Natural Language Processing in Industry
Successes and Challenges

Yves Peirsman, NLP Town

nlp.town

NLP Town

We provide guidance

to the NLP domain to companies that would like to develop their AI software in-house.

We develop software

and train NLP models for challenging or domain-specific applications.

We bridge the gap

between state-of-the-art NLP research and innovative industry applications.

nlp.town

Building an NLP Community

Meetups

4 times every year we organize the Belgium NLP Meetup

Hackathon

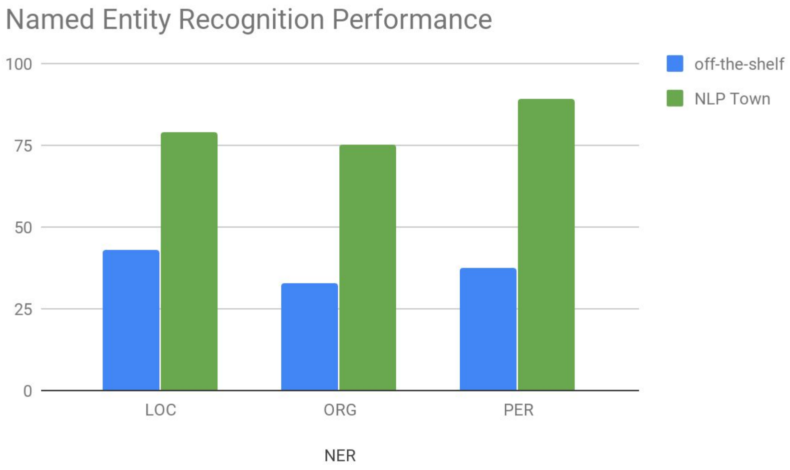
Earlier this year we co-organized the #NLP4Gov hackathon together with the Flemish Government.

Named Entity Recognition

- Collaboration with a Dutch media company
- NER: identification of people, locations, organizations
- Specific domain: Dutch financial news

Named Entity Recognition

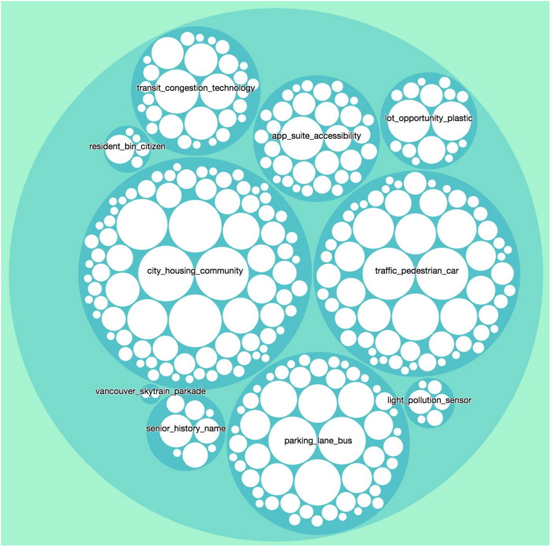
- Start from a pre-trained solution,
- Collect a large set of news articles automatically,
- Label a small set of financial articles manually.



nlp.town

Text exploration

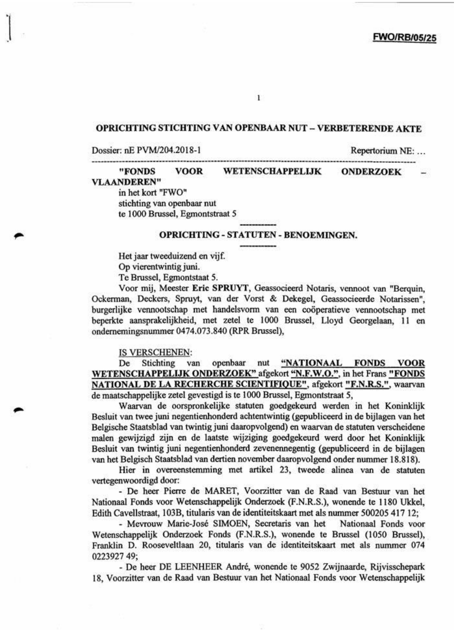
- Explore messages people send to their city council
- No labelled data
- Identification and clustering of similar messages
- Clustering does not need labelled data



nlp.town

Document parsing

- Collaboration with FEDNOT
- Parsing of founding deeds
- Combination of named entity recognition + classification


