

Developing NLP models without labelled data

Pierre Lison
plison@nr.no

Norsk Forening for Kvantitativ Finans
4.. mars 2020



“FinAI”

- ▶ R&D project in collaboration with Exabel:
 - “FinAI: Artificial Intelligence tool to monitor global financial markets”
- ▶ Development of a range of novel statistical / machine learning techniques for financial investors
 - Turn massive streams of data into actionable insights!
 - Including both text data and other sources

Exabel

 **Forskningsrådet**

NLP for finance

- ▶ NLP is a crucial tool to extract structured information from unstructured text sources
 - News articles, social media, financial reports, etc.
- ▶ Pretrained NLP models available for tasks such as *Named Entity Recognition*
 - At least for English (and a few other languages)
- ▶ However, their performance degrade rapidly when applied to texts that differ from the type of texts observed in the training set



Example (with Spacy)

```
import spacy, annotations

# We load the spacy model
nlp = spacy.load("en")
doc = nlp(news_text)

# Visualising the entities
annotations.display_entities(doc)
```

ATLANTA ORG (Reuters ORG) - Retailer Best Buy Co, seeking new ways to appeal to cost-conscious shoppers, said on Tuesday DATE it is selling refurbished versions of Apple Inc's ORG iPhone 3G at its stores that are priced about \$50 MONEY less than new iPhones ORG . The electronics chain said the used iPhones LOC , which were returned within 30 days DATE of purchase, are priced at \$ 149 MONEY for the model with 8 CARDINAL gigabytes of storage, while the 16-gigabyte version is \$ 249 MONEY . A two-year DATE service contract with AT&T Inc ORG is required. New iPhone ORG 3Gs currently sell for \$ 199 MONEY and \$ 299 MONEY at Best Buy Mobile ORG stores. "This is focusing on customers' needs, trying to provide as wide a range of products and networks for our consumers," said Scott Moore PERSON , vice president of marketing for Best Buy Mobile ORG . Buyers of first ORDINAL -generation iPhones can also upgrade to the faster refurbished 3 CARDINAL G models at Best Buy ORG , he said. Moore PERSON said AT&T ORG , the exclusive wireless provider for the iPhone ORG , offers refurbished iPhones ORG online. The sale of used iPhones LOC comes as Best Buy, the top consumer electronics chain, seeks ways to fend off increased competition from discounters such as Wal-Mart Stores Inc ORG , which began selling the popular phone late last month DATE . Wal-Mart ORG sells a new 8-gigabyte iPhone 3 CARDINAL G for \$197 and \$297 MONEY for the 16-gigabyte model. The iPhone ORG is also sold at Apple ORG stores and AT&T ORG stores. Moore said Best Buy's ORG move was not in response to other retailers' actions. (Reporting by Karen Jacobs PERSON ; Editing by Andre Grenon PERSON)

The data problem

- ▶ NLP models can be quite “brittle” and suffer from even small mismatch (e.g. differences in punctuation or casing) between training data and the actual “test” data
- ▶ How can we develop NLP models that are better tailored to the type of texts we must deal with?
 - a. Collect & annotate our own training data and train new model on it → Expensive & time-consuming!
 - b. Use *transfer learning* to adapt an existing model → Cheaper, but need a good model to start with, along with labelled data from the target domain
 - c. Use *weak supervision* to automatically annotate data from our target domain



Outline

- ▶ A short intro to weak supervision
- ▶ Weak supervision for named entity recognition
- ▶ Results & conclusion

Based on the paper “Named Entity Recognition without Labelled Data: A weak supervision approach”, currently under review for ACL 2020



