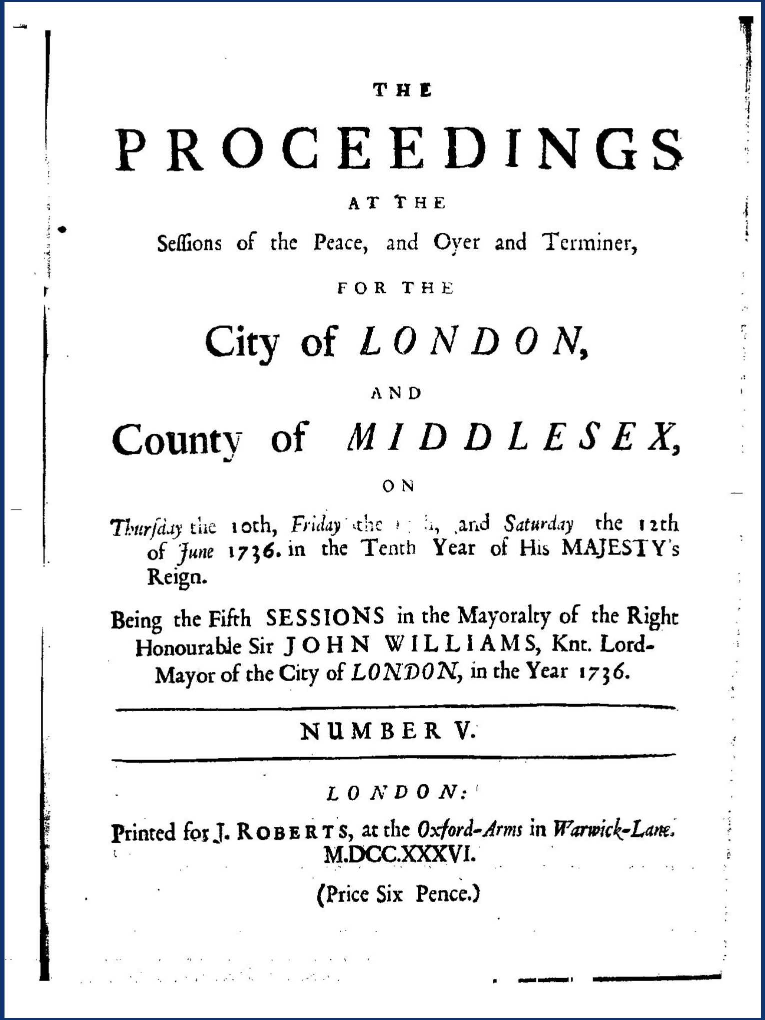
THE
DIGITAL
HUMANITIES
INSTITUTE.
SHEFFIELD

Jamie McLaughlin

Entity Tagging and Relation Extraction From Historical Texts Using Deep Learning

https://github.com/jjsmclaughlin/rsecon25



THE

PROCEEDINGS

AT THE

Seſſions of the Peace, and Oyer and Terminer,

FOR THE

City of *LONDON*,

AND

County of *MIDDLESEX*,

ON

Thurſday the 10th, *Friday* the 11th, and *Saturday* the 12th of *June* 1736. in the Tenth Year of His MAJESTY's Reign.

Being the Fifth SESSIONS in the Mayoralty of the Right Honourable Sir JOHN WILLIAMS, Knt. Lord-Mayor of the City of *LONDON*, in the Year 1736.

NUMBER V.