FWO/RB/05/25

1

OPRICHTING STICHTING VAN OPENBAAR NUT – VERBETERENDE AKTE

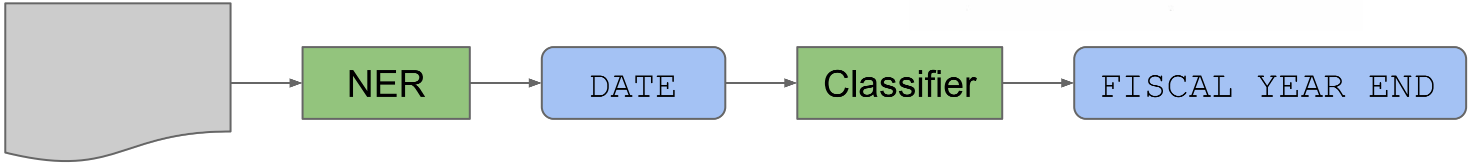
Dossier: nE PVM/204.2018-1 Repertorium NE: …

"FONDS VOOR WETENSCHAPPELIJK ONDERZOEK – VLAANDEREN"
in het kort "FWO"
stichting van openbaar nut
te 1000 Brussel, Egmontstraat 5

OPRICHTING - STATUTEN - BENOEMINGEN.

Het jaar tweeduizend en vijf.
Op vierentwintig juni.
Te Brussel, Egmontstaat 5.
Voor mij, Meester **Eric SPRUYT**, Geassocieerd Notaris, vennoot van "Berquin, Ockerman, Deckers, Spruyt, van der Vorst & Dekegel, Geassocieerde Notarissen", burgerlijke vennootschap met handelsvorm van een coöperatieve vennootschap met beperkte aansprakelijkheid, met zetel te 1000 Brussel, Lloyd Georgelaan, 11 en ondernemingsnummer 0474.073.840 (RPR Brussel),

IS VERSCHENEN:
De Stichting van openbaar nut **"NATIONAAL FONDS VOOR WETENSCHAPPELIJK ONDERZOEK"** afgekort **"N.F.W.O."**, in het Frans **"FONDS NATIONAL DE LA RECHERCHE SCIENTIFIQUE"**, afgekort **"F.N.R.S."**, waarvan de maatschappelijke zetel gevestigd is te 1000 Brussel, Egmontstraat 5,
Waarvan de oorspronkelijke statuten goedgekeurd werden in het Koninklijk Besluit van twee juni negentienhonderd achtentwintig (gepubliceerd in de bijlagen van het Belgische Staatsblad van twintig juni daaropvolgend) en waarvan de statuten verscheidene malen gewijzigd zijn en de laatste wijziging goedgekeurd werd door het Koninklijk Besluit van twintig juni negentienhonderd zevenennegentig (gepubliceerd in de bijlagen van het Belgisch Staatsblad van dertien november daaropvolgend onder nummer 18.818).
Hier in overeenstemming met artikel 23, tweede alinea van de statuten vertegenwoordigd door:
- De heer Pierre de MARET, Voorzitter van de Raad van Bestuur van het Nationaal Fonds voor Wetenschappelijk Onderzoek (F.N.R.S.), wonende te 1180 Ukkel, Edith Cavellstraat, 103B, titularis van de identiteitskaart met als nummer 500205 417 12;
- Mevrouw Marie-José SIMOEN, Secretaris van het Nationaal Fonds voor Wetenschappelijk Onderzoek Fonds (F.N.R.S.), wonende te Brussel (1050 Brussel), Franklin D. Rooseveltlaan 20, titularis van de identiteitskaart met als nummer 074 0223927 49;
- De heer DE LEENHEER André, wonende te 9052 Zwijnaarde, Rijvisschepark 18, Voorzitter van de Raad van Bestuur van het Nationaal Fonds voor Wetenschappelijk



Challenges

Great expectations

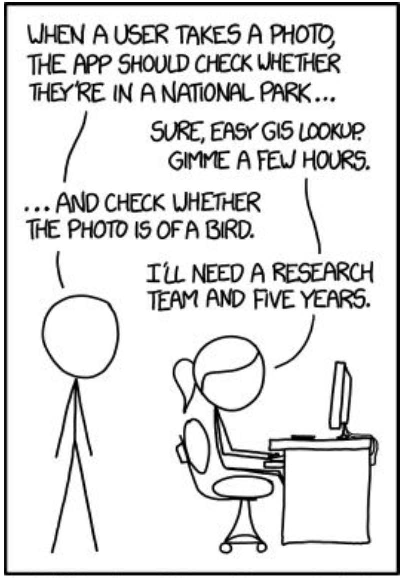
- AI hype has overblown expectations.
- Things that sound easy may be difficult and vice versa.

No one size fits all

- NLP models are language and domain-specific.
- Most resources/research are for English only.

Lack of data

- For most tasks there is a lack of labelled data.



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

nlp.town


BIG DATA BIG DATA BIG DATA BIG DATA BIG DATA BIG DATA

Data labelling

The problem with labelling

- Labelling training data is time-consuming and expensive.
- It is hard to label consistently.
- You may need several iterations you get your labels right.

We would like to reduce the labelling effort

- Distant supervision
- Active learning
- Transfer learning

nlp.town

Distant supervision

Distant supervision automatically creates labelled data by relying on external knowledge.

