

*“We have also obtained a glimpse  
of another crucial idea  
about languages and program design.*

*This is the thing  
which I greatly feared  
is come upon us,  
desolation and destruction.”*

# New AI fake text generator may be too dangerous to release, say creators

**The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse**



Source: <https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction>

# From Research to Production

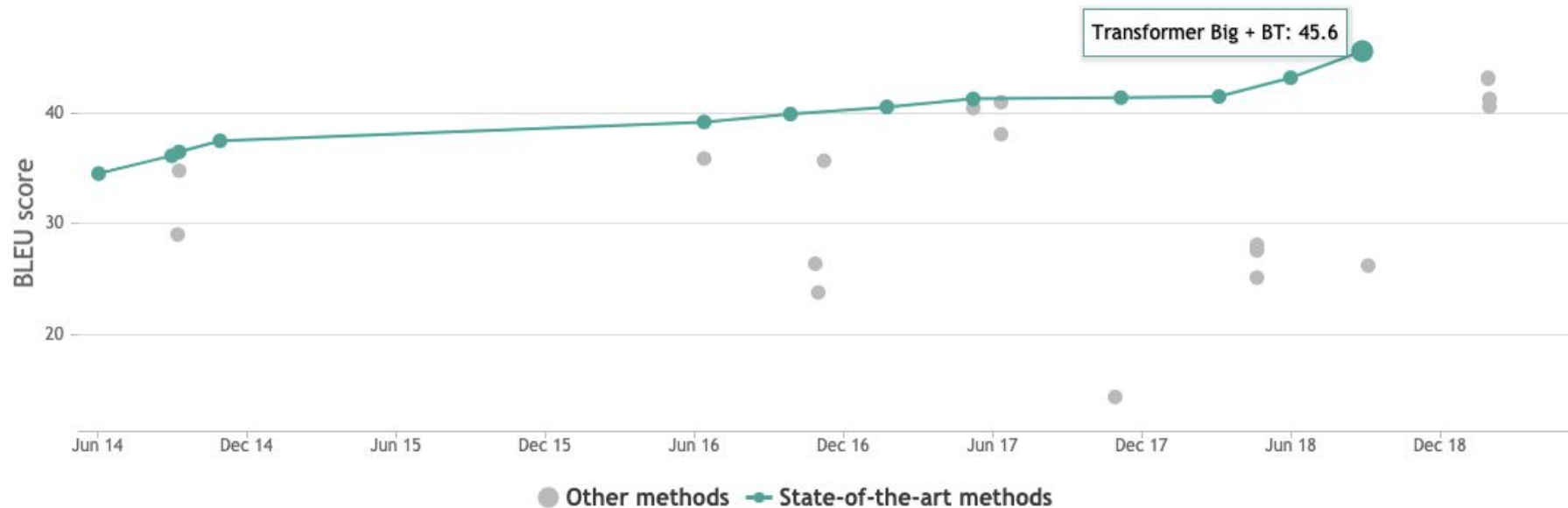
Matti Lyra

Research Engineer

@mattilyra / @comtravo\_tech

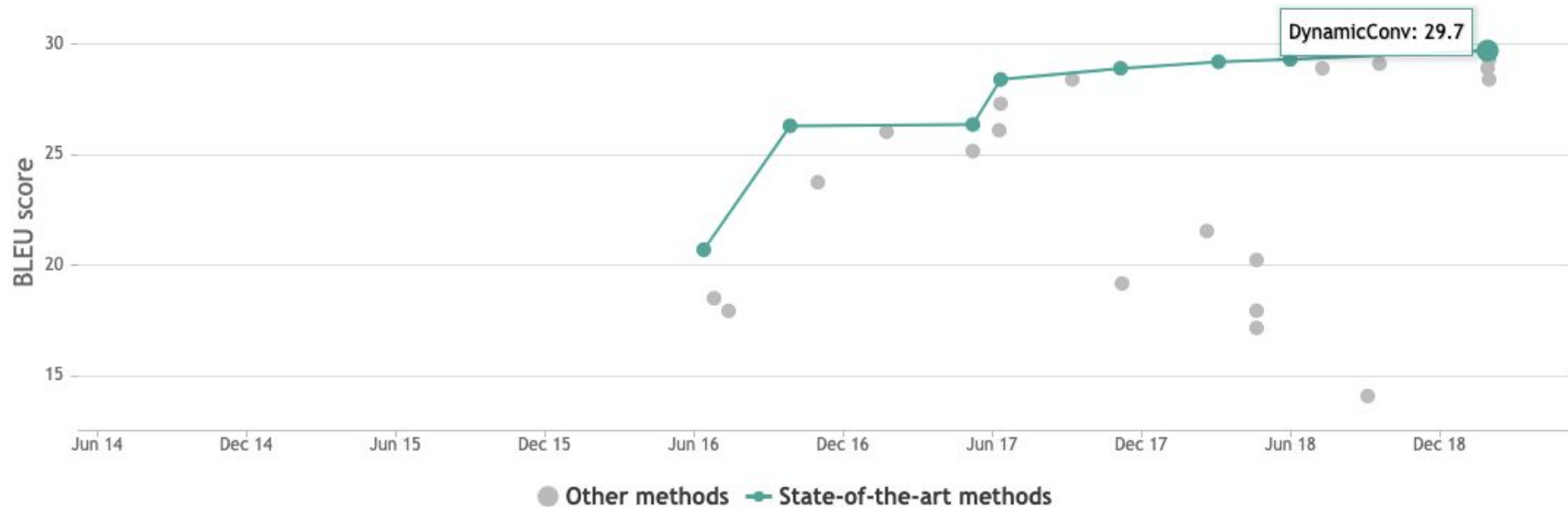


# Machine Translation (WMT14 EN-FR)



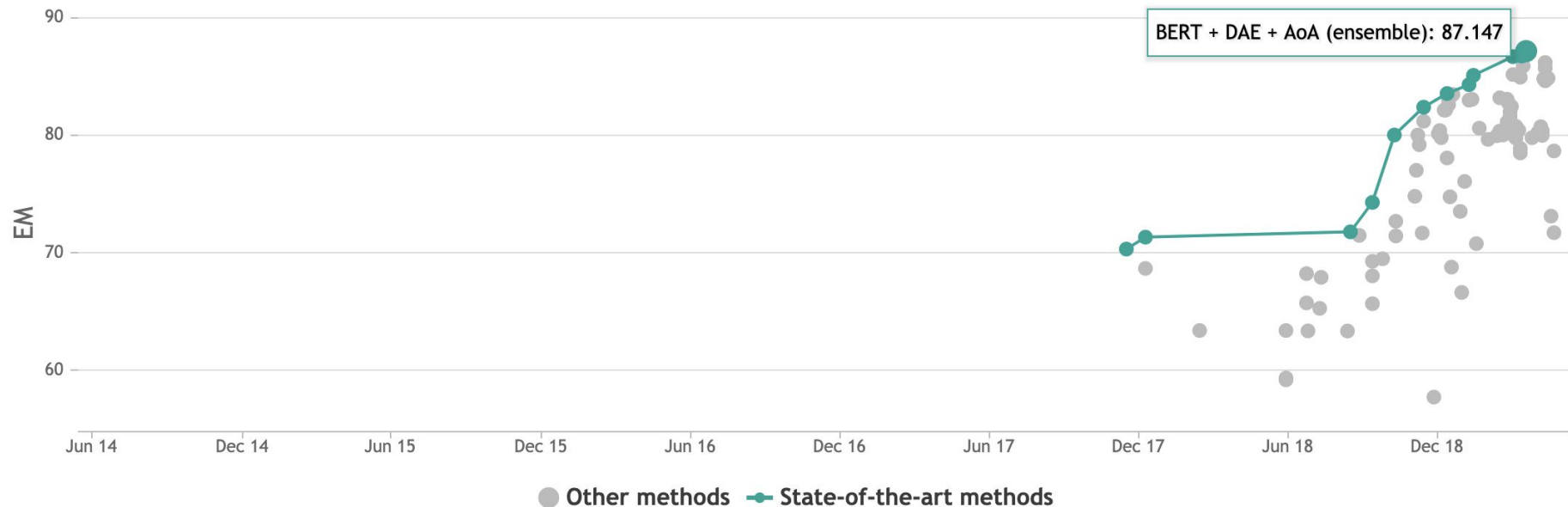
Source: <https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-french>

