07.10.19

Data:
What do we need
and how can we get it?

Kiel.AI

Big Data

vs.

Small Data

APIs for Annotated Text and Pretrained Word Vectors

Stanford(Core)NLP	https://stanfordnlp.github.io/stanfordnlp/
SpaCy	https://spacy.io/

SpaCy

Non-destructive **tokenization**

Named entity recognition

pretrained **word vectors**

Built in **visualizers** for syntax and NER

Part-of-speech tagging

Labelled dependency parsing

Syntax-driven sentence segmentation

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun, in an interview with Recode earlier this week.

Noun phrases

'Sebastian Thrun', 'self-driving cars', 'Google', 'few people', 'the company', 'him', 'I', 'you', 'very senior CEOs', 'major American car companies', 'my hand', 'I', 'Thrun', 'an interview', 'Recode‘

Verbs

'start', 'work', 'drive', 'take', 'tell', 'shake', 'turn', 'talk', 'say‘

Entities

Sebastian Thrun	PERSON
Google	ORG
2007	DATE
American	NORP
Thrun	ORG
Recode	PRODUCT
earlier this week	DATE

Performance on Named Entity Recognition (NER) over time

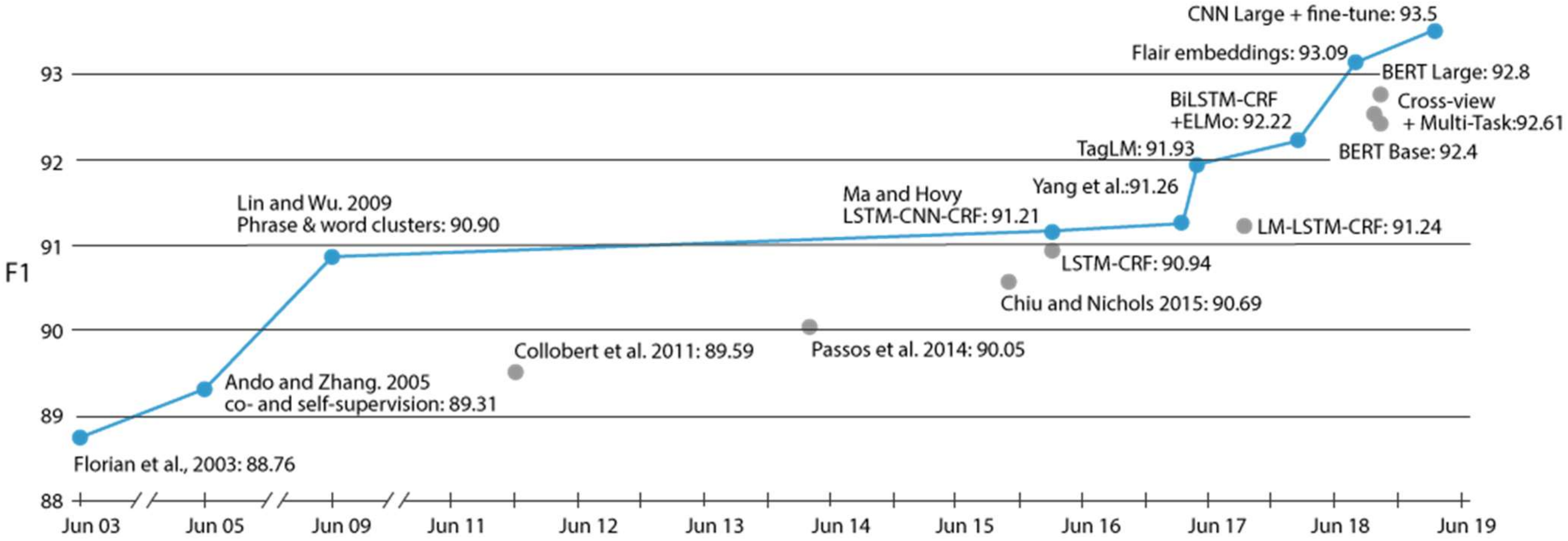
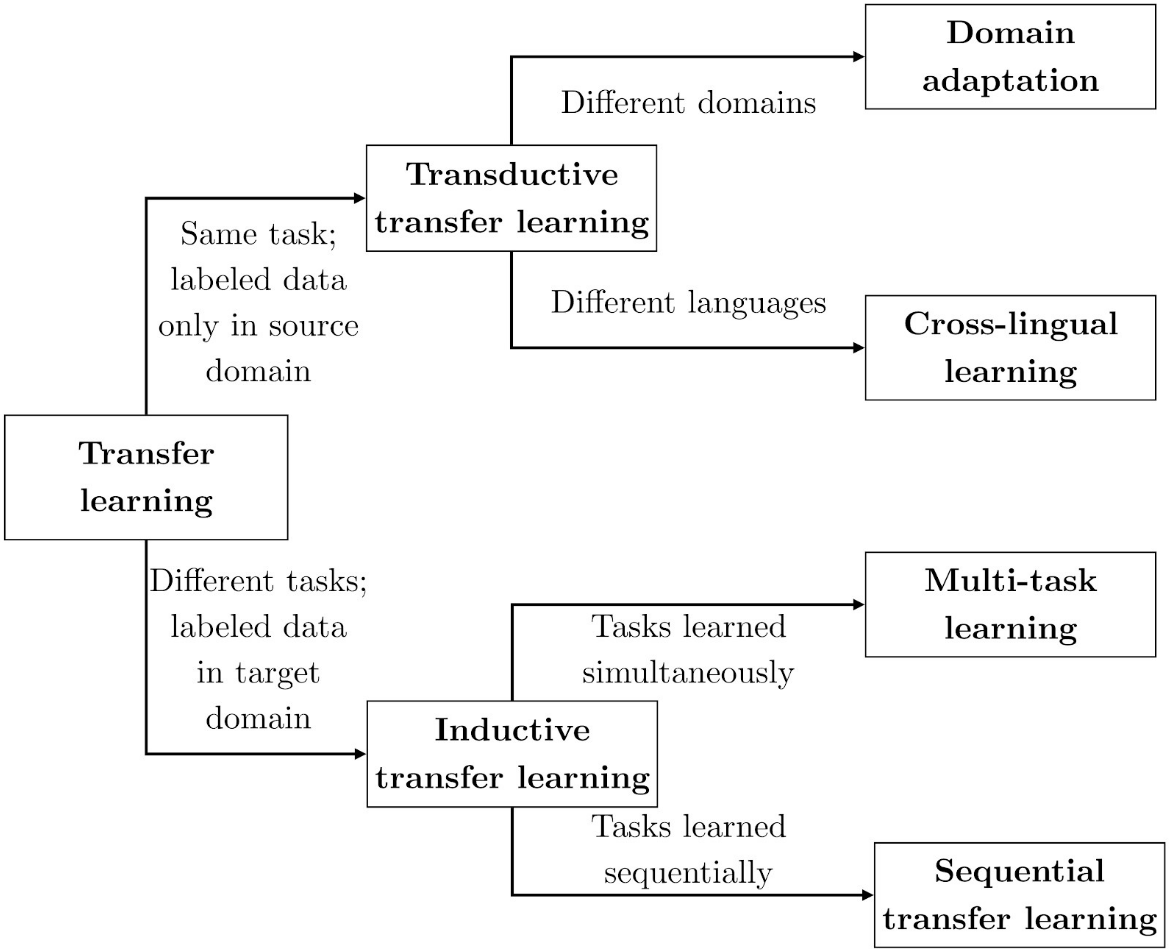
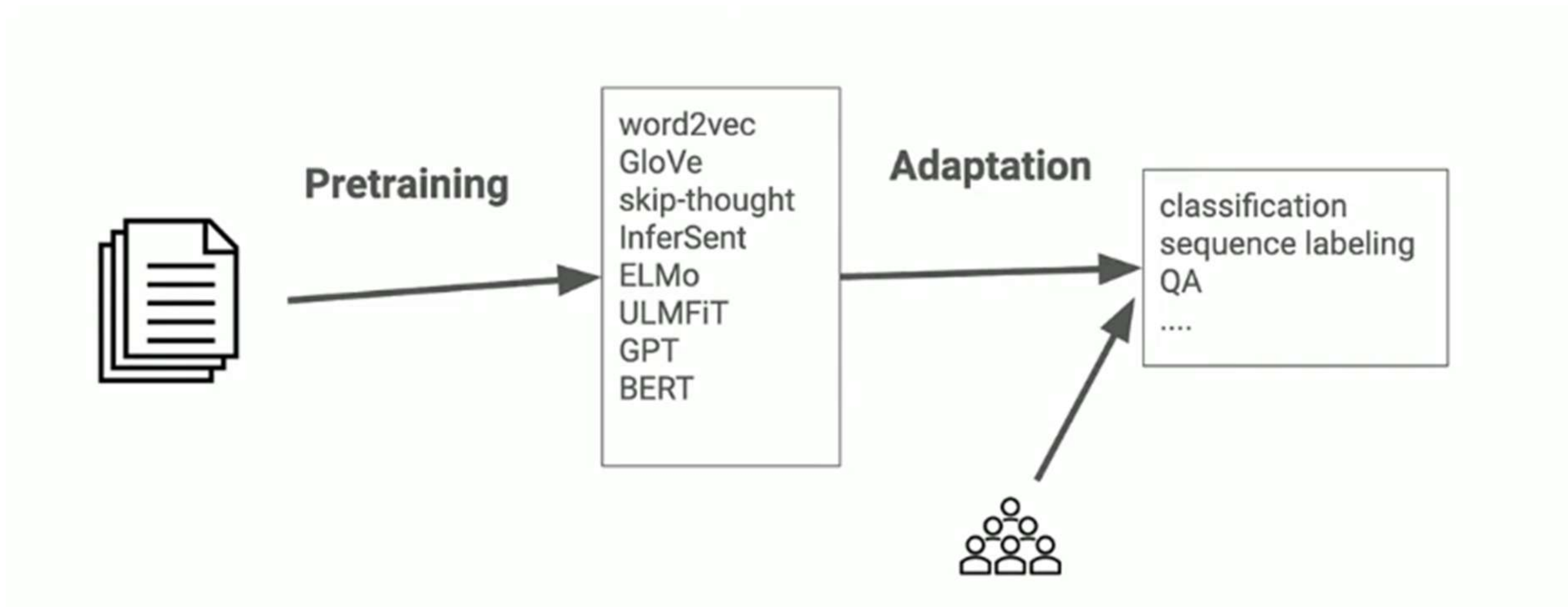


