Event Detection

Michal Starý

...in multilingual setting on uncommon domain

Context

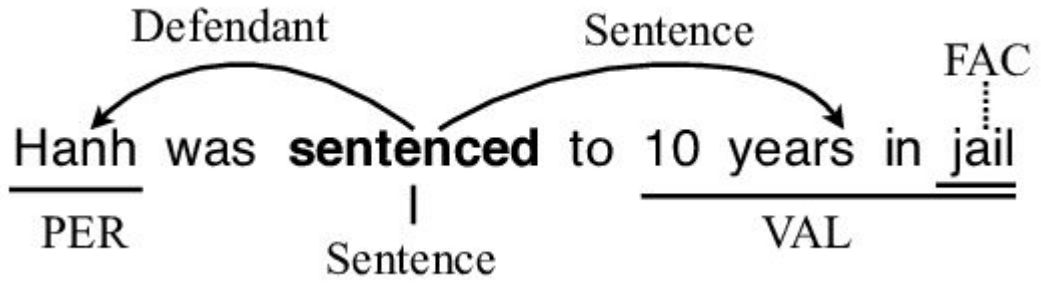
...that is necessary

Context

NLP

- Information Extraction
  - Event Extraction
    - **Event Detection**
    - Event Argument Extraction
    - …
  - ...

Example from ACE05



Event Miner

Temporal Expression recognition [DONE]

Temporal Expression normalization [DONE]

Event Detection

Relation Extraction

Event Miner

Multilingual

(not only) Office domain

Finetunable

Modular

Event detection

Usually simpler than event argument extraction

Kinds

- Close domain
- Open domain

Main domains

News

Medical data

Event definition

Not clear at all

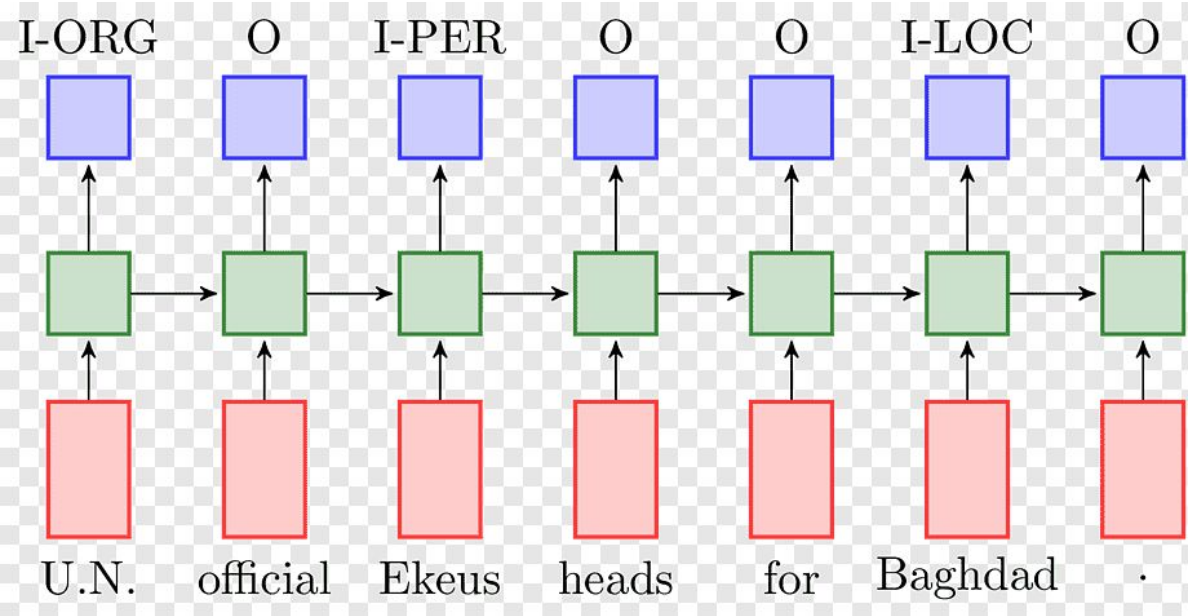
Domain dependent

Approaches

...that was proposed

Sequence tagging (token classification) task

Technically, ED is an instance of sequence tagging (=token classification task)



Classification to BILOU scheme

B - 'beginning'

I - 'inside'

L - 'last'

O - 'outside'

U - 'unit'

Give 50$ to John tomorrow.

-> (Give, B), (50, I), ($, L), (to, O), (John, O), (tommorow, O), (., O)

Published approaches for EE (ML/DL focus)

- traditional ML on top of all kinds of features (POS, NER, depency tree)
- CNN
- LSTM/GRU
- Tree-LSTM utilizing the dependency tree
- Graph NN (useful for event arguments)
- Graph-CNN
- CNN-LSTM
- GAN
- Attention over edges of SDT
- BERT, ELMO, …
  - but not that much dominant as usual nowadays
  - probably due to the “mutually dependent” nature of argument extraction

Data

...that are relevant

Datasets (generic news oriented)

Free

- Language Understanding Annotation Corpus (2009)
- ECB+ (2014)
- MAVEN (2020)
- RAMS (2020)
- TBAQ + TE3
- MEANTIME (5lang)

Licensed (LDC)

- **ACE05 (2005)**
- TAC-KBP 2016
- TDT CORPORA
- Richer Event Description (2016)

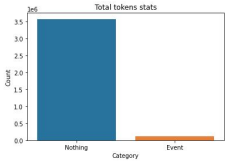
Proprietary (KM) - [OFFICE domain]

- Task dataset

Available datasets detail

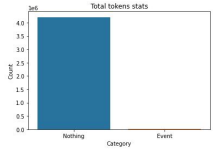
MAVEN (2020)

- 4,480 documents - wikipedia
- 118,732 event mention instances
- 168 event types.
- *4M tokens*



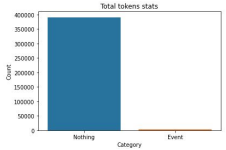
RAMS (2020)

- 8,000 documents - news
- 9,124 annotated events from news
- based on an ontology of 139 event types and 65 roles.
- *4M tokens*



ECB+ (2014)