# Machine Translation (WMT14 EN-DE)



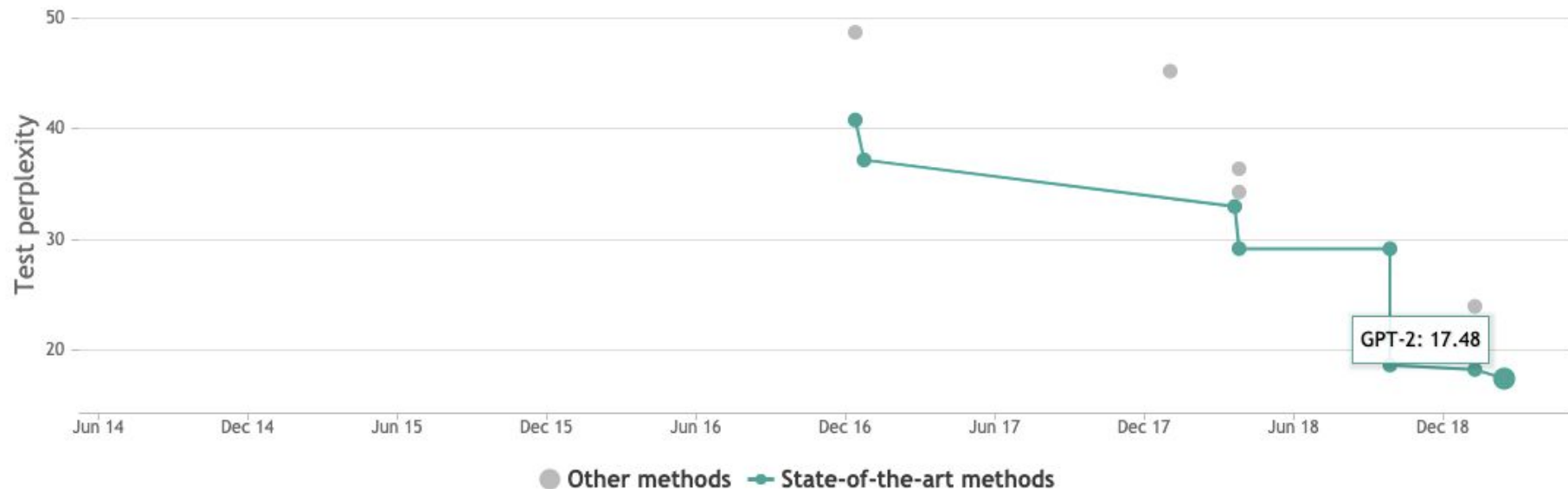
Source: <https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german>

# Question Answering (SQuAD2.0)



Source: <https://paperswithcode.com/sota/question-answering-on-squad20>

# Language Modeling (EN-DE)



Source: <https://paperswithcode.com/sota/language-modelling-on-wikitext-103>

# Predict the Next Element in a Sequence

I bet you can't guess what I'm about to say

...	
to	5%
<b>next</b>	<b>12%</b>
after	4%
tomorrow	1%
apple	0.04%
!	0.00025%
...	