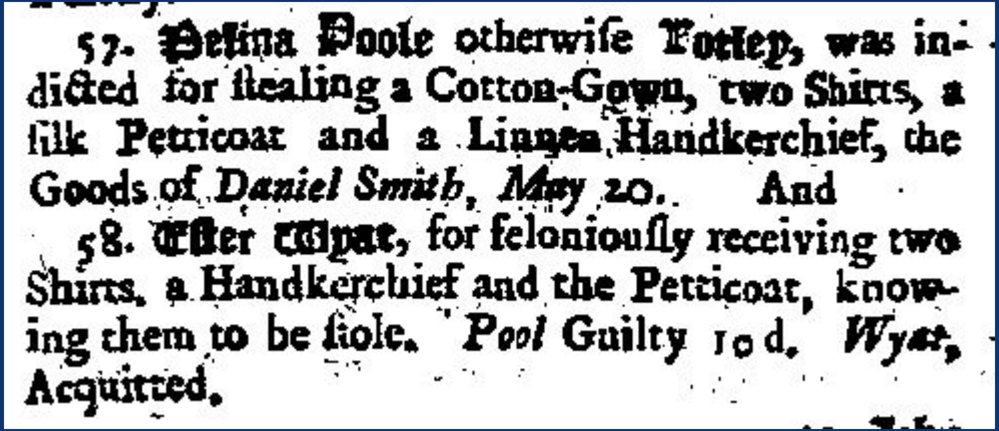
LONDON:
Printed for J. ROBERTS, at the *Oxford-Arms* in *Warwick-Lane*.
M.DCC.XXXVI.

(Price Six Pence.)

The Proceedings of the Old Bailey

Accounts of criminal trials.

- 1674 - 1911
- 197,752 trials
- ~ 127 Million words
- ~ 635 MB text



57. Delina Poole otherwiſe Foxley, was indicted for ſtealing a Cotton-Gown, two Shirts, a ſilk Petticoat and a Linnen Handkerchief, the Goods of *Daniel Smith*, *May* 20. And

58. Eſter Wyat, for feloniouſly receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be ſtole. *Pool* Guilty 10 d. *Wyat*, Acquitted.

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

DEFENDANT	Delina Poole otherwise Totley
DEFENDANT	Ester Wyat
VICTIM	Daniel Smith
THEFT	stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief
RECEIVING	feloniously receiving two Shirts, a Handkerchief and the Petticoat ..
GUILTY	Guilty
NOTGUILTY	Acquitted

Delina Poole	>>	**DEFOFF**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	>>	**DEFOFF**	>>	feloniously receiving two Shirts ..
Delina Poole	>>	**DEFVER**	>>	Guilty
Ester Wyat	>>	**DEFVER**	>>	Acquitted

spaCy

- Python NLP library
- Open Source
- First release 2016
- Added CNNs 2017
- Added transformers 2021
- Added LLMs (still beta) in 2023

- Provides document serialisation and annotation conventions = less “glue” code.
- Can be used with any back end library or technology.
- Built in scripts for training and evaluation (eg PRF scores).
- Has a built in component for Named Entity Recognition called EntityRecognizer.

Recognising defendants, victims and verdicts

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

DEFENDANT	Delina Poole otherwise Totley
DEFENDANT	Ester Wyat
VICTIM	Daniel Smith
GUILTY	Guilty
NOTGUILTY	Acquitted

This is what we want the EntityRecognizer to find in our unstructured plain text.

https://colab.research.google.com/drive/1nJ-nPbWfUPYWZCBVbis0IL9QvgSe9I2j

Result

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

DEFENDANT	Delina Poole otherwise Totley
DEFENDANT	Ester Wyat
VICTIM	Daniel Smith
GUILTY	Guilty
NOTGUILTY	Acquitted

For this trial, 100% success!

Why does this work so well?

- **Transition based**. Maintains a state machine as it parses the document sequentially and predicts the likelihood of the next token being the start or end of an entity.
- Because it predicts **transitions** in the document it can seem surprisingly context aware.
- It also copes well when the **length of an entity can vary**.
- The transition based approach means that the Entity Recognizer predicts **binary outcomes**. A single EntityRecognizer cannot predict overlapping entities. But you can just use more of them.

Recognising offence descriptions

Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

THEFT	stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief
RECEIVING	feloniously receiving two Shirts, a Handkerchief and the Petticoat ..

This is what we want the EntityRecognizer to find in our unstructured plain text now.

EntityRecognizer

“named real-world objects, like persons, companies or locations.”

“... If your entities are long and characterized by tokens in their middle, the component will likely not be a good fit for your task.”

The published specification of the EntityRecognizer makes it sound like it is **not a good fit** for this task.