Outline

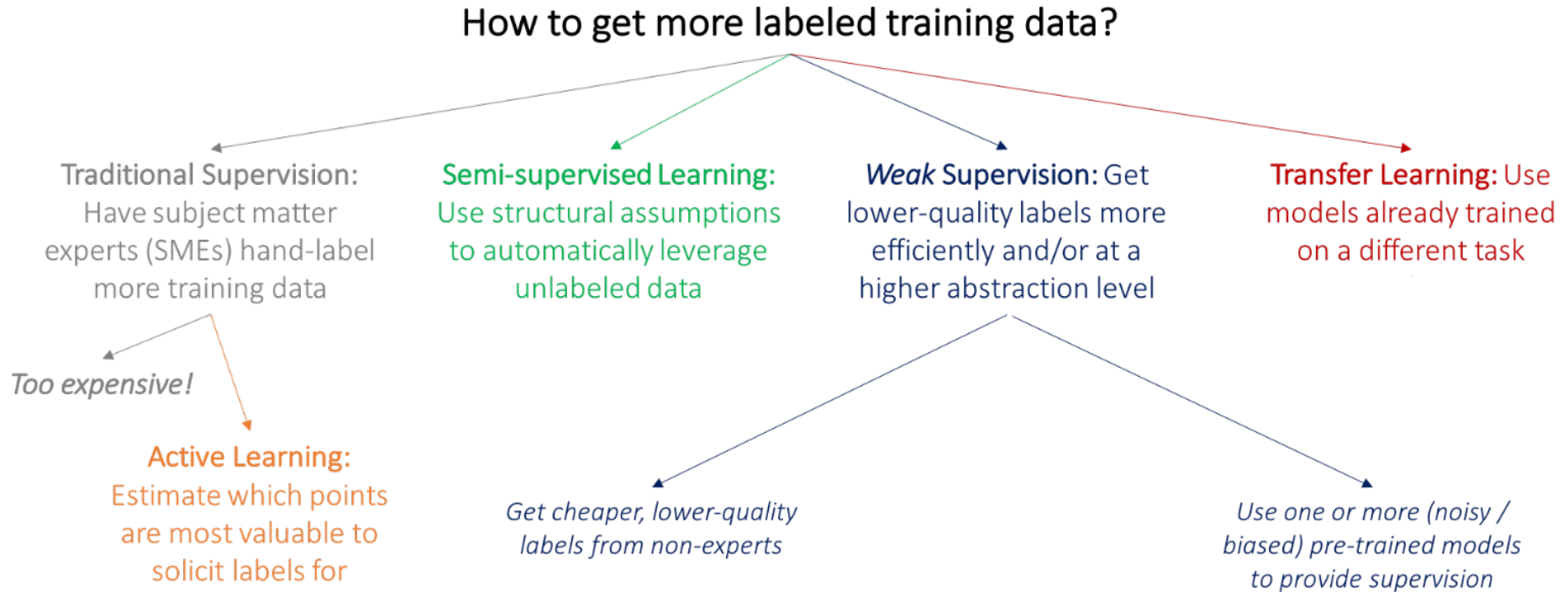
- ▶ **A short intro to weak supervision**
- ▶ Weak supervision for named entity recognition
- ▶ Results & conclusion

Based on the paper “Named Entity Recognition without Labelled Data: A weak supervision approach”, currently under review for ACL 2020