- Entities in documents can be labelled on the basis of their metadata
- Semantic relations can be collected from databases such as Freebase
- Tweets can be labelled with sentiment on the basis of their hashtags and smileys.

nlp.town

Distant supervision

Tartifletan @RomaneAstruc · Sep 28
A tous les vegans qui cassent les vitrines des boucheries si vous continuez nous on va allez à Jardiland et brûler toutes les graines de fruits et légumes compris ?! #Pascontent
Translate Tweet
1 2

Julien B @JulienB · Sep 28
#SFR #SFRenpanne #pascontent vous pourriez au moins informer les clients sur la panne en cours plutôt que de faire l'autruche....
Translate Tweet

Cherwoo21 @woolies21 · Sep 27
Mais wtff #twitch vous etes serieux avec les maj surprise que si on ne fait pas gaffe avec vos systèmes de tag on est plus référence.... super quoi alalala #pascontent :P
Translate Tweet
3

Nicolas Mercier @NicoMercier1 · Sep 27
Il n'est #pascontent

> **Fabien Cazenave** @FabienCazenave
> Michel Onfray claque la porte de l'Université populaire via @ouestfrance ouest-france.fr/normandie/caen…

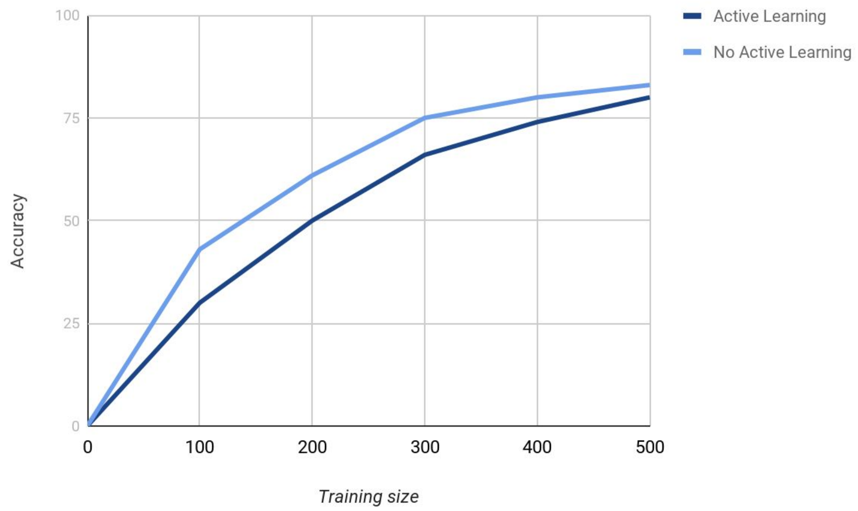
1

Active learning

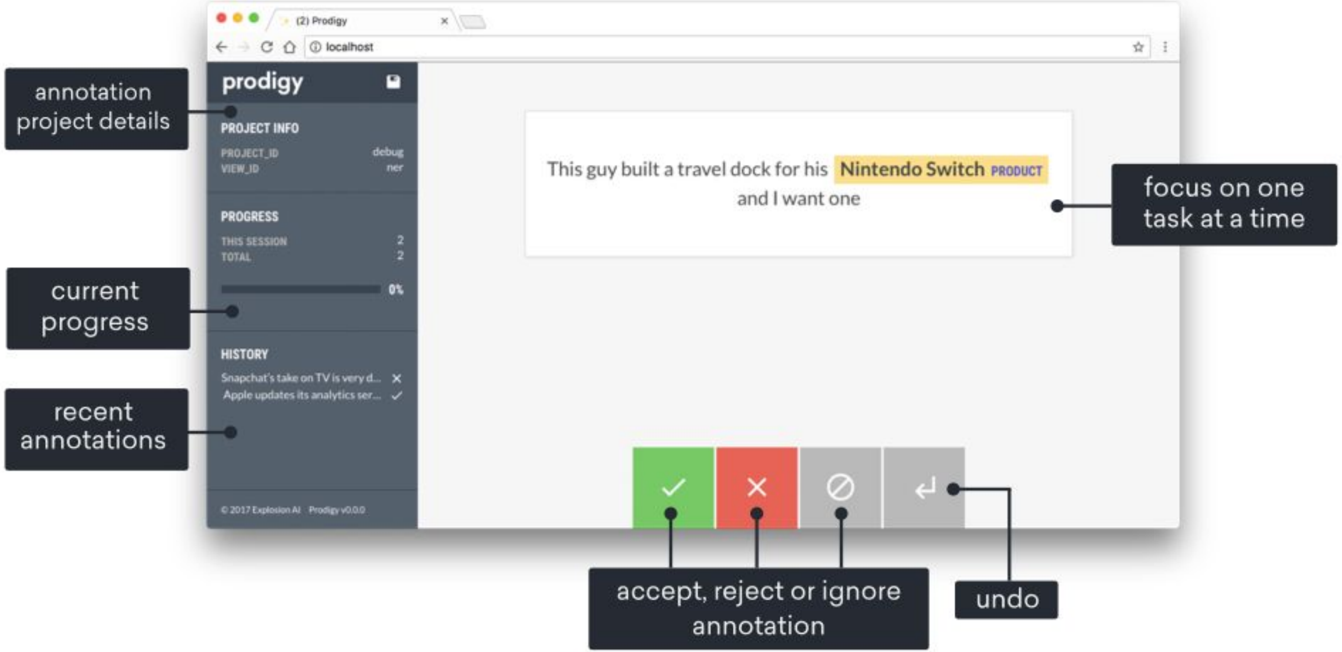
- Distantly supervised data sets are large, but potentially noisy.
- Active learning focuses on *few*, *high-quality* examples.
- Reduce the labelling effort by selecting *informative* examples.
    - Uncertainty sampling: label uncertain examples first.