Image Credit: NAACL 2019 Transfer Learning in NLP Tutorial

OPENCAMPUS.sh


Transfer learning
Same task; labeled data only in source domain
Transductive transfer learning
Different domains
Domain adaptation
Different languages
Cross-lingual learning
Different tasks; labeled data in target domain
Inductive transfer learning
Tasks learned simultaneously
Multi-task learning
Tasks learned sequentially
Sequential transfer learning

Ruder, S. (2019). *Transfer Learning in Open-Source Natural Language Processing*. Gehalten auf der spaCy IRL 2019, Berlin. Abgerufen von https://www.youtube.com/watch?v=hNPwRPg9BrQ&list=PLBmcuObd5An4UC6jvK_-eSl6jCvP1gwXc&index=2&t=0s



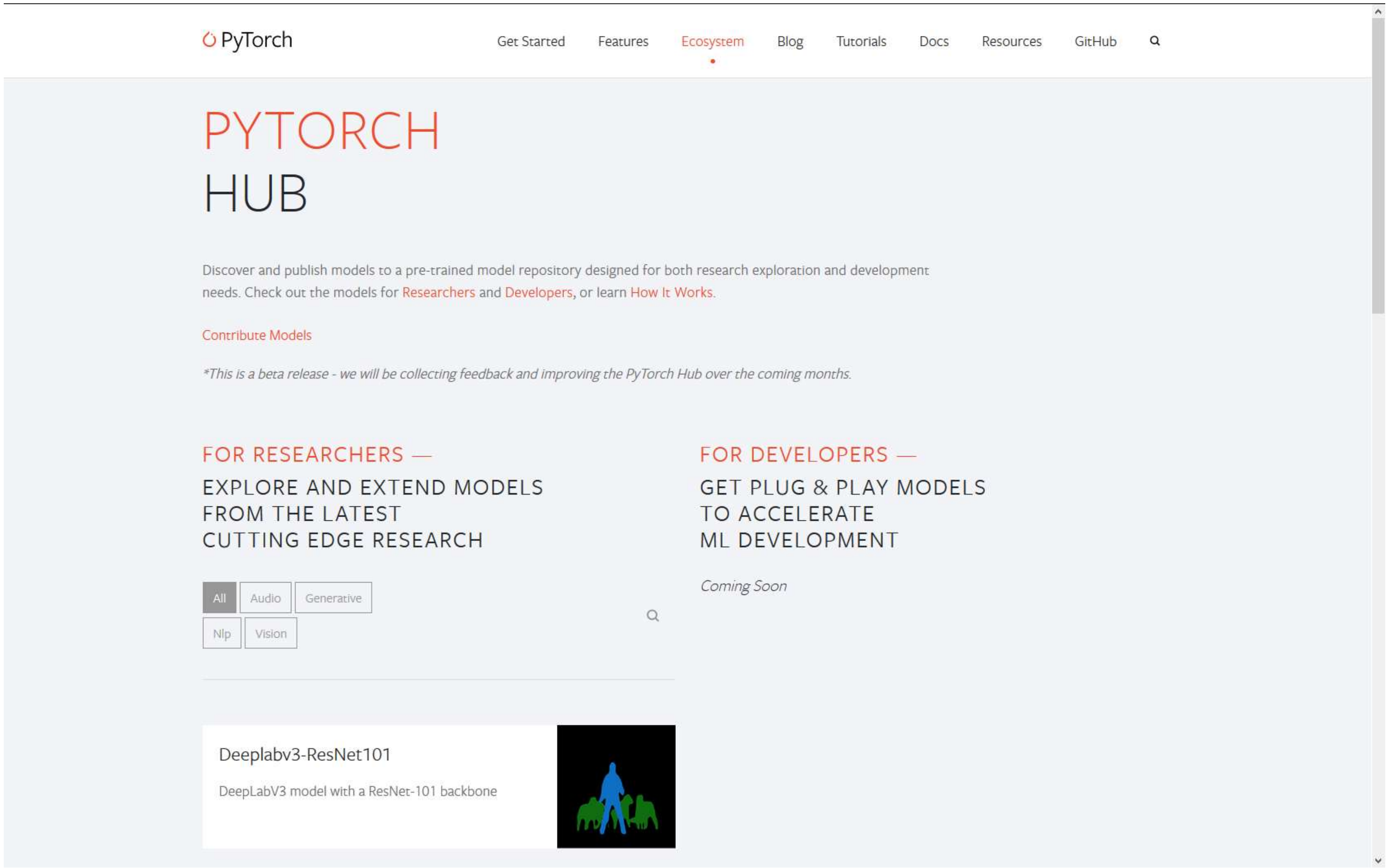
Pretraining

word2vec
GloVe
skip-thought
InferSent
ELMo
ULMFiT
GPT
BERT

Adaptation

classification
sequence labeling
QA
....

Ruder, S. (2019). *Transfer Learning in Open-Source Natural Language Processing*. Gehalten auf der spaCy IRL 2019, Berlin. Abgerufen von https://www.youtube.com/watch?v=hNPwRPg9BrQ&list=PLBmcuObd5An4UC6jvK_-eSl6jCvP1gwXc&index=2&t=0s



PYTORCH
HUB

Discover and publish models to a pre-trained model repository designed for both research exploration and development needs. Check out the models for Researchers and Developers, or learn How It Works.