- 502 documents - 43 topics of the ECB
- “action annotation”
- 4,006 events
- *400k tokens*



Language Understanding Annotation (2009)

- "informal input,".
- excerpts from newswire stories, telephone conversation transcripts, emails, contracts and written instructions.
- *12k tokens*

Available datasets detail

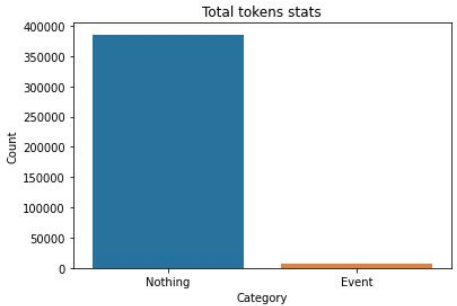
Task

769 documents - emails

7,856 events

“task annotation”

400k tokens



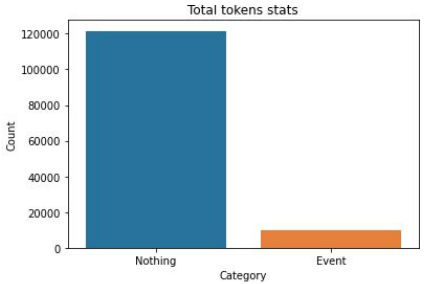
TBAQ

257 documents - news

10,007 events

“verb annotation”

120k tokens



Target data - office

Email

Hi Andrea et al,

I hope you are keeping well. Did you have any question from the documentation?

Let us know what the next steps would be, as we are very excited about this project.

Many Thanks,

Kind Regards

Vasilis

Contract

T ORIGINS, ENTER ORIGINS AGAINST ITEMS IN 12.

9. Conditions of Sale and Terms of Payment (ie. Sales, Consignment Shipment, Leased Goods, etc.).

INLAND CLEARANCE

10A. Currency of Settlement

BORDER CLEARANCE

11. No. of HM 12. Specification of Commodities & HS Tariff Classification Code (Kind of Packages, Marks and Numbers, Pkgs. General Description and Characteristics ie. Grade, Quality)

12A Weight

18. If any of fields 1 to 17 are included on an attached commercial invoice, check this box

Commercial Invoice No. 19. Exporter's Name, Address and Phone# (if other than Vendor)

23. If included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of direct shipment to Canada. (ii) Costs for construction, erection and assembly incurred after importation into Canada. (iii) Export Packing.
24. Shipper's No. or B.O.L. No.
25. If not included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of dir

Model zoo

...that support zero-shot multilinguality

Dealing with multilingual inputs

- language recognition + language specific model
- translation into english + english model (+ translation back)
- ...
- **multilingual model**

Multilingual transfomer based models

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

ProphetNet

https://arxiv.org/pdf/2010.11934.pdf

multiBERT

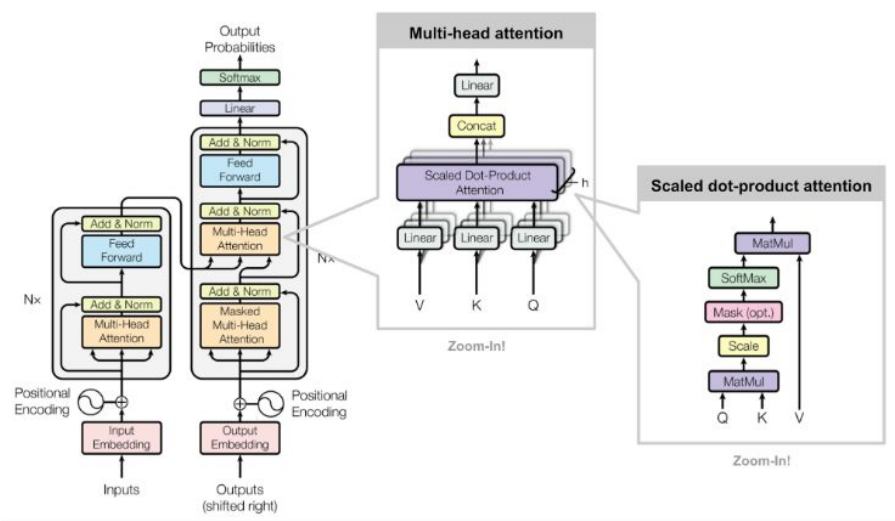
BERT base cased

12-layer

768-hidden

12-heads

110M parameters



multilingual corpus (wikipedia) - 104 languages

MLM + NSP

On NER: ~60 F1 zero-shot, ~90 F1 finetuned

XLM (croX lingual Language Model)

2 versions

- with specific language embedding
- fully multilingual

multilingual corpus (wikipedia) - 104 languages

- a causal language modeling (CLM) objective (next token prediction),
- a masked language modeling (MLM) objective (BERT-like), or
- a Translation Language Modeling (TLM) object (extension of BERT's MLM to multiple language inputs)