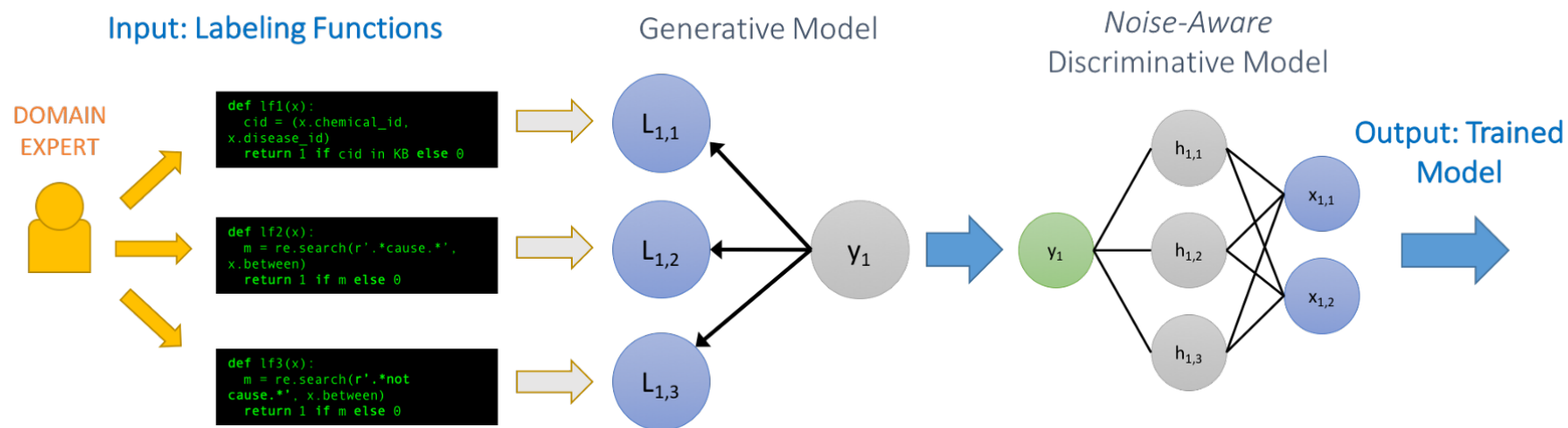
Weak supervision

- ▶ **Goal:** train machine learning through “weaker / noisier” supervision signals obtained from different sources
 - Instead of relying on a single “gold standard”
- ▶ Approach:
 1. We start with some raw, unlabelled data
 2. We apply several “*labelling functions*” (heuristics etc.) to obtain some (imperfect) labels of our data
 3. We then *aggregate* together the labels
 4. And we train our ML model on these aggregated labels

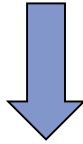
Relation to other ML paradigms



One popular weak supervision framework is **Snorkel**:
(www.snorkel.org, originally from Stanford U.)

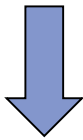


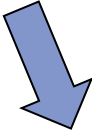
One popular weak supervision framework is **Snorkel**:
(www.snorkel.org, originally from Stanford U.)



However, Snorkel has some limitations:

- Assumes all data points are i.i.d
- Cannot take into account “probabilistic” labels



Not well suited for sequence labelling tasks such as Named Entity Recognition (consecutive words in a sentence are not i.i.d.!) 

We have developed a *novel weak supervision approach* tailored to NER and other sequence labelling tasks

Outline

- ▶ A short intro to weak supervision
- ▶ **Weak supervision for named entity recognition**
- ▶ Results & conclusion

Based on the paper “Named Entity Recognition without Labelled Data: A weak supervision approach”, currently under review for ACL 2020

Weak supervision for NER

- ▶ The goal of Named Entity Recognition is to detect entities related to persons, organisations, places, dates
 - Each entity may consist of one or several words

Donald J. Trump PERSON replied to North Korea's GPE leader Kim Jong-un PERSON
on January 6 DATE , according to the White House ORG .

- Named Entity Recognition is a crucial step for most text mining tasks
- Framed as a *sequence labelling problem* (decide for each word whether it is part of a named entity or not)

Approach

1. We first collect large amounts of text data from the domain we are interested in
 1. E.g. financial news from Reuters and Bloomberg
2. We then define many alternative “labelling functions” that will automatically annotate the texts with named entities
3. We then aggregate all these annotations
4. And finally train a sequence labelling model (a standard convolutional neural net) on these aggregated annotations

Labelling functions

1. NER models with out-of-domain data:

- 4 distinct models, using training data from news articles, broadcast, tweets, SEC reports, etc.

2. Gazetteers:

- Wikipedia (persons, organisations, places, products)
- Geonames
- Crunchbase
- DBPedia
- Curated list of countries, languages, nationalities, political and religious groups

Labelling functions

3. Heuristic functions:

- Recognition of proper names
- Recognition of dates, money amounts, percents
- Recognition of legal references
- Etc.

4. Document-level relations:

- Label consistency
- Document history

named entities occurring multiple times through a document have a high probability of belonging to the same category