Result

Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

GRANDLARCENY	stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen ..
GRANDLARCENY	feloniously receiving two Shirts, a Handkerchief and the Petticoat ..

The EntityRecognizer identifies the start and end tokens of our offence descriptions perfectly, but it gives them the wrong labels.

cri_lg_t2v_ner 3,211 Training Docs 921 Evaluation Docs (No maximum length)

cri_test_lg		45.82	43.98	44.88
cri_test_lg	GRANDLARCENY	42.69	94.30	58.78
cri_test_lg	PETTYLARCENY	0.00	0.00	0.00
cri_test_lg	THEFT	0.00	0.00	0.00
cri_test_lg	THEFTFROMPLACE	0.00	0.00	0.00
cri_test_lg	BURGLARY	45.83	57.89	51.16
cri_test_lg	MURDER	25.00	33.33	28.57
cri_test_lg	POCKETPICKING	50.00	80.77	61.76
cri_test_lg	RECEIVING	86.67	54.17	66.67
cri_test_lg	SHOPLIFTING	0.00	0.00	0.00
cri_test_lg	HIGHWAYROBBERY	67.65	82.14	74.19
cri_test_lg	ROBBERY	0.00	0.00	0.00
cri_test_lg	ANIMALTHEFT	0.00	0.00	0.00
cri_test_lg	FORGERY	0.00	0.00	0.00
cri_test_lg	HOUSEBREAKING	0.00	0.00	0.00
cri_test_lg	BIGAMY	0.00	0.00	0.00
cri_test_lg	PERJURY	0.00	0.00	0.00
cri_test_lg	FRAUD	0.00	0.00	0.00

Convolutional Neural Network (CNN)

- Good for local feature extraction. Older architecture.
- Should capture the contextual meaning of words to some extent, but over a **smaller distance** than a transformer.
- Can be trained using the **CPU** and RAM of a modest computer.

Transformer

- Good for long range dependency modelling and global context understanding. Newer architecture. Computationally expensive.
- Imports an existing, **pre-trained** transformer model from HuggingFace/transformers.
- Should be more aware of the **long range** contextual meaning of words (model is orders of magnitude larger and we know transformers scale better than CNNs)
- Massively more resource intensive than a CNN, in practice requiring a **GPU**. Can also require a lot of VRAM.

cri_tra_ner 2,289 Training Docs 663 Evaluation Docs (length restricted to 1,000 characters)

cri_test_lg		84.21	83.53	83.87	+38.99
cri_test_lg	GRANDLARCENY	88.34	91.14	89.72	+30.94
cri_test_lg	PETTYLARCENY	50.00	20.00	28.57	+28.57
cri_test_lg	THEFT	95.05	94.12	94.58	+94.58
cri_test_lg	THEFTFROMPLACE	84.00	89.36	86.60	+86.60
cri_test_lg	BURGLARY	78.95	78.95	78.95	+27.79
cri_test_lg	MURDER	16.67	16.67	16.67	-11.90
cri_test_lg	POCKETPICKING	85.19	88.46	86.79	+25.03
cri_test_lg	RECEIVING	81.82	75.00	78.26	+11.59
cri_test_lg	SHOPLIFTING	88.46	85.19	86.79	+86.79
cri_test_lg	HIGHWAYROBBERY	71.88	82.14	76.67	+2.48
cri_test_lg	ROBBERY	100.00	57.14	72.73	+72.73
cri_test_lg	ANIMALTHEFT	78.95	75.00	76.92	+76.92
cri_test_lg	FORGERY	66.67	50.00	57.14	+57.14
cri_test_lg	HOUSEBREAKING	75.00	60.00	66.67	+66.67
cri_test_lg	BIGAMY	66.67	66.67	66.67	+66.67
cri_test_lg	PERJURY	20.00	20.00	20.00	+20.00
cri_test_lg	FRAUD	60.00	100.00	75.00	+75.00

Relation Extraction

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

Delina Poole	>>	**DEFOFF**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	>>	**DEFOFF**	>>	feloniously receiving two Shirts ..