Active learning

- Models learn much more quickly with examples selected through active learning.



Active learning: Prodigy

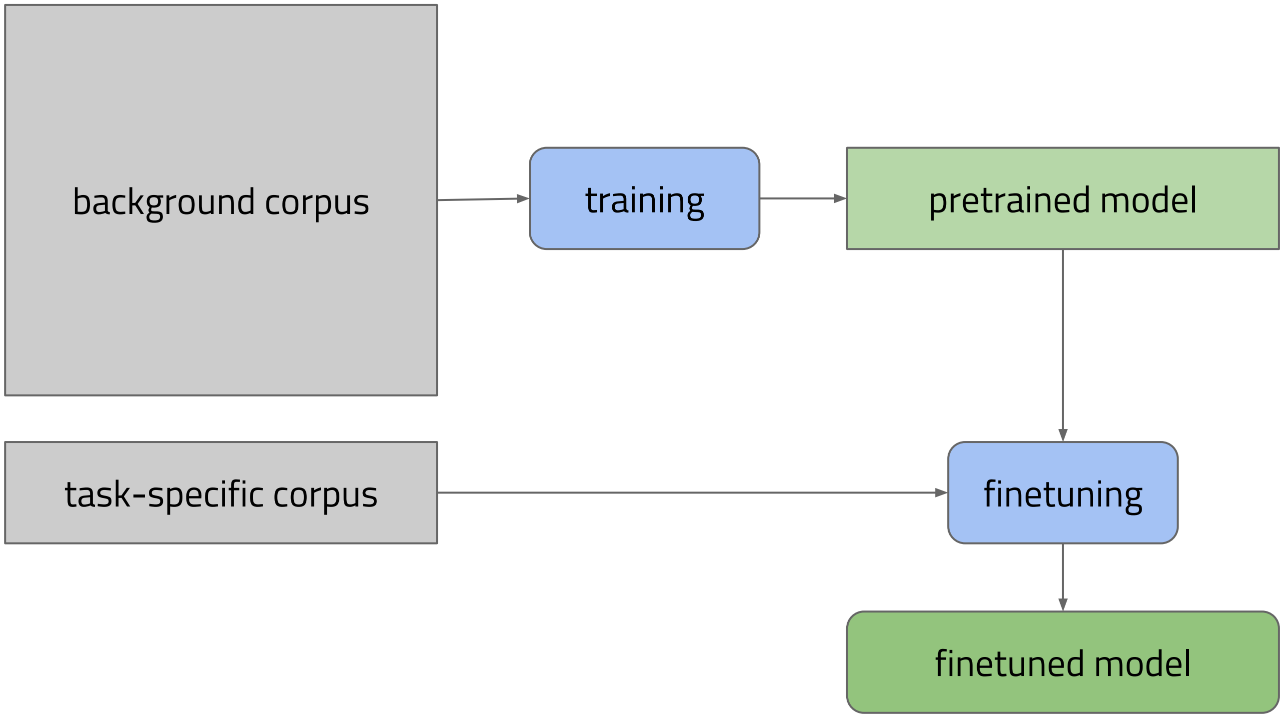


Transfer learning

- Transfer learning is one of the most exciting trends in NLP today.
- The central idea is to learn knowledge from one data source and task, and apply it to another task.
- The background data source can be labelled (see: NER before) or unlabelled (see: word embeddings).
- This can drastically reduce the number of training examples we need to build a high-quality model for the final task.