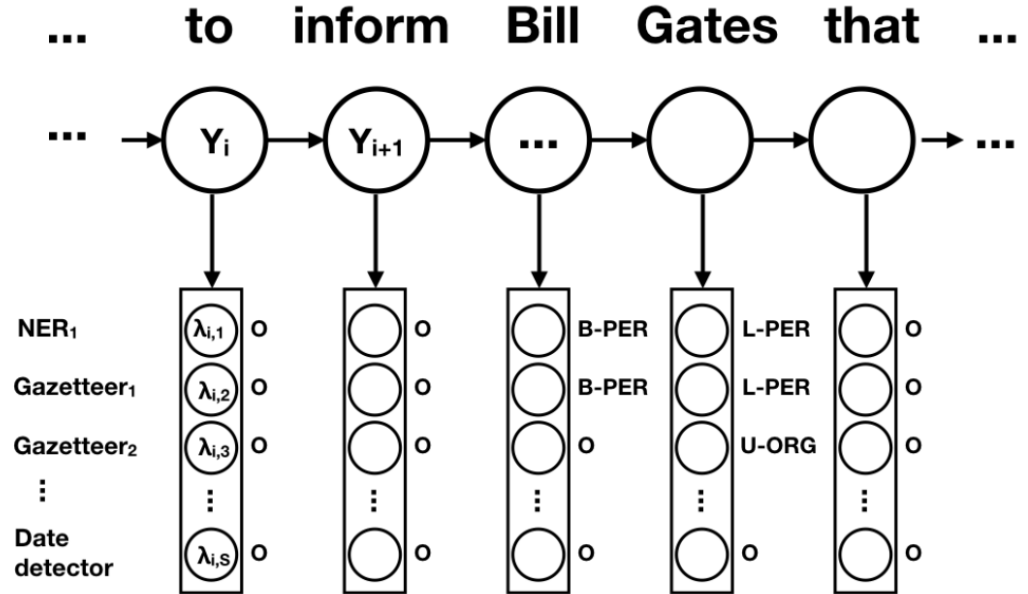
Entities are often introduced in “long-form” in a document, then in shorter form afterwards

Example (gazetteer from Crunchbase)

ATLANTA (Reuters COMPANY) - Retailer Best Buy Co, seeking new ways to appeal to cost-conscious shoppers, said on Tuesday it is selling refurbished versions of Apple Inc COMPANY 's iPhone 3G at its stores that are priced about \$50 less than new iPhones. The electronics chain said the used iPhones, which were returned within 30 days of purchase, are priced at \$149 for the model with 8 gigabytes of storage, while the 16-gigabyte version is \$249. A two-year service contract with AT&T Inc COMPANY is required. New iPhone 3Gs currently sell for \$199 and \$299 at Best Buy Mobile stores. "This is focusing on customers' needs, trying to provide as wide a range of products and networks for our consumers," said Scott Moore PERSON , vice president of marketing for Best Buy Mobile. Buyers of first-generation iPhones can also upgrade to the faster refurbished 3G models at Best Buy COMPANY , he said. Moore COMPANY said AT&T COMPANY , the exclusive wireless provider for the iPhone, offers refurbished iPhones online. The sale of used iPhones comes as Best Buy COMPANY , the top consumer electronics chain, seeks ways to fend off increased competition from discounters such as Wal-Mart Stores Inc, which began selling the popular phone late last month. Wal-Mart sells a new 8-gigabyte iPhone 3G for \$197 and \$297 for the 16-gigabyte model. The iPhone is also sold at Apple stores and AT&T stores. Moore COMPANY said Best Buy COMPANY 's move was not in response to other retailers' actions. (Reporting by Karen Jacobs PERSON ; Editing by Andre Grenon)

Aggregation

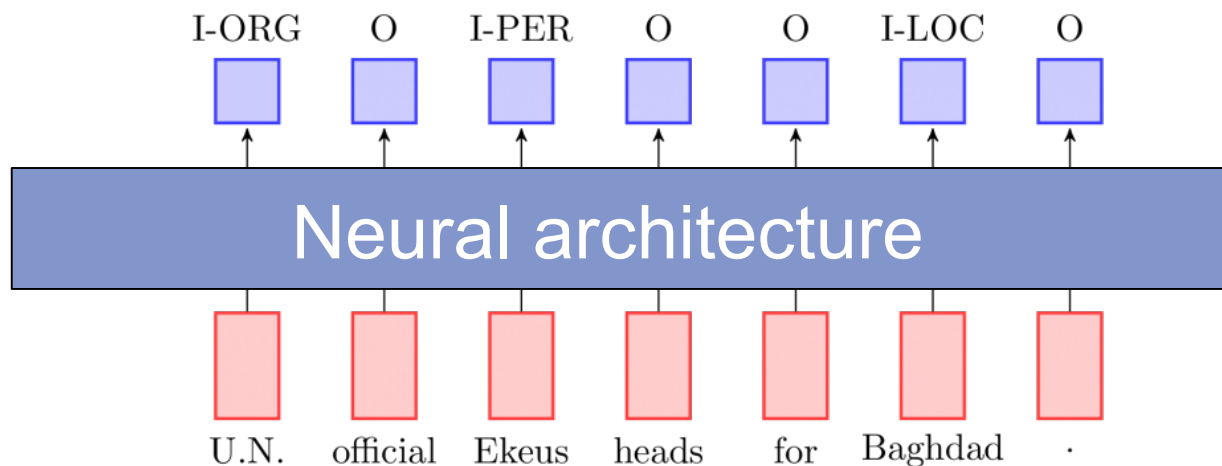
- ▶ Once we have defined all labelling functions, we need to **aggregate** them