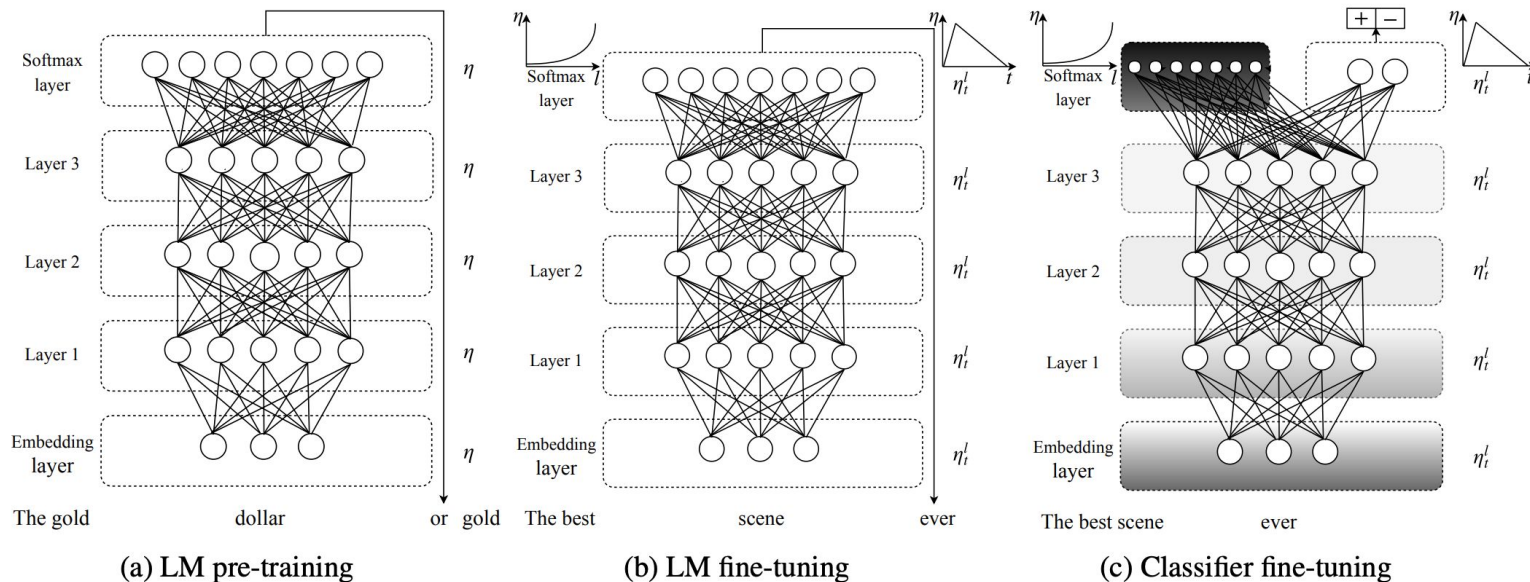
*"The train up to Hamburg today was really slow, but I had time to enjoy the scenery"*

*"The train up to Hamburg today was really slow, but it was also uncomfortable"*

"Das geht einfach nicht so! Doch, es geht!"



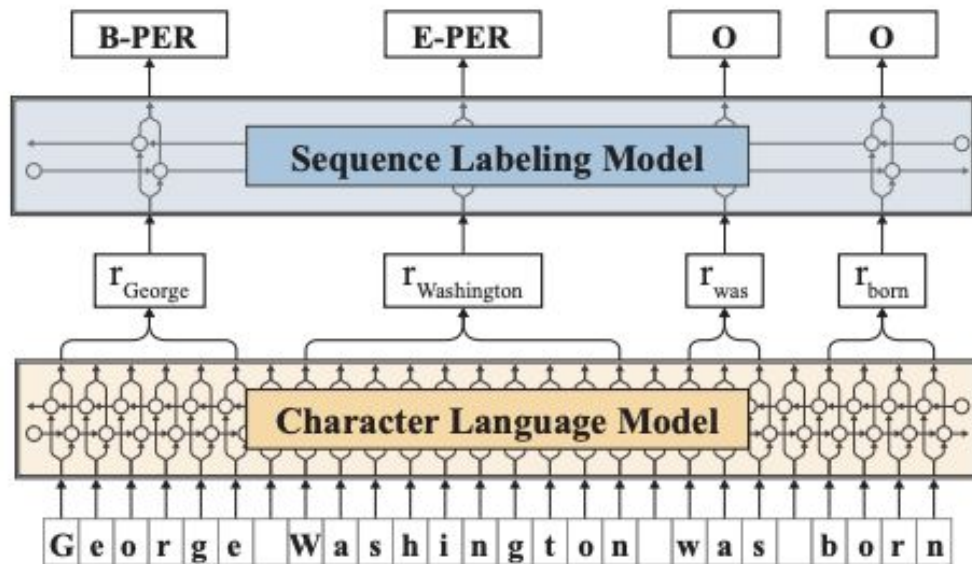
# Language Model Pretraining



"Universal Language Model Fine-tuning for Text Classification" (Howard et al. 2018) ACL'18

<https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html>

# Language Model Pretraining



*"Contextual String Embeddings for Sequence Labeling"* (Akbik et al. 2018) COLING'18

*"Pooled Contextualized Embeddings for Named Entity Recognition"* (Akbik et al. 2019) NAACL'19

<https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html>

Hallo Comtravo.

Buchungsanfrage für ein Hotelzimmer incl.  
Frühstück in Eisenach vom 08.05.2019 bis  
10.05.2019.