Contribute Models

**This is a beta release - we will be collecting feedback and improving the PyTorch Hub over the coming months.*

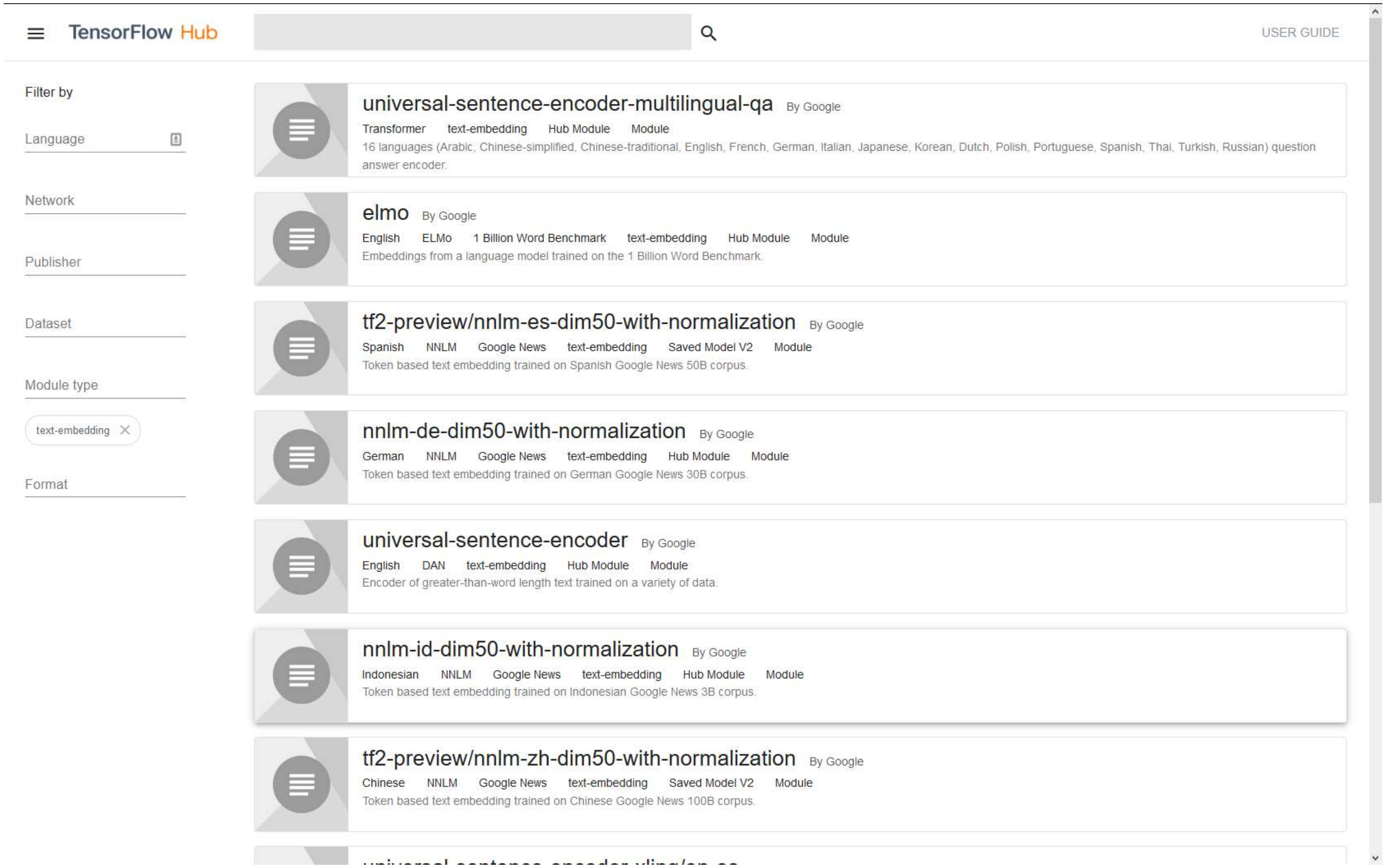
FOR RESEARCHERS —
EXPLORE AND EXTEND MODELS FROM THE LATEST CUTTING EDGE RESEARCH

Deeplabv3-ResNet101

DeepLabV3 model with a ResNet-101 backbone

FOR DEVELOPERS —
GET PLUG & PLAY MODELS TO ACCELERATE ML DEVELOPMENT

Coming Soon



universal-sentence-encoder-multilingual-qa By Google
Transformer text-embedding Hub Module Module
16 languages (Arabic, Chinese-simplified, Chinese-traditional, English, French, German, Italian, Japanese, Korean, Dutch, Polish, Portuguese, Spanish, Thai, Turkish, Russian) question answer encoder.

elmo By Google
English ELMo 1 Billion Word Benchmark text-embedding Hub Module Module
Embeddings from a language model trained on the 1 Billion Word Benchmark.

tf2-preview/nnlm-es-dim50-with-normalization By Google
Spanish NNLM Google News text-embedding Saved Model V2 Module
Token based text embedding trained on Spanish Google News 50B corpus.

nnlm-de-dim50-with-normalization By Google
German NNLM Google News text-embedding Hub Module Module
Token based text embedding trained on German Google News 30B corpus.

universal-sentence-encoder By Google
English DAN text-embedding Hub Module Module
Encoder of greater-than-word length text trained on a variety of data.

nnlm-id-dim50-with-normalization By Google
Indonesian NNLM Google News text-embedding Hub Module Module
Token based text embedding trained on Indonesian Google News 3B corpus.

tf2-preview/nnlm-zh-dim50-with-normalization By Google
Chinese NNLM Google News text-embedding Saved Model V2 Module
Token based text embedding trained on Chinese Google News 100B corpus.

universal-sentence-encoder-xling/en-es

General Dataset Repositories / Directories

skymind	https://skymind.ai/wiki/open-datasets
Kaggle	https://www.kaggle.com/datasets
OpenML	https://www.openml.org/search?type=data
Google Dataset Search	https://toolbox.google.com/datasetsearch

Working With Wordnets

Natural Language Toolkit (NLTK) for Python http://www.nltk.org/

WordNet (English)	https://wordnet.princeton.edu/
GermaNet (German)	http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml
Open DE WordNet (German)	https://github.com/hdaSprachtechnologie/odenet