- ▶ We rely on a simple Hidden Markov Model (HMM)
 - The states are the underlying (unknown) labels
 - One emission per labelling function
- ▶ Model estimation using the Baum-Welch algorithm

Final model

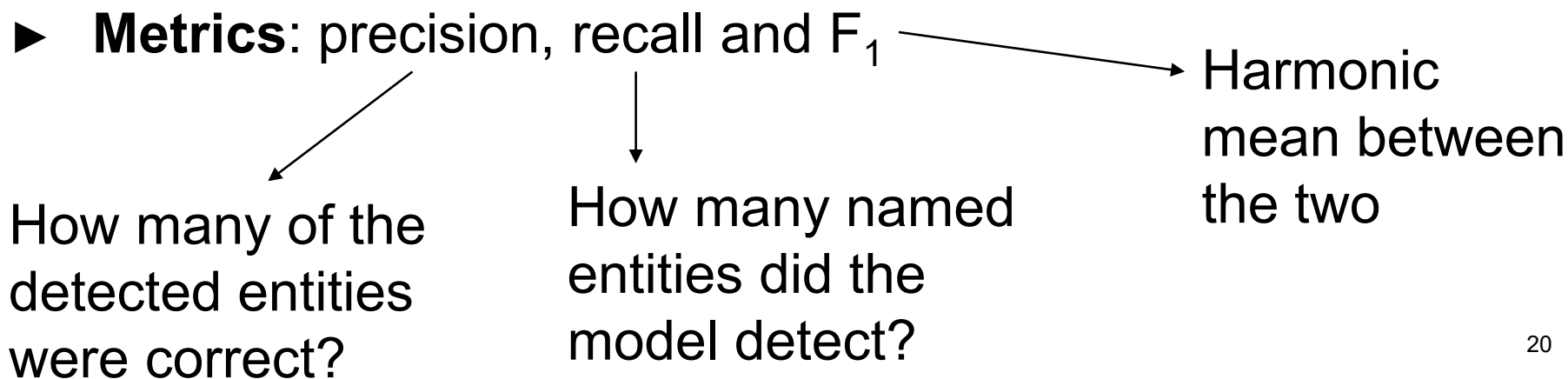
- ▶ Finally, the machine learning model is trained on the (probabilistic) aggregated annotations:



- ▶ We employed a convolutional network (with four layers) in our experiments, but any other architecture may be used

Evaluation

- ▶ We evaluated our approach on two datasets:
 - A corpus of older news articles from ConLL 2003
 - A newer collection of texts from Reuters and Bloomberg annotated via crowdsourcing
- ▶ We compared our approach to several baselines, including the current state-of-the-art in unsupervised domain adaptation: “AdaptaBERT” (presented in 2019)



Evaluation (ConLL 2003)

Model	Token-level				Entity-level		
	P	R	F_1	CEE	P	R	F_1
Ontonotes-trained NER	0.719	0.706	0.712	2.671	0.694	0.620	0.654
MV -aggregated labels	0.815	0.675	0.738	2.047	0.751	0.619	0.678
acc -aggregated labels	0.704	0.689	0.696	2.818	0.662	0.603	0.632
CM -aggregated labels	0.704	0.689	0.696	2.818	0.662	0.603	0.632
CV -aggregated labels	0.705	0.691	0.698	2.811	0.660	0.604	0.630
seq -aggregated labels	0.682	0.666	0.674	2.851	0.617	0.576	0.596
Snorkel-aggregated labels	0.710	0.661	0.684	2.264	0.714	0.621	0.664
AdaptaBERT (OntoNotes)	0.693	0.733	0.712	2.280	0.652	0.736	0.691
HMM-aggregated labels (only NER models)	0.658	0.720	0.688	2.653	0.642	0.599	0.620
HMM-aggregated labels (only gazetteers)	0.759	0.394	0.518	3.678	0.687	0.367	0.478
HMM-aggregated labels (heuristics)	0.722	0.771	0.746	1.989	0.718	0.683	0.700
HMM-aggregated labels (all but doc-level)	0.714	0.778	0.744	1.878	0.713	0.693	0.702
HMM-aggregated labels (all functions)	0.719	0.794	0.754	1.812	0.721	0.713	0.716
Neural net trained on HMM-agg. labels	0.712	0.790	0.748	2.282	0.715	0.707	0.710