Bevorzugtes Hotel wäre hierfür das Hotel  
„pentahotel Eisenach“ oder das Steigenberger  
Hotel Thüringer Hof


Reisender: Jari Litmanen

Bitte buchen Sie auf die Kostenstelle 73934.

Vielen Dank.

\*\*\*\*\*

*Mit freundlichen Grüßen / Best regards  
Matti Lyra  
Technischer Außendienst Division PKW/  
Technical Sales Support Division LV*



Dear Mr Litmanen,

please find below your options:

### Hotel in Eisenach

#1

Steigenberger Hotel Thüringer Hof

**Steigenberger Hotel Thüringer Hof \*\*\*\***  
Karlsplatz 11 Eisenach, DE

CHECKIN	CHECKOUT
<b>May 8th</b> Wednesday	<b>May 10th</b> Friday
incl. breakfast / incl. wifi / non cancelable / Checkin from 3pm	

PAYMENT DETAILS

1 Twin/Double Room

2 nights X **€ 99.30**

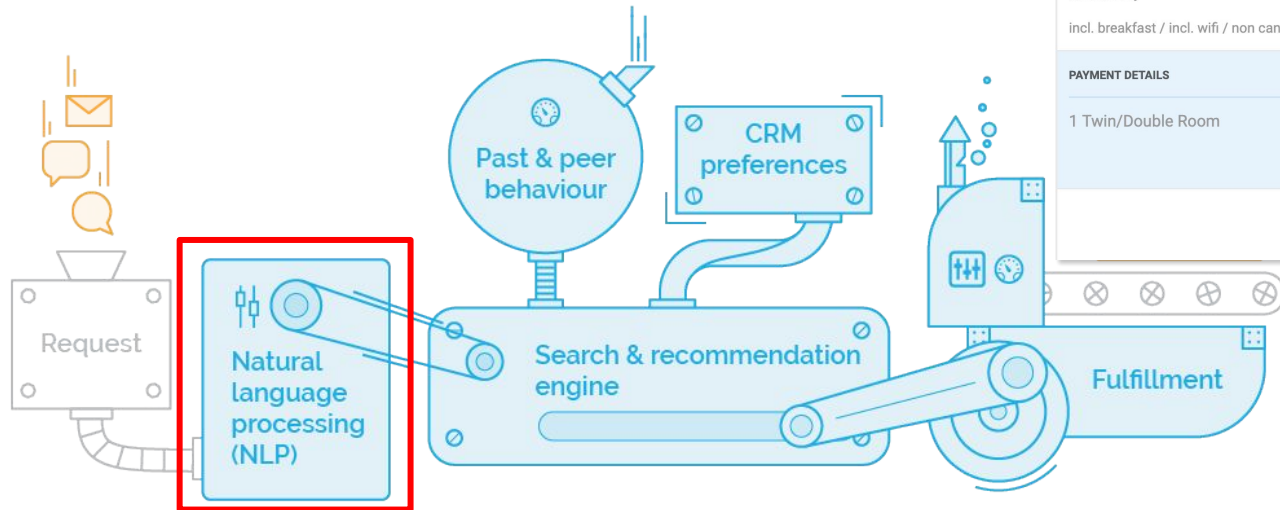
**Total (gross): € 198.60**


CONFIRM THIS OPTION→

Hallo Comtravo.

Buchungsanfrage für ein Hotelzimmer incl.  
Frühstück in Eisenach vom 08.05.2019 bis  
10.05.2019.

Bevorzugtes Hotel wäre hierfür das Hotel  
„pentahotel Eisenach“ oder das Steigenberger  
Hotel Thüringer Hof ...





Dear Mr Litmanen,

please find below your options:

### Hotel in Eisenach

#1 **Steigenberger Hotel Thüringer Hof**

**Steigenberger Hotel Thüringer Hof \*\*\*\***  
[Karlsplatz 11 Eisenach, DE](#)

CHECKIN

**May 8th**  
Wednesday

CHECKOUT

**May 10th**  
Friday

incl. breakfast / incl. wifi / non cancelable / Checkin from 3pm

PAYMENT DETAILS

1 Twin/Double Room

2 nights X € 99.30

**Total (gross): € 198.60**

CONFIRM THIS OPTION→

# Formalised Request

	Hotel Booking	
H	Hotel location	Eisenach
H	Hotel Name	Pentahotel Eisenach
H	Hotel check in	8.5.2019
H	Hotel check out	10.5.2019
H	Hotel number of rooms	1
V	Wifi	Yes
P	Parking	No
E	Breakfast	Yes