On NER: ~60 F1 zero-shot

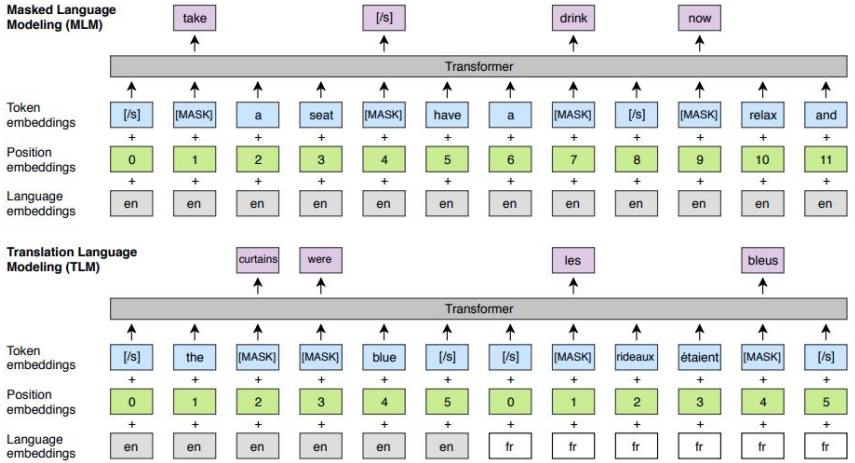


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

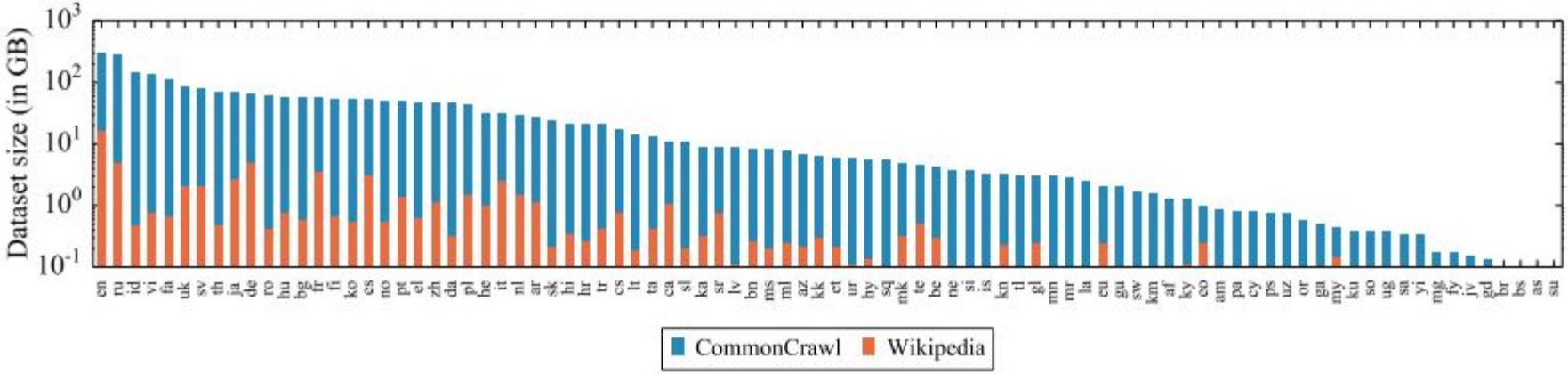
XLM-RoBERTa

based on RoBERTa

- It builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

trained on 100 languages - common crawl

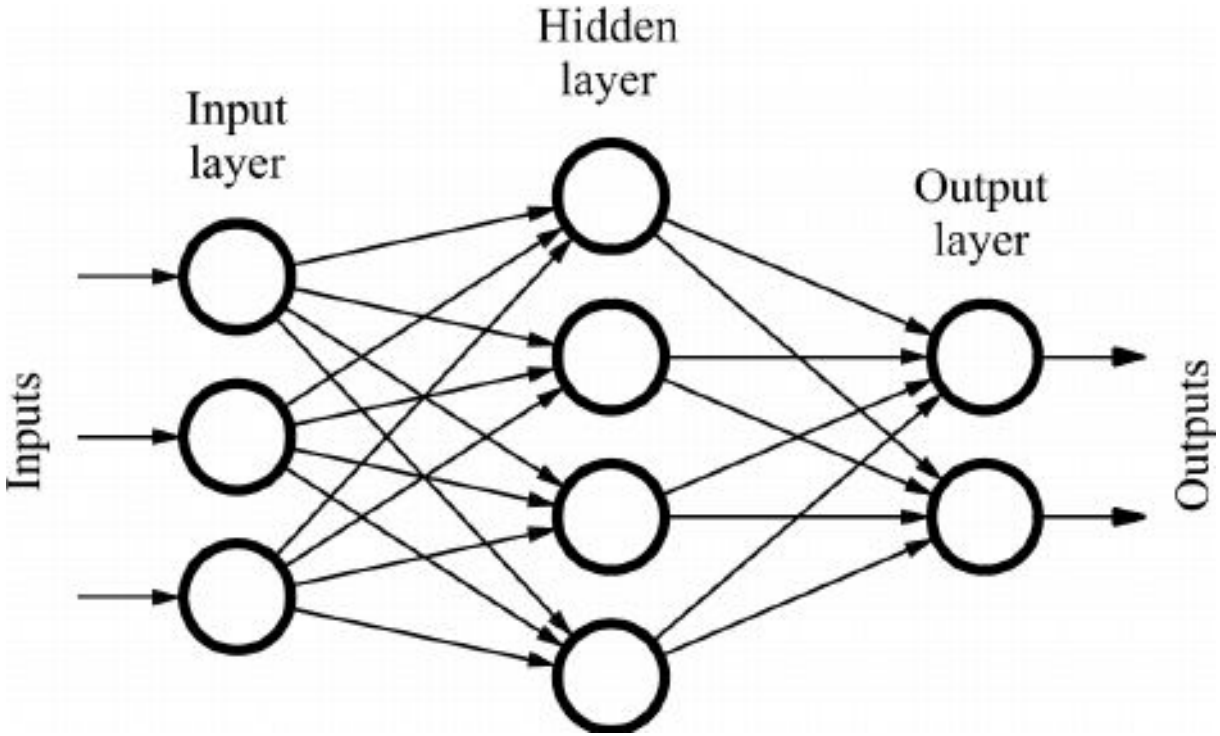
On NER: ~65 F1 zero-shot



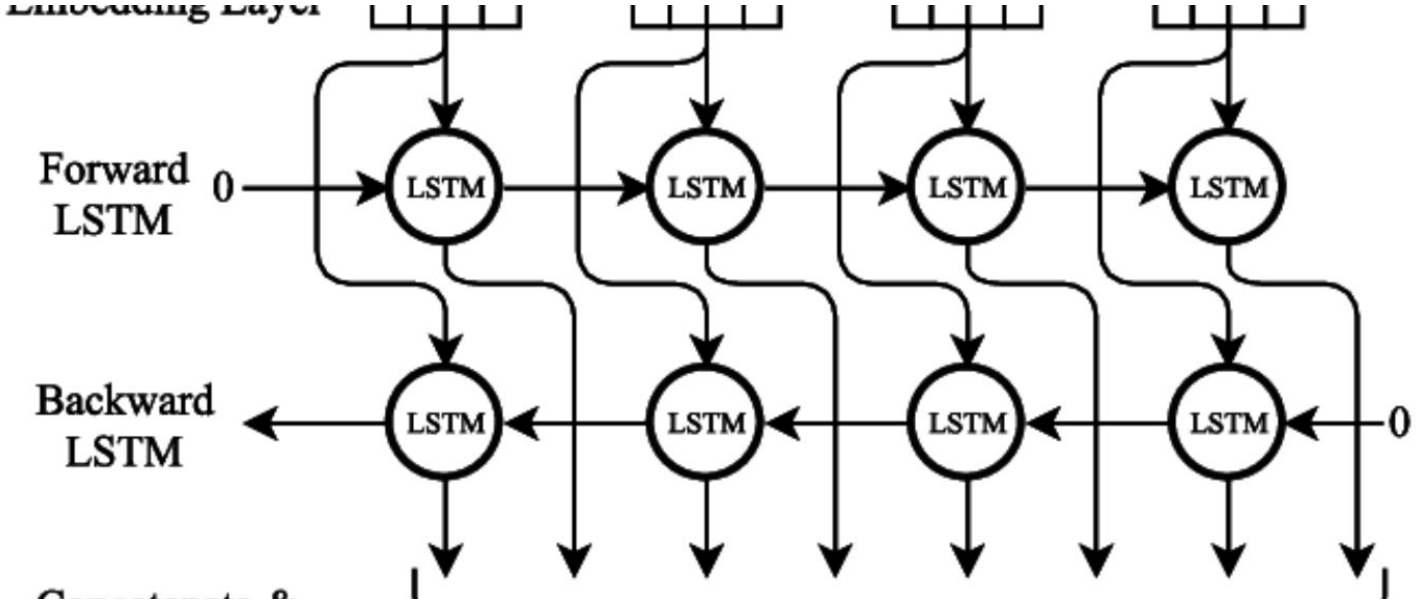
Adding some more layers

...from contextualized embeddings to class

Linear



biLSTM


Forward
LSTM
0
LSTM
LSTM
LSTM
LSTM
Backward
LSTM
LSTM
LSTM
LSTM
LSTM
0

Leveraging the inter-labels dependencies

PROBLEM

tokens = tokenize("**Náměstí 28.října**")

tokens.words

-> ["Náměstí", "28", ".", "října"]