nlp.town

Transfer learning



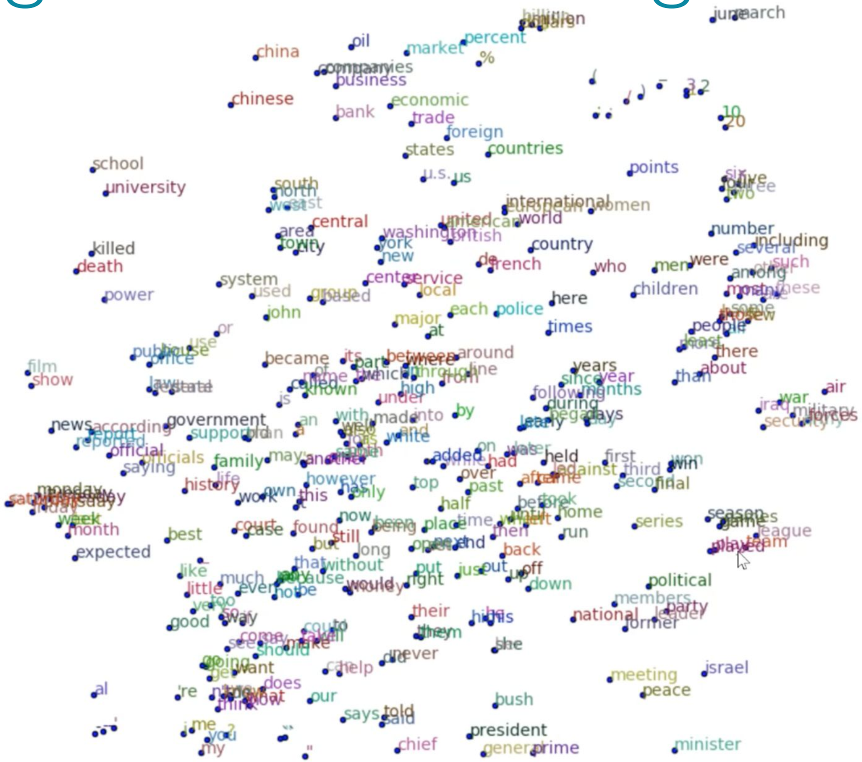
Transfer learning

Many flavours of transfer learning have been applied:

- Word embeddings
- Multilingual word embeddings
- Language models

Transfer learning: word embeddings

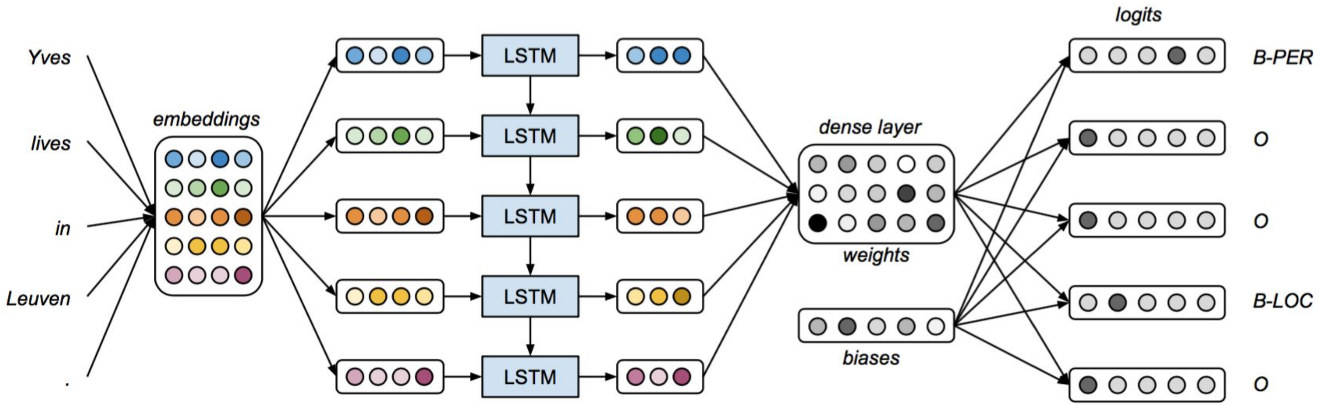
- Word embeddings have taken NLP by storm.
- Word meaning is modelled on the basis of context.



