSouza,] DSouza, D. L. Ner using bert - pytorch.

[niklub and nik,] niklub and nik. label-studio-ml-backend.

[Staples,] Staples, E. M. Understanding the label studio json format.