model(tokens)

-> [B-Loc, B-Date, O, L-Date]

Náměstí 28.října

DESIRED SOLUTION

tags = model(tokens)

tags

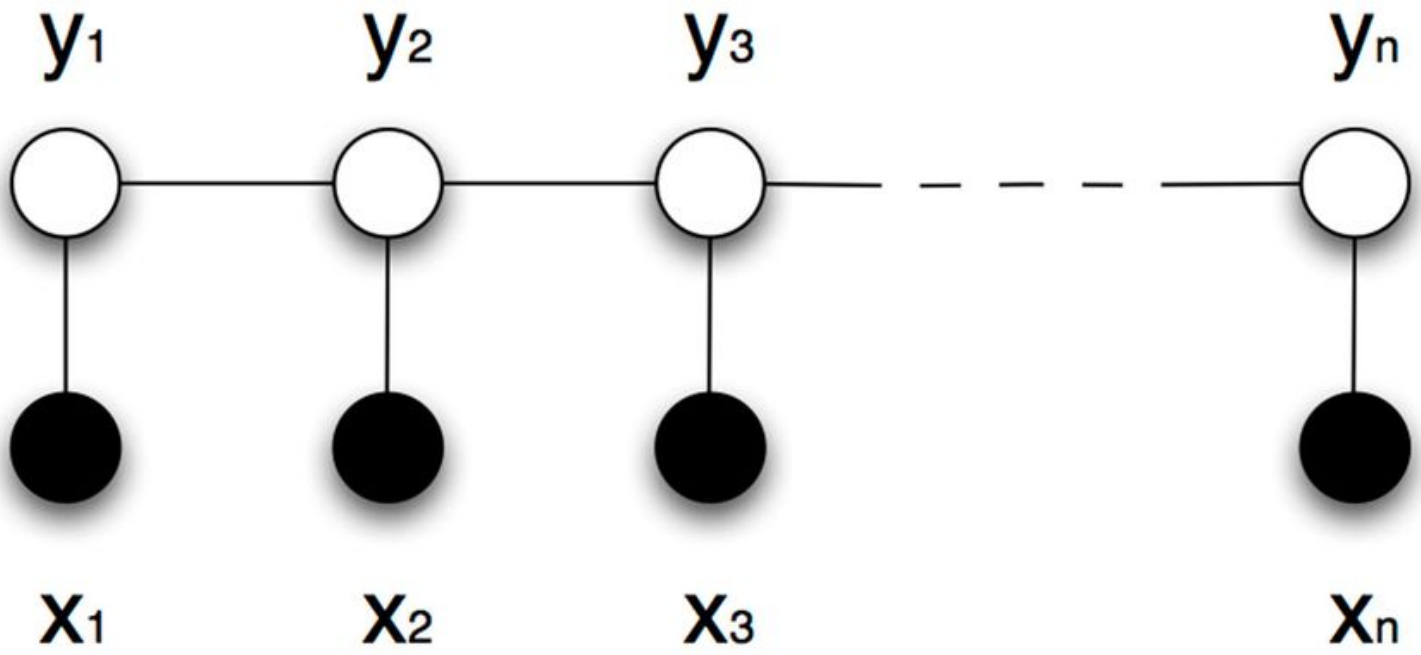
-> [B-Loc, B-Date, O, L-Date]

correct_tags(tags)

-> [B-Loc, I-Loc,I-Loc, L-Loc]

Náměstí 28.října

CRF



Our models (part1)

multiBERT + Linear

Implemented, Trained

(BertForTokenClassification from Hugging Face)

multiBERT + biLSTM + Linear

Implemented, Not trained yet

(BertModel from Hugging Face, nn.LSTM and nn.linear from Pytorch)

multiBERT + Linear + CRF

TODO

multiBERT + biLSTM + Linear + CRF

TODO

Our models (part2)

XLM-RoBERTa + Linear

Implemented, Trained

(XLMRobertaForTokenClassification from Hugging Face)