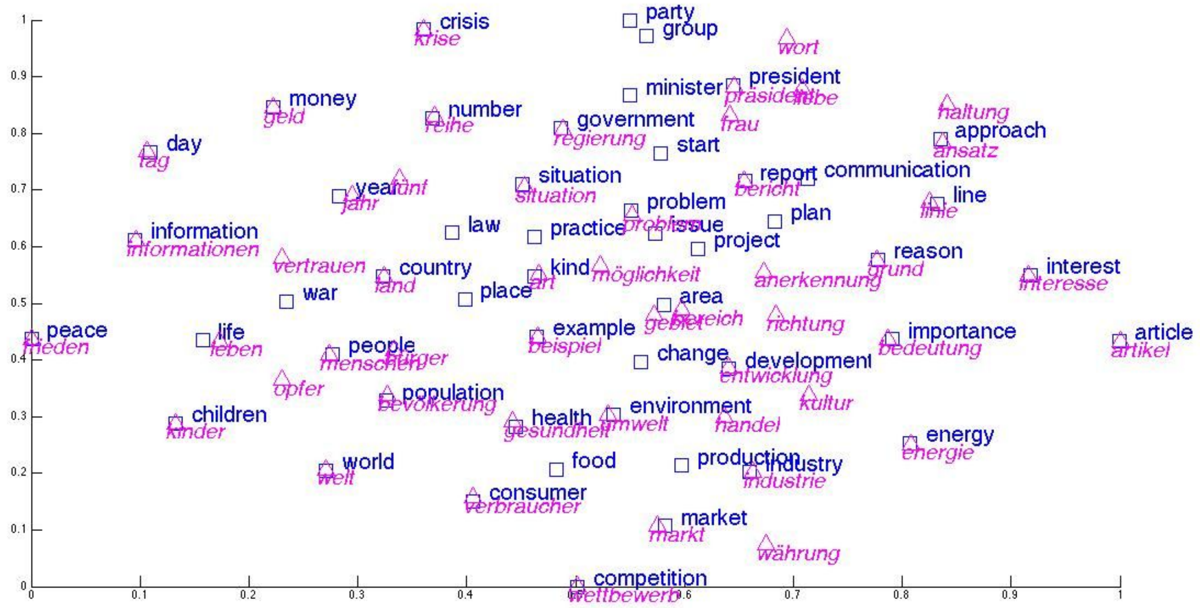
nlp.town

Transfer learning: word embeddings

- Most Deep Learning models start with an embedding layer.
- Pre-trained embeddings usually improve performance.



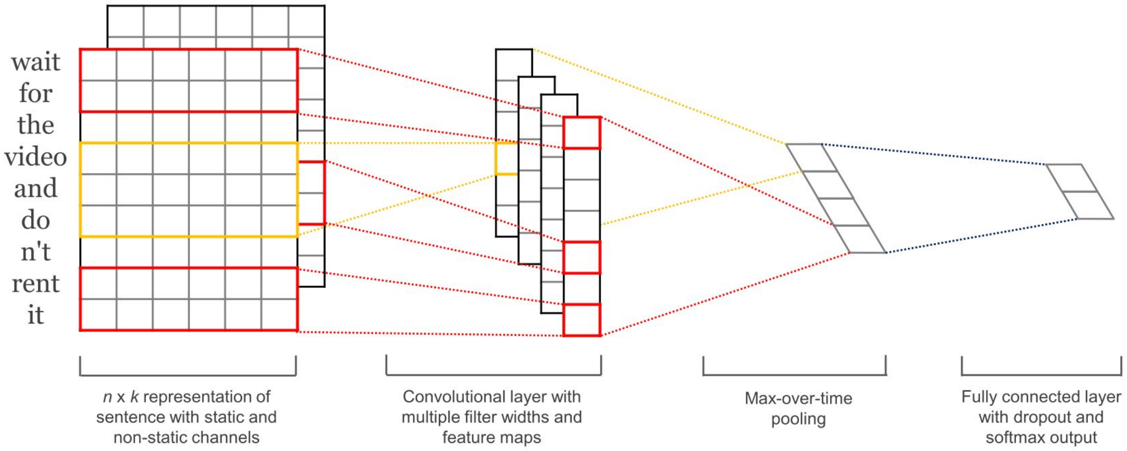
Transfer learning: cross-lingual



Luong et al., 2015

Transfer learning: cross-lingual

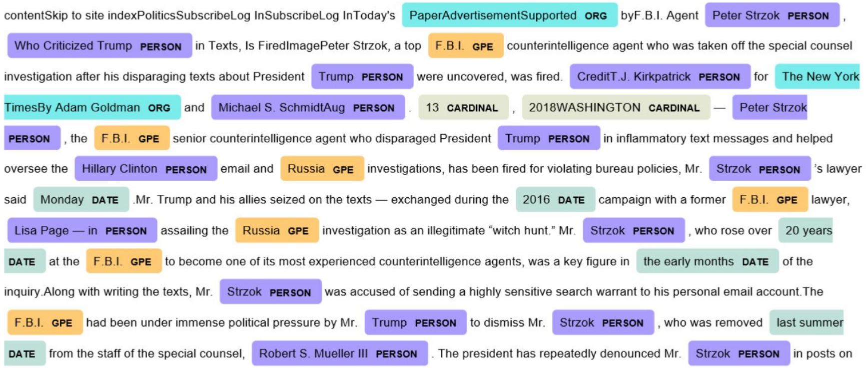
- Multilingual embeddings help us generalize across languages.
- We can swap the embedding layer and leave the rest intact.