We want to train a model to **add relationship annotations** to the document, assigning defendants to the crime they committed, like our human transcribers did.

relation_extractor

Not a built in spaCy component. A userspace custom component built as a proof of concept by SpaCy engineers.

https://explosion.ai/blog/relation-extraction

“we take two entities previously predicted by a named entity recognizer and try to determine whether there is a semantic relationship between them and, if so, label it.”

Laura has bought a house in Boston	**LIVES**
Laura travelled through South America	**VISIT**
Laura met him in the evening before he flew to London	**UNRELATED**

The relation_extractor, despite being essentially a tutorial, sounds like a **good fit for our problem**.

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool
Guilty 10 d. Wyat, Acquitted.

Delina Poole	>>	**DEFOFF 0.588**	>>	stealing a Cotton-Gown, two Shirts ..
Delina Poole	>>	**DEFOFF 0.519**	>>	feloniously receiving two Shirts ..
Ester Wyat	>>	**DEFOFF 0.599**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	>>	**DEFOFF 0.530**	>>	feloniously receiving two Shirts ..

41, 42. William Lee, was indicted for stealing 16 lb. of Hog's Bristles, value 16 s. in the Parish of St. Magnus the Martyr, the Goods of William Thorp, April 5. And, Robert Davidson, for receiving the same, knowing them to be stole. Both Acquitted.

William Lee	>>	**DEFOFF 0.588**	>>	stealing
William Lee	>>	**DEFOFF 0.453**	>>	receiving the same ..
Robert Davidson	>>	**DEFOFF 0.586**	>>	stealing
Robert Davidson	>>	**DEFOFF 0.451**	>>	receiving the same ..

relation_extractor

.. Eleanor Williams was indicted for feloniously stealing ..

[[-0.42 ...], [1.93 ...] [3.84 ...], [2.59 ...] [3.35 ...], [-1.51 ...]]

[[0.87 ...] [2.14 ...]]

[[0.87 ... , 2.14 ...]
[2.14 ... , 0.87 ...]] = **DEFOFF**

The connecting words are **not** included in the tensor which is evaluated. The relation_extractor is only as context aware as the token embeddings **within the entities themselves**.

What we would really like would be a relation extractor which evaluated the words **between** the entities.

relation_extractor_context

.. Eleanor Williams was indicted for feloniously stealing ..

[[-0.42 ...], [1.93 ...] **[3.84 ...], [2.59 ...]** [3.35 ...], [-1.51 ...]]

[**[3.37 ...]**]

[**[3.37 ...]**] = **DEFOFF**

dcr_mu_t2v_rcx	145 Training Docs		50 Evaluation Docs	(multiple defendants)
dcr_test_mu	98.04	81.97	**89.29**	

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool
Guilty 10 d. Wyat, Acquitted.

Delina Poole	>>	**DEFOFF 0.999**	>>	stealing a Cotton-Gown, two Shirts ..
Delina Poole	>>	**DEFOFF 0.239**	>>	feloniously receiving two Shirts ..
Ester Wyat	>>	**DEFOFF 0.000**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	>>	**DEFOFF 0.999**	>>	feloniously receiving two Shirts ..

dcr_mu_tra_rcx 145 Training Docs 50 Evaluation Docs (multiple defendants)

dcr_test_mu 92.19 96.72 **94.40** **+ 5.11**

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool
Guilty 10 d. Wyat, Acquitted.

Delina Poole	>>	**DEFOFF 0.999**	>>	stealing a Cotton-Gown, two Shirts ..
Delina Poole	>>	**DEFOFF 0.000**	>>	feloniously receiving two Shirts ..
Ester Wyat	>>	**DEFOFF 0.000**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	>>	**DEFOFF 0.999**	>>	feloniously receiving two Shirts ..

DEFENDANT >> VERDICT

Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

George Butterfield, Edward Mould and Elizabeth Cook, of St. Ann's Westminster, were indicted for feloniously stealing 27 Saws the Property of several Persons, viz. 2 of John White 's, 5 of William Keys, 4 of William Anderson 's, 6 of Robert Raper 's, 6 of Anthony Sampson 's, and 5 of James Brody 's, the 20th of September last. To which Indictment George Butterfield pleaded Guilty ; but there not being sufficient Evidence against the two others, they were acquitted.