XLM-RoBERTa + biLSTM + Linear

Implemented, Not trained yet

(XLMRobertaModel from Hugging Face, nn.LSTM and nn.linear from Pytorch)

XLM-RoBERTa + Linear + CRF

TODO

XLM-RoBERTa + biLSTM + Linear + CRF

TODO

Training

...that raises more questions than it answers

Stats and settings

5 datasets

~20k documents

~7,3M tokens in train set

~120k event tokens in train set

Individual training + eval for each dataset

Joint training + eval on all of them

Finetuned on turnus03 - 1x 2080 GPU

Adam optimizer

batch_size = 4

lr = 1e-5

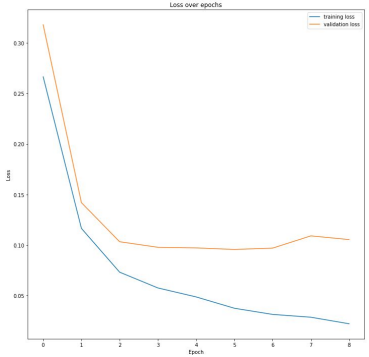
wdecay=1.2e-6

Joint dataset takes ~15mins/epoch

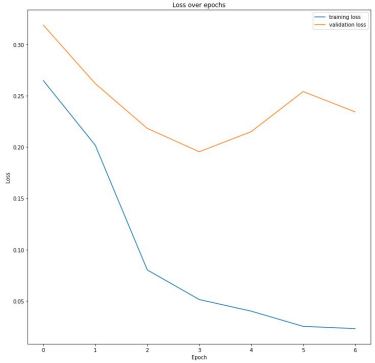
Very few (1-5 epochs) until convergence

Training progresses (multiBERT+Linear)

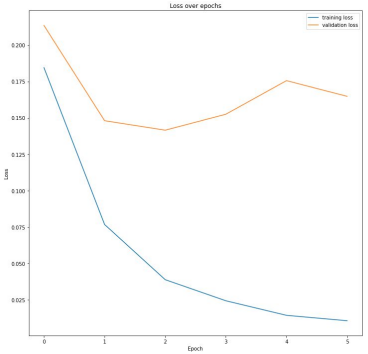
TBAQ



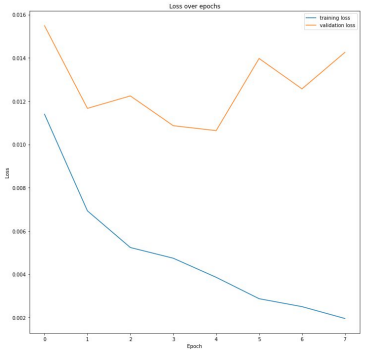
Task



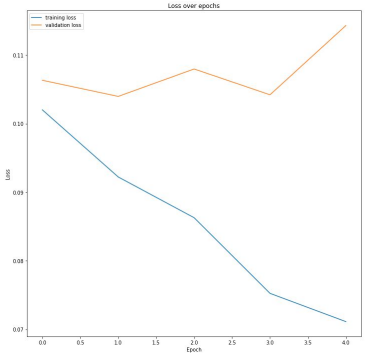
ECB+



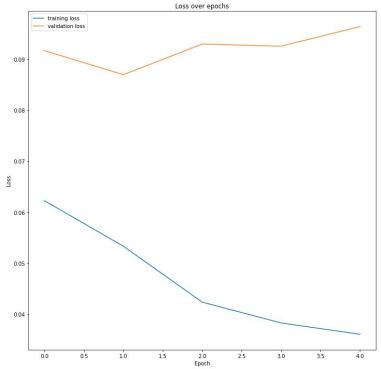
RAMS



MAVEN

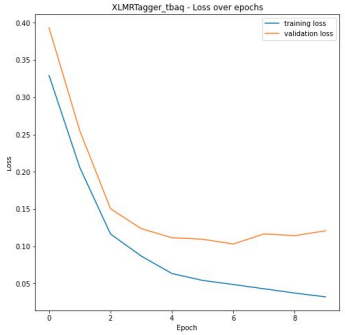


Joint

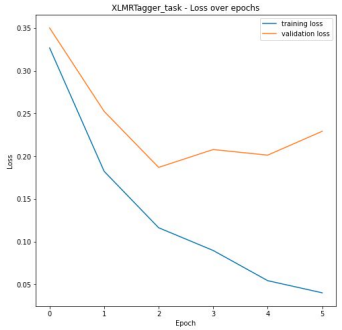


Training progresses (XLM-Roberta+Linear)