Table 1: Evaluation results on CoNLL 2003. MV=Majority Voter, P=Precision, R=Recall, CEE=Cross-entropy Error (lower is better). The results are micro-averaged on all labels (PER, ORG, LOC and MISC).

Evaluation (Reuters & Bloomberg)

Model	Token-level				Entity-level		
	P	R	F_1	CEE	P	R	F_1
OntoNotes-trained NER	0.793	0.791	0.792	2.648	0.694	0.635	0.664
MV -aggregated labels	0.847	0.394	0.538	4.015	0.750	0.388	0.512
CM -aggregated labels	0.781	0.783	0.782	2.592	0.688	0.633	0.660
AdaptaBERT (OntoNotes)	0.799	0.801	0.800	2.351	0.668	0.734	0.699
HMM-aggregated labels (all functions)	0.804	0.823	0.814	2.219	0.749	0.697	0.722
Neural net trained on HMM-agg. labels	0.805	0.827	0.816	2.448	0.749	0.701	0.724

Table 2: Evaluation results on 1094 crowd-annotated sentences from Reuters and Bloomberg news articles. The results are micro-averaged on 8 labels (PERSON, NORP, ORG, LOC, PRODUCT, DATE, PERCENT, and MONEY).

Output example

```
import labelling
full_annotator.annotate(doc)
unified_model.annotate(doc)
annotations.display_entities(doc, "HMM")
```

ATLANTA GPE (Reuters COMPANY) - Retailer Best Buy Co ORG , seeking new ways to appeal to cost-conscious shoppers, said on Tuesday DATE it is selling refurbished versions of Apple Inc COMPANY 's iPhone 3G PRODUCT at its stores that are priced about \$50 MONEY less than new iPhones PRODUCT . The electronics chain said the used iPhones PRODUCT , which were returned within 30 days DATE of purchase, are priced at \$149 MONEY for the model with 8 gigabytes QUANTITY of storage, while the 16-gigabyte CARDINAL version is \$249 MONEY . A two-year DATE service contract with AT&T Inc COMPANY is required. New iPhone 3Gs PRODUCT currently sell for \$199 MONEY and \$299 MONEY at Best Buy Mobile ORG stores. "This is focusing on customers' needs, trying to provide as wide a range of products and networks for our consumers," said Scott Moore PERSON , vice president of marketing for Best Buy Mobile COMPANY . Buyers of first-generation iPhones PRODUCT can also upgrade to the faster refurbished 3 CARDINAL G models at Best Buy COMPANY , he said. Moore PERSON said AT&T COMPANY , the exclusive wireless provider for the iPhone PRODUCT , offers refurbished iPhones PRODUCT online. The sale of used iPhones PRODUCT comes as Best Buy COMPANY , the top consumer electronics chain, seeks ways to fend off increased competition from discounters such as Wal-Mart Stores Inc COMPANY , which began selling the popular phone late last month DATE . Wal-Mart LOC sells a new 8-gigabyte iPhone 3G PRODUCT for \$197 MONEY and \$297 MONEY for the 16-gigabyte CARDINAL model. The iPhone PRODUCT is also sold at Apple COMPANY stores and AT&T COMPANY stores. Moore PERSON said Best Buy COMPANY 's move was not in response to other retailers' actions. (Reporting by Karen Jacobs PERSON ; Editing COMPANY by Andre Grenon PERSON)

Conclusion

- ▶ Weak supervision is a powerful approach to train machine learning models in the absence of labelled data
- ▶ Allow us to inject *expert knowledge* (from heuristics, existing resources, etc.) into the model
- ▶ Demonstrated here for Named Entity Recognition, but can be applied to a wide range of problems