Transfer learning: language models

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate “witch hunt.” Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Transfer learning: language models

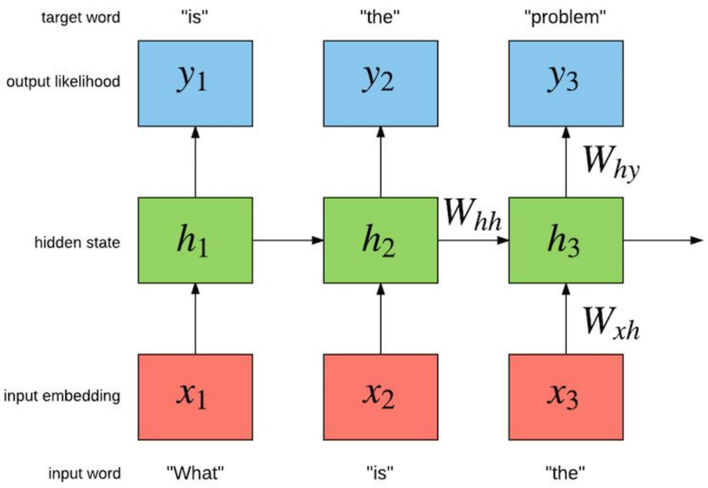


Unsupervised language model

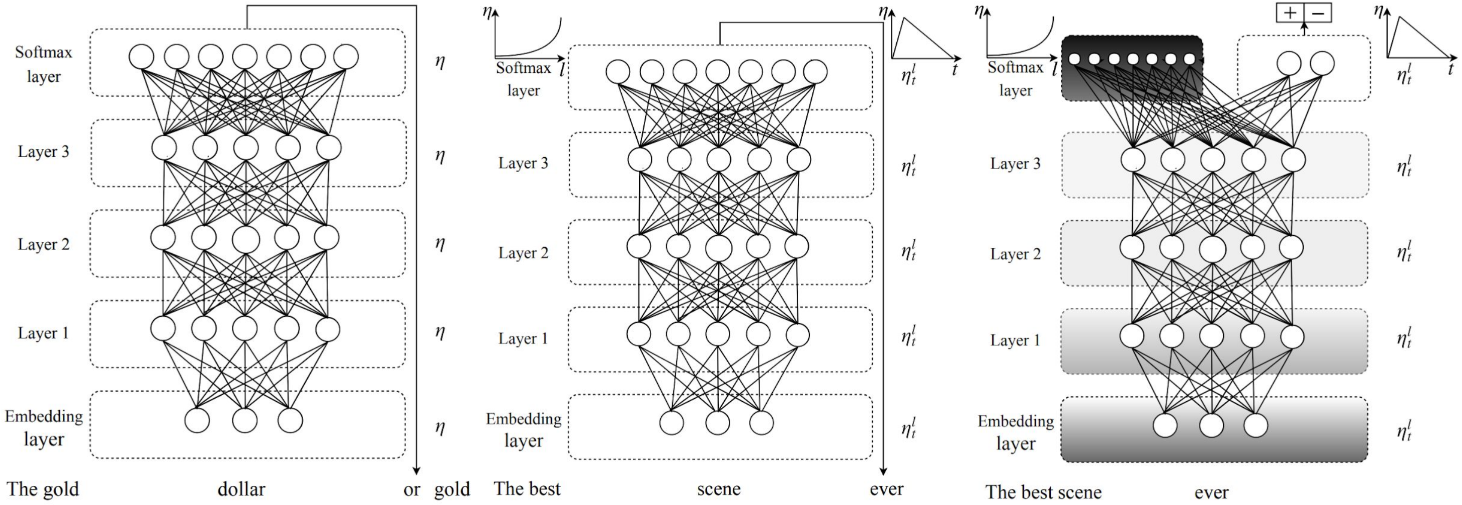
Supervised finetuning

Transfer learning: language models

- Language models learn a lot about text by trying to predict the next word.
- This could be NLP's ImageNet moment.
- Many flavours of models:
  - ELMo Embeddings
  - ULMFit
  - BERT