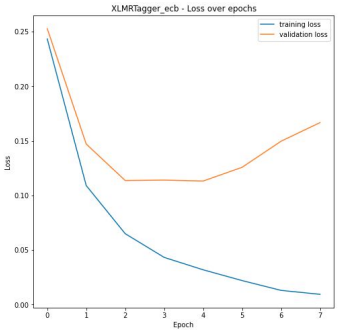
TBAQ



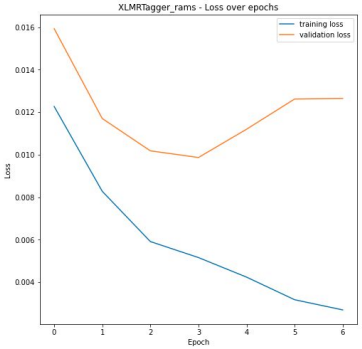
Task



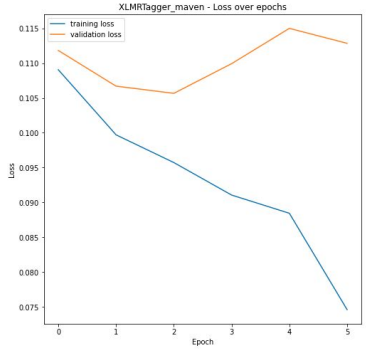
ECB+



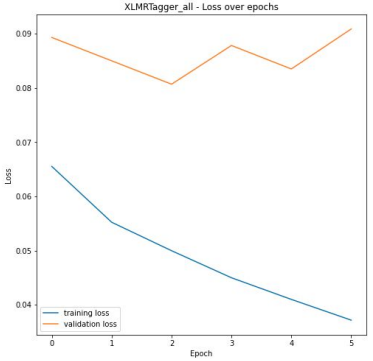
RAMS



MAVEN



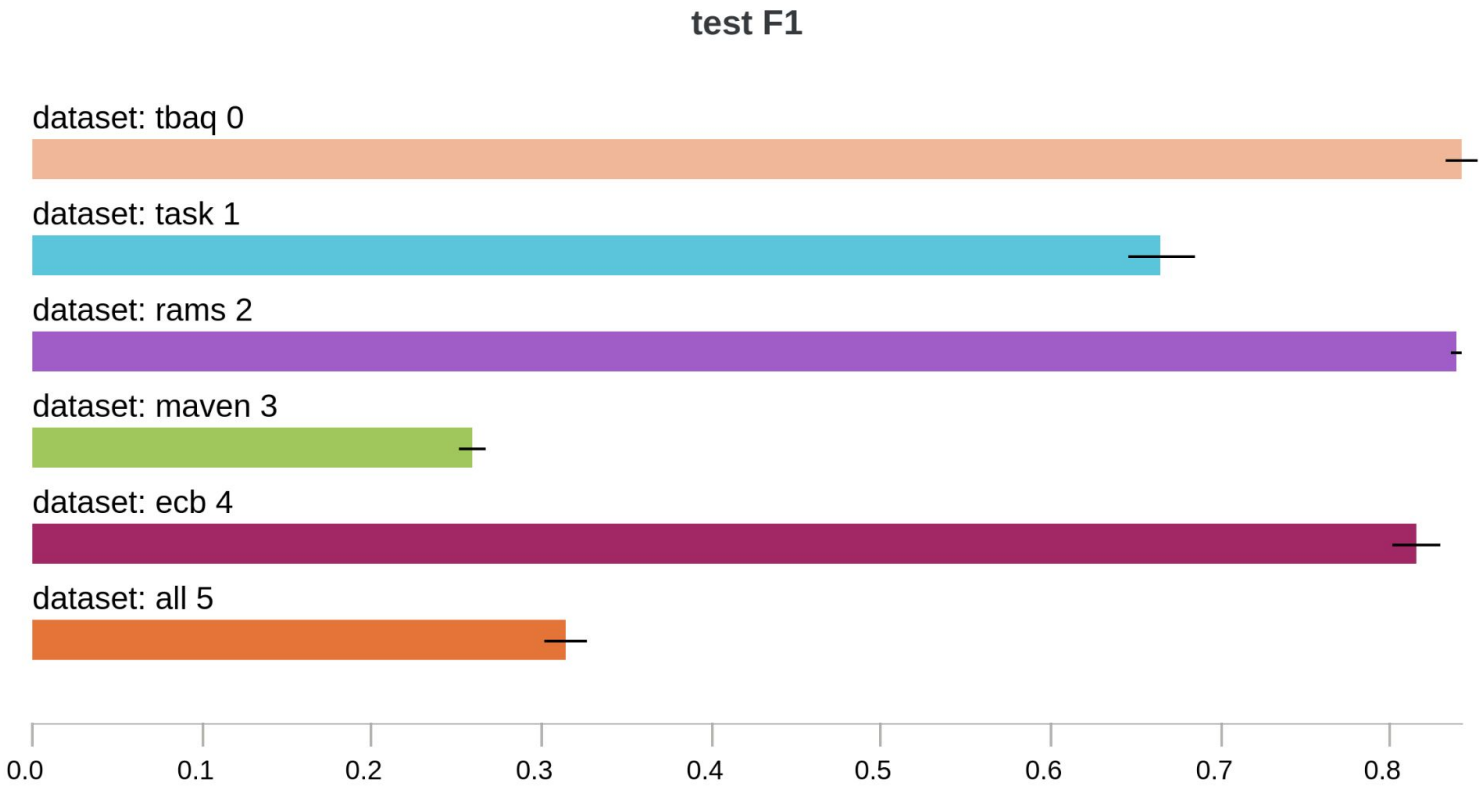
Joint



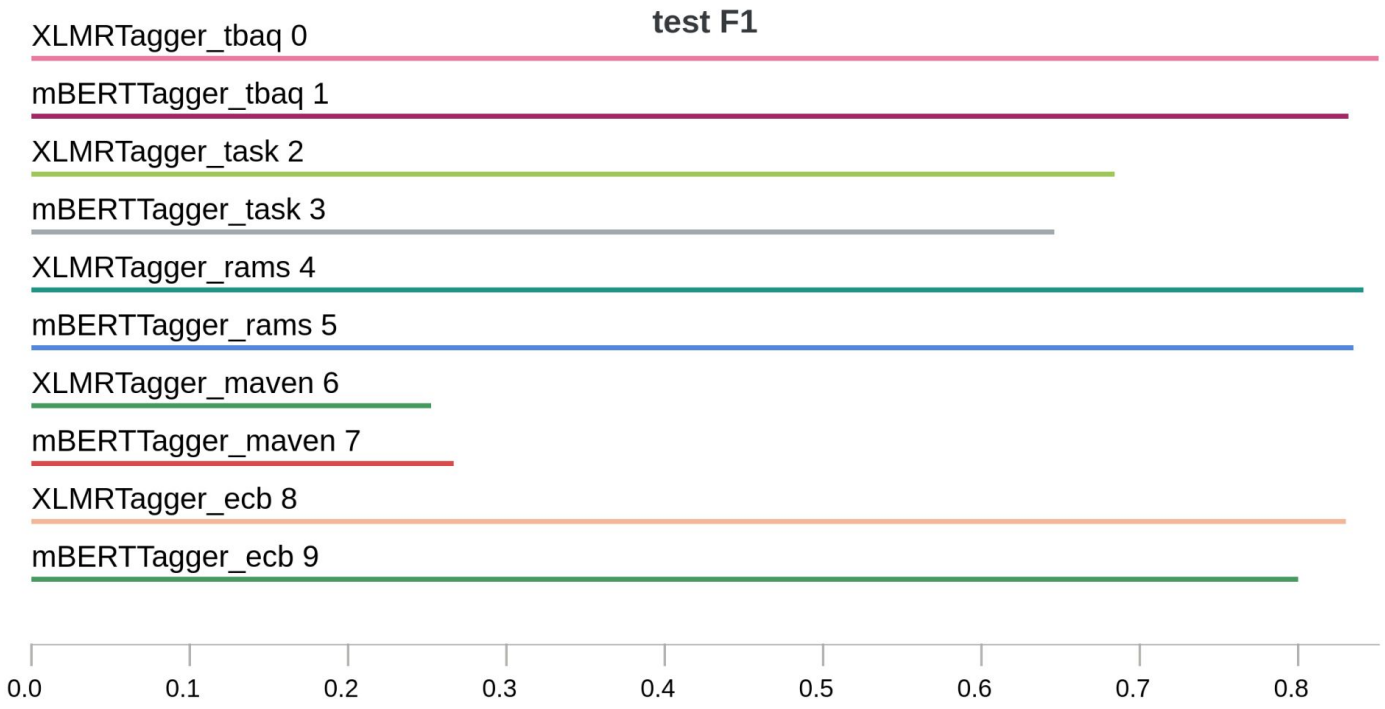
(partial) Evaluation