# Machine Translation

*"Hi Comtravos,*

*I need a morning flight Berlin - Helsinki early on the 10th or an afternoon / evening flight after 6pm the day before back on the 12th and a hotel close to Finlandia Hall, cheapest.*

*Best, Matti"*

---

*<flight.dep **early morning 10th**> <OR> <flight.dep **after 6pm 9th**>*

*<flight.orig **Berlin**> <flight.dest **Helsinki**>*

*<flight.orig **Helsinki**> <flight.dest **Berlin**> <flight.dep **12th**>*

# Question Answering

*"Hi Comtravos,*

*I need a morning flight Berlin - Helsinki early on the 10th or an afternoon / evening flight after 6pm the day before back on the 12th and a hotel close to Finlandia Hall, cheapest.*

*Best, Matti"*

- 
- Fly from Berlin to Helsinki
  - I want to fly early morning on the 10th
  - I want to fly back on the 12th
  
  - Q: What is the origin of the flight?
  - Q: What is the destination of the flight?

*"Dynamic Memory Networks for Visual and Textual Question Answering"* (Xiong et al. 2016) ICML'16

<https://www.youtube.com/watch?v=T3octNTE7Is>

# Sequence Prediction (a.k.a. NER)

*"Hi Contravos,*

*I need a morning flight Berlin - Helsinki early on the 10th or an afternoon / evening flight after 6pm the day before back on the 12th and a hotel close to Finlandia Hall, cheapest.*

*Best, Matti"*

---

<O> <O> <O> <U-flight.dep> <O> <U-flight.orig> <U-flight.dest> <B-flight.dep> <I-flight.dep>  
I need a **morning** flight **Berlin** **Helsinki** **early** **on**

<I-flight.dep> <I-flight.dep> <O> <O> <U-flight.dep>  
**the** **10th** or an **afternoon / evening** ...

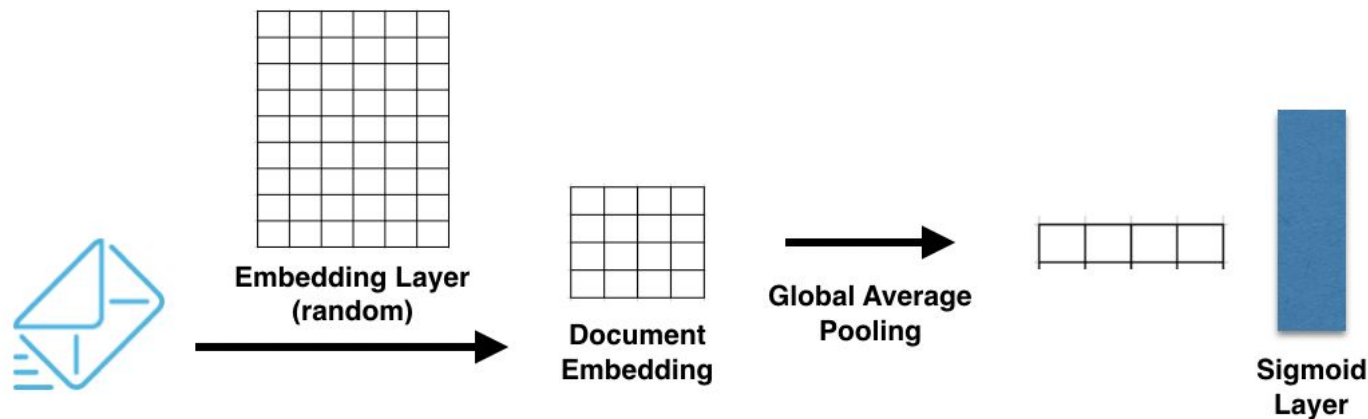
*"Contextual String Embeddings for Sequence Labeling"* (Akbik et al. 2018) COLING'18



# Considerations

- How well does the model fit to what we need to do?
  - How convoluted is the evaluation going to be?
- How well can we identify, isolate and fix errors?
  - How easy is it to tweak some specific behaviour to suit business requirements?
- Deployment?
  - Memory requirements
  - Specialised hardware
- Is there a reliable implementation available?