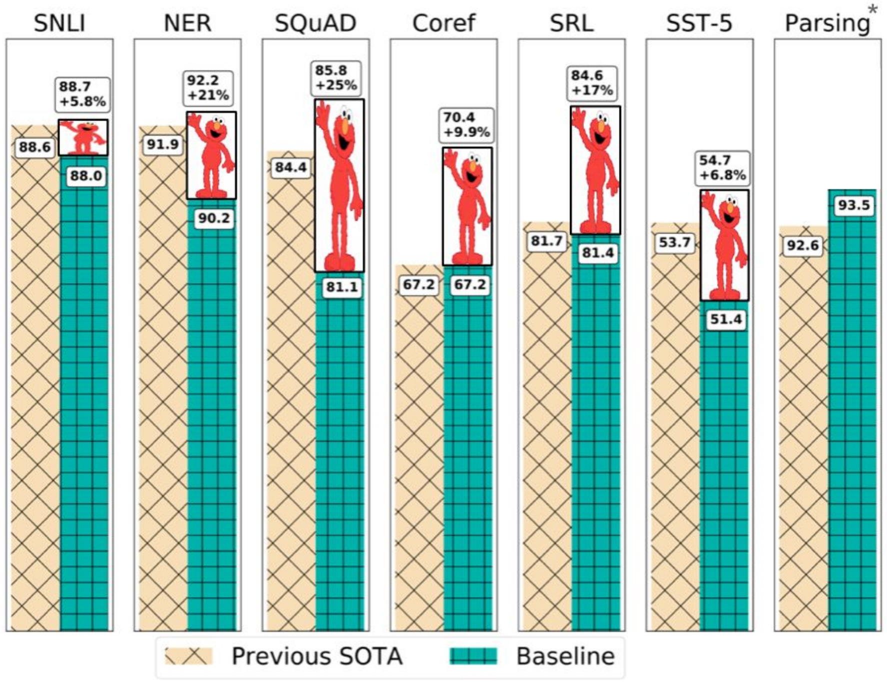
Transfer learning: language models



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

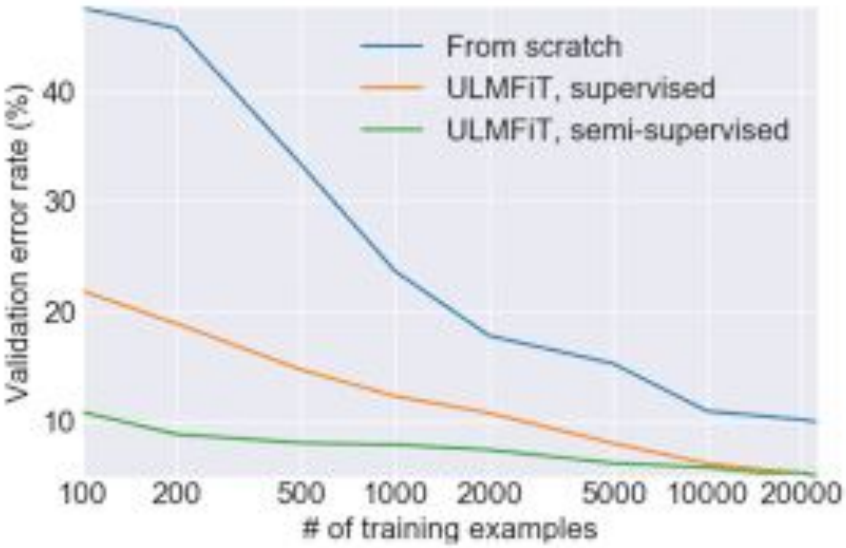
ULMFit, Howard and Ruder 2018.

Transfer learning: language models



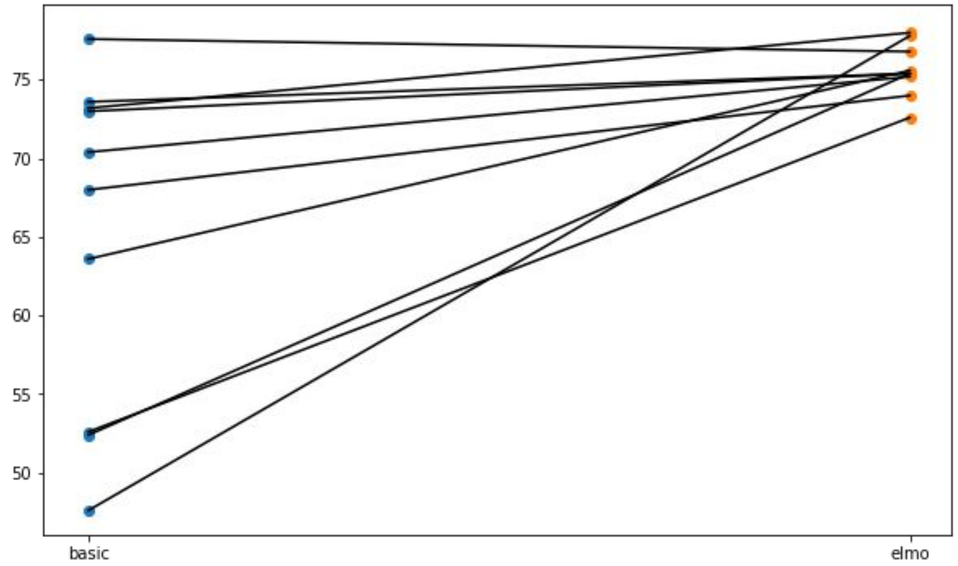
*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

Transfer learning: language models



Transfer learning: language models

Training sentiment analysis on just 200 examples:



Conclusions

- For most practical NLP applications, there is a clear lack of data.
- Organizations need to wake up to the importance of *labelled* data from the *correct domain*.
- Invest more time in your data than in your algorithms.
- Established and emerging techniques help us deal with data scarcity.

nlp.town

nlp.town

www.nlp.town
yves@nlp.town
+32 485 024 334