...that raises more questions than it answers

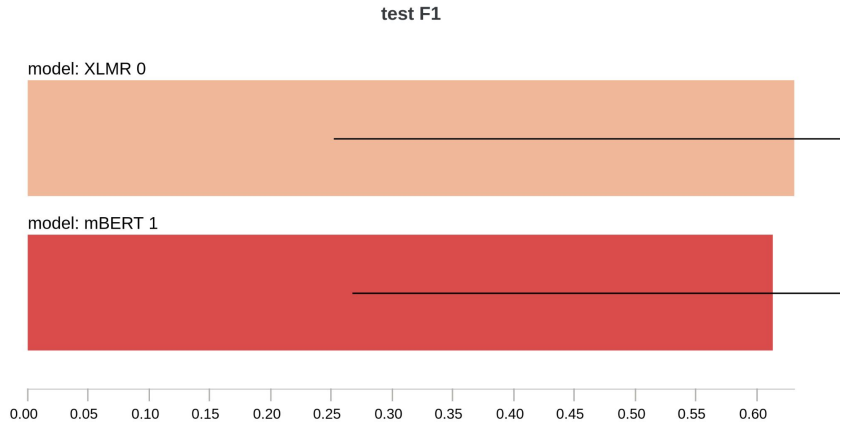
Datasets

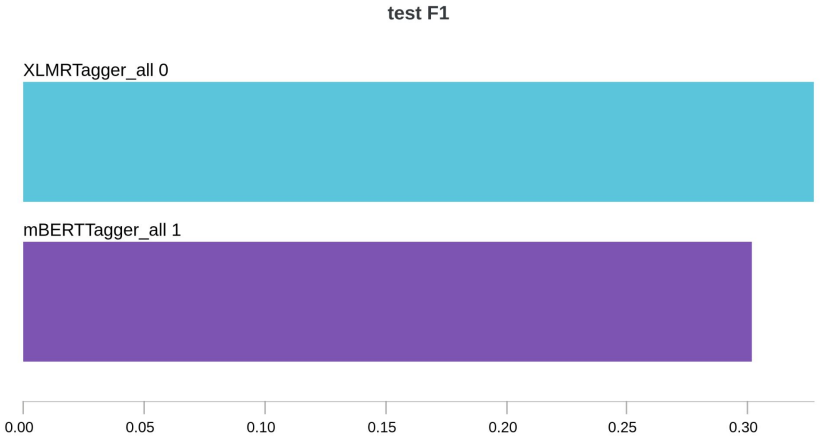

test F1
dataset: tbaq 0
dataset: task 1
dataset: rams 2
dataset: maven 3
dataset: ecb 4
dataset: all 5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Quantitative results - individual datasets