Assigning the correct verdict to the correct defendant requires **reasoning across the whole trial**. Or at least, comparing the first and final sections, with some ability to reason.

spacy-llm

Introduced in 2023. Still in beta. Allows you to use an LLM to perform NLP tasks within SpaCy driven by prompts.

https://github.com/explosion/spacy-llm

- Components implement only two functions:
  - **generate_prompts** takes a list of Doc objects and transforms them into a list of prompts.
  - **parse_responses** transforms the LLM outputs into annotations on the Doc.

- Supports both remotely hosted and locally hosted LLMs.

microsoft/phi-2 Quantised to 16 bit floats, will run on a 6 GB GPU.

“Instruct: You are an expert Natural Language Processing system. Your task is to extract structured information from the following legal case text.

For each defendant, output one line in the following format: [Defendant Name]: [Verdict]. Do not put any other text in your answer.

Text to analyze:

"57. Delina Poole otherwise Totley ..."

```
Delina Poole: Guilty
Ester Wyat: Acquitted
```

6. Mary Hughes, of St. Brides was indicted for stealing 2 Diaper Table-cloths, 2 Linnen Sheets, 2 Napkins, 2 Pewter Dishes, 6 Pewter Plates, 1 pair of Chints- Curtains, 1 double Cambrick Handkerchief, 6 Drinking Glasses, 2 China Chocolate-cups, 2 China Tea-cups, 2 Saucers, 3 Pounds of Candles, 1 pair of Laced double Ruffles, the Goods of Anthony Daffey, the 12th of February. And, 7, 8, 9. Grace Hughes, Edward Williams, and Diana, his Wife, were indicted for receiving part of the said Goods knowing them to be stolen. Mrs. Daffey. About Christmas last Mary Hughes came to me to be hired as a Servant ; she said, she had not been above a Fortnight in Town, but Mr. Williams the Prisoner, was a Relation of her's; and as he had liv'd in the Neighbourhood 14 Years, I thought I might depend upon him for her Character.

...

The Prisoner Mary owned the taking the Goods with a design to carry them off, and if I would go to Williams's House I should find them there; but Williams and his Wife told us, Grace Hughes had carry'd them all away in the Night. Mary said, she believ'd we might find her in Barnaby-Street; Williams and his Wife went to see for her, and about 10 o'Clock at night, they came to Mr. Daffey's House with Grace Hughes, and the Goods which Mrs. Daffey claim'd. Mrs. Daffey. Mary Hughes had left some Boxes, full of other things at Williams's House, but upon Mary's being taken up, Grace had pick'd out all that were mine, and carry'd them away in the Night. Mary Hughes had nothing material to say in her Defence: and Grace said, she expected Mr. Rawlinson of Hackney to appear to her Character; but no one appearing for either of them; the Jury found both the Hughe's Guilty. Several Persons giving Williams a fair Character, he and his Wife were Acquitted. [Transportation. See summary.]

1,141 words. 6,074 characters.

microsoft/phi-2 Quantised to 16 bit floats, will run on a 6 GB GPU.

...

”Defendants: Mary Hughes. Grace Hughes. Edward Williams.

Verdicts: Mary Hughes had nothing material to say in her Defence: and Grace said, she expected Mr. Rawlinson of Hackney to appear to her Character; but no one appearing for either of them; the Jury found both the Hughe's Guilty. Several Persons giving Williams a fair Character, he and his Wife were Acquitted. "

Mary Hughes: Guilty. Grace Hughes: Guilty. Edward Williams: Not Guilty.

dvr_llm_rel	microsoft/phi-2 with custom wrapper		
dvr_test_mu	~86.40	~82.09	**~84.19**

Workflow

- Create an initial gold standard corpus using **human annotators**. This is essential.
- Model Development:
  - An LLM can often produce helpful **zero-shot** annotations.
  - EntityRecognizer can produce useful results for some entity types after **only 100 - 200 examples**.
  - **Simple relation extraction** can be handled by relation_extractor_context after a few hundred examples.
  - Complex, distant relation extraction requires a **cleverer transformer solution**, or even an **LLM**.