# Document Level Classification



- **Booking**
- **Content:** Flight? Hotel? Train? Rental Car? Transfer? Other?
- **In Scope:** is this a request we handle

*"Bag of Tricks for Efficient Text Classification"* (A Joulin et al. 2017) EACL'17

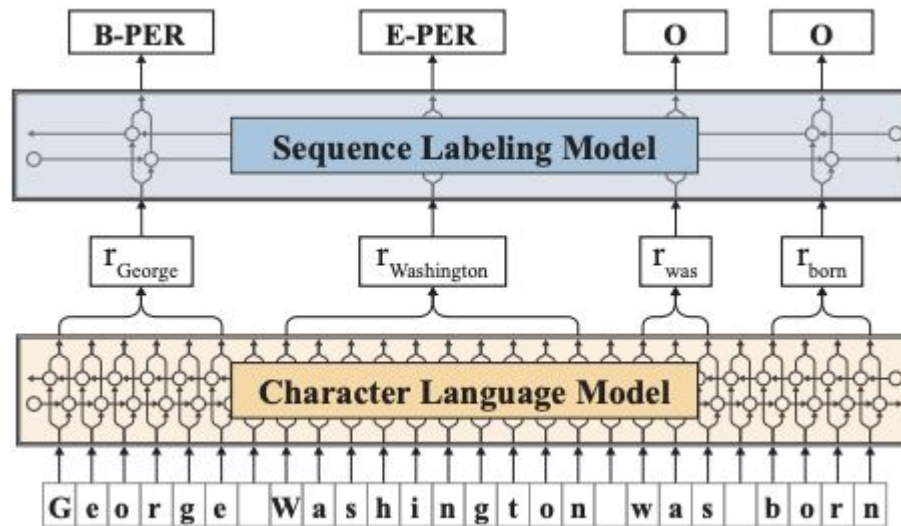
# Information Extraction / Slot Filling

Current solution inspired by Kaggle winners, developed in-house

- CNN + Bidirectional RNN (GRU) Ensemble
- Does not account for label dependencies

To be replaced *soon* with

- BiLSTM + CRF
- average absolute improvement in F1  
~13% over 71 labels
- Takes ~week to train ;(



"Contextual String Embeddings for Sequence Labeling" (Akbik et al. 2018) COLING'18

"Pooled Contextualized Embeddings for Named Entity Recognition" (Akbik et al. 2019) NAACL'19

# Semantics / Time Expressions

- supports German and English expressions
- expressions are relative to some pre-defined reference times
- resolutions are in the future relative to the reference time, unless explicitly specified in the time expression

## ctparse - Parse natural language time expressions in python

build passing codecov 97% pypi v0.0.36 pyup 1 update docs passing

This code is in early alpha stage. There can and will be potentially breaking changes right on the ``master`` branch

- Free software: MIT license
- Documentation: <https://ctparse.readthedocs.io>.

# Ongoing Research Efforts

- Document level classification
  - Tight coupling with pre-trained language models
  - Hierarchical Attention Networks
    - *"Hierarchical Attention Networks for Document Classification"* (Yang et al. 2016)  
NAACL'16
  - LSTM + MaxPooling
    - *"Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling"* (Zhou et al. 2016) COLING'16
- Time Parsing
- Recommendation ( we're hiring !!! )
- ...

# Pitfalls

- Start with simple models
- State-of-the-art is often questionable for production purposes
  - Evaluation data / methodology
  - Getting a single slot wrong invalidates entire process -> human machine hybrid
- Establish a baseline !!!
  - *"Bag of Tricks for Efficient Text Classification"* (A Joulin et al. 2017) EACL'17
- Don't be afraid to have and test crazy ideas

# Recommended Reading

- *NLP's ImageNet Moment* by Sebastian Ruder
  - <http://ruder.io/nlp-imagenet/>
- *Generalized Language Models* by Lilian Weng
  - <https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html>
- *NLP's Generalisation Problem* by Ana Marasović
  - <https://thegradient.pub/frontiers-of-generalization-in-natural-language-processing/>

# Thank you

Matti Lyra

Research Engineer

@mattilyra / @comtravo\_tech

David S. Batista

Andreas Grever

Anne Matthies

Sebastian Mika