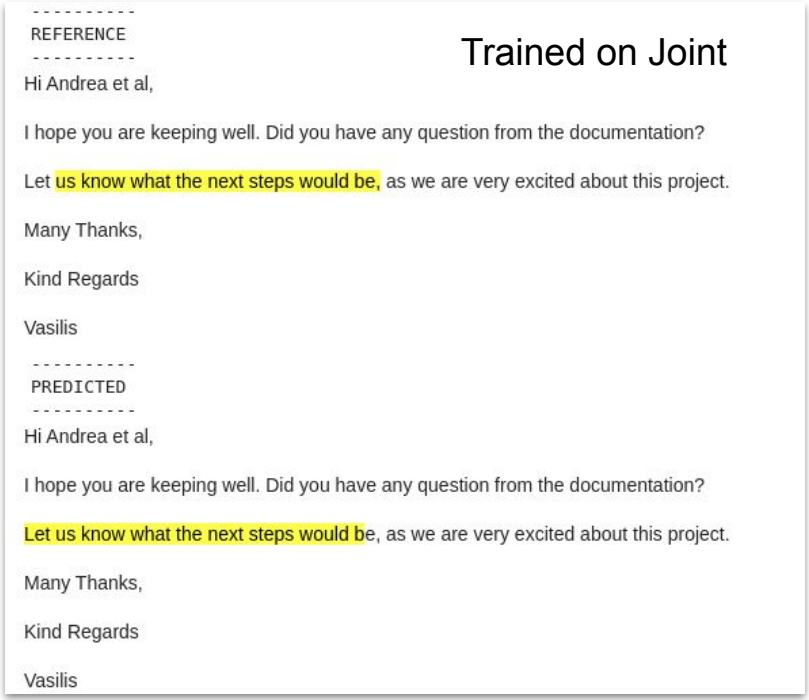


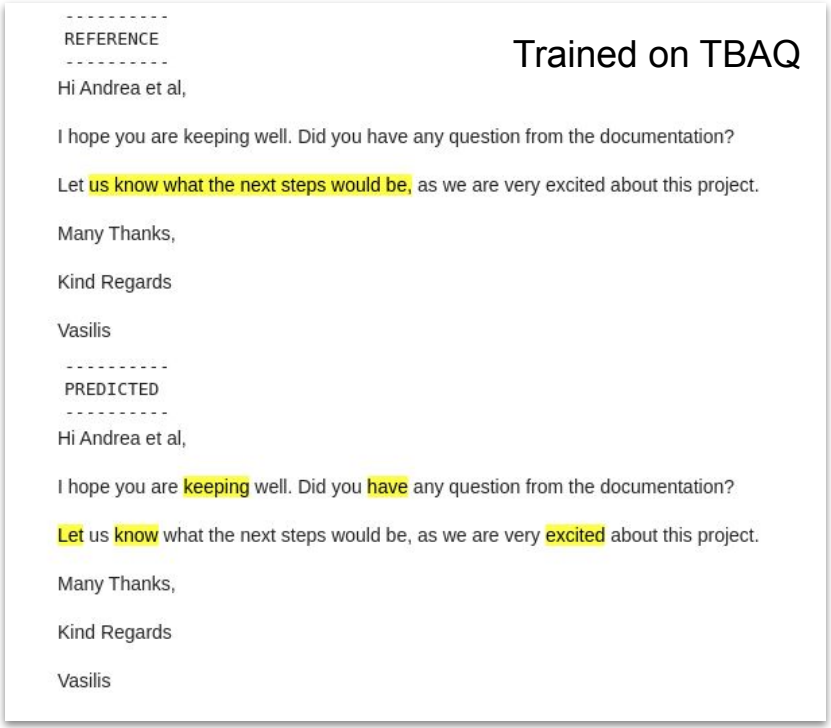
Quantitative results - joint dataset





Qualitative results





Qualitative OOD results

T ORIGINS, ENTER ORIGINS AGAINST ITEMS IN 12.

9. Conditions of Sale and Terms of Payment (ie. Sales, Consignment Shipment, Leased Goods, etc.)

10.

INLAND CLEARANCE

10A. Currency of Settlement

BORDER CLEARANCE

11. No. of HM 12. Specification of Commodities & HS Tariff Classification Code (Kind of Packages, Marks and Numbers, Pkgs. General Description and Characteristics ie. Grade, Quality)

12A Weight

18. If any of fields 1 to 17 are included on an attached commercial invoice, check this box

Commercial Invoice No.
19. Exporter's Name, Address and Phone# (if other than Vendor)

20. If included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of direct shipment to Canada. (ii) Costs for construction, erection and assembly incurred after importation into Canada. (iii) Export Packing.

21. Shipper's No. or B.O.L. No.

22. If not included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of dir

Trained on Joint

S IN 12.

ie. Sales, Consignment Shipment, Leased Goods, etc.)

Trained on TBAQ

BORDER CLEARANCE

11. No. of HM 12. Specification of Commodities & HS Tariff Classification Code (Kind of Packages, Marks and Numbers, Pkgs. General Description and Characteristics ie. Grade, Quality)

12A Weight

18. If any of fields 1 to 17 are included on an attached commercial invoice, check this box

Commercial Invoice No. 19. Exporter's Name, Address and Phone# (if other than Vendor)

23. If included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of direct shipment to Canada. (ii) Costs for construction, erection and assembly incurred after importation into Canada. (iii) Export Packing.

24. Shipper's No. or B.O.L. No.

25. If not included in field 17 indicate amount: (i) Transportation charges, expenses and insurance to the place of dir

Qualitative zero-shot cross lingual results

Trained on Joint

Kate, please, can you confirm the meeting at 6 PM tomorrow? It's about time to prepare our talk at the NLP conference, it's on the 27th of November, remember? We need to submit the presentation 2 weeks before the conference itself. PS. I've talked to John on Wed. and the vacation won't be a problem. See you, Michal

Kate, prosím, můžete potvrdit schůzku zítra v 18:00? Je čas připravit naši přednášku na konferenci NLP, je to 27. listopadu, pamatujete? Prezentaci musíme předložit 2 týdny před samotnou konferencí. PS. Mluvil jsem s Johnem ve středu a dovolená nebude problém. Uvidíme se, Michal

Kate, Bitte, können Sie das Treffen morgen um 18 Uhr bestätigen? Es ist an der Zeit, unseren Vortrag auf der NLP-Konferenz vorzubereiten. Es ist am 27. November, erinnerst du dich? Wir müssen die Präsentation 2 Wochen vor der Konferenz selbst einreichen. PS. Ich habe am Mittwoch mit John gesprochen. und der Urlaub wird kein Problem sein. Wir sehen uns, Michal

Kate, s'il vous plaît, pouvez-vous confirmer la réunion à 18 heures demain? Il est temps de préparer notre discours à la conférence de la PNL, c'est le 27. novembre, tu te souviens? Nous devons soumettre la présentation 2 semaines avant la conférence elle-même.

Trained on TBAQ

Kate, please, can you confirm the meeting at 6 PM tomorrow? It's about time to prepare our talk at the NLP conference, it's on the 27th of November, remember? We need to submit the presentation 2 weeks before the conference itself. PS. I've talked to John on Wed. and the vacation won't be a problem. See you, Michal

Kate, prosím, můžete potvrdit schůzku zítra v 18:00? Je čas připravit naši přednášku na konferenci NLP, je to 27. listopadu, pamatujete? Prezentaci musíme předložit 2 týdny před samotnou konferencí. PS. Mluvil jsem s Johnem ve středu a dovolená nebude problém. Uvidíme se, Michal

Kate, Bitte, können Sie das Treffen morgen um 18 Uhr bestätigen? Es ist an der Zeit, unseren Vortrag auf der NLP-Konferenz vorzubereiten. Es ist am 27. November, erinnerst du dich? Wir müssen die Präsentation 2 Wochen vor der Konferenz selbst einreichen. PS. Ich habe am Mittwoch mit John gesprochen. und der Urlaub wird kein Problem sein. Wir sehen uns, Michal

Kate, s'il vous plaît, pouvez-vous confirmer la réunion à 18 heures demain? Il est temps de préparer notre discours à la conférence de la PNL, c'est le 27. novembre, tu te souviens? Nous devons soumettre la présentation 2 semaines avant la conférence elle-même.

Final notes

Lot of work to do

Any ideas/ tips welcomed!

Whole pipeline allows multiclass sequence tagging (NER, POS, ..)

Can be possibly used to compare the performance of multilingual vs language dedicated (csalbert, csgpt2) pretrained models on such kinds of tasks