EntityRecognizer

- Uses a CNN or a transformer to try to learn likely start and end points of entities.
- Intended to recognise: **“named real-world objects, like persons, companies or locations.”**
- Cannot identify overlapping entities
- Does not provide confidence scores.

https://spacy.io/api/entityrecognizer

Training, testing and evaluating a SpaCy pipeline

- **prodigy_to_docbin.py**
  - Process your documents into a SpaCy DocBins. Some QOL features are very useful here:
  - Splitting your data into train, dev and test sets.
  - Filtering documents by length.
  - Filtering documents by type and number of spans and relations.

- **spacy init config / spacy init fill-config**
  - Automatically create / complete a SpaCy config file by describing the pipeline.

- **spacy train**
  - Specify a **config file**, **training DocBin** and **dev DocBin**. Spacy runs the training.

- **test.py / test_rcx.py / test_llm.py**
  - Specify a **model** and a **test DocBin**. Sanity check the output.

- **spacy evaluate**
  - Specify a **model** and a **test DocBin**. Get PRF scores.

https://github.com/jjsmclaughlin/rsecon25

dvv_t2v_ner 2,314 Training Docs 669 Evaluation Docs (length restricted to 1000 characters)

dvv_test	94.24	93.57	**93.91**
DEFENDANT	95.72	97.19	**96.45**
VICTIM	92.72	90.44	**91.56**
GUILTY	97.35	96.49	**96.92**
NOTGUILTY	88.64	86.67	**87.64**

Precision What percentage of the annotations the model made were right?
Punishes wrong annotations.

Recall What percentage of the annotations in the example documents did the model find?
Punishes missed annotations.

F1 Score The harmonic mean of precision and recall.

dvv_t2v_ner 2,314 Training Docs 669 Evaluation Docs (length restricted to 1000 characters)

dvv_test		94.24	93.57	93.91	
dvv_test_lg		85.27	92.02	88.52	
dvv_test_xl		68.41	88.24	77.07	
dvv_test_xl	DEFENDANT	80.20	94.74	86.86	
dvv_test_xl	VICTIM	49.58	81.25	61.58	
dvv_test_xl	GUILTY	79.55	84.34	81.87	
dvv_test_xl	NOTGUILTY	84.85	91.80	88.19	

dvv_lg_t2v_ner 3,270 Training Docs 934 Evaluation Docs (No length restriction)

dvv_test		96.10	92.32	94.17	+ 0.26
dvv_test_lg		93.73	90.94	92.32	+ 3.80
dvv_test_xl		88.16	87.58	87.87	+ 10.80
dvv_test_xl	DEFENDANT	96.43	94.74	95.58	
dvv_test_xl	VICTIM	77.78	77.78	77.78	
dvv_test_xl	GUILTY	87.50	84.34	85.89	
dvv_test_xl	NOTGUILTY	90.62	95.08	92.80	

dvv_lg_t2v_ner 3,270 Training Docs 934 Evaluation Docs (No length restriction)

dvv_test		96.10	92.32	94.17
dvv_test_lg		93.73	90.94	92.32
dvv_test_xl		88.16	87.58	87.87
dvv_test_xl	DEFENDANT	96.43	94.74	95.58
dvv_test_xl	VICTIM	77.78	77.78	77.78
dvv_test_xl	GUILTY	87.50	84.34	85.89
dvv_test_xl	NOTGUILTY	90.62	95.08	92.80

dvv_md_tra_ner 3,174 Training Docs 902 Evaluation Docs (length restricted to 10,000 characters)

dvv_test		96.14	95.54	95.84	+ 1.67
dvv_test_lg		94.47	95.12	94.79	+ 2.47
dvv_test_xl		90.57	94.12	92.31	+ 4.44
dvv_test_xl	DEFENDANT	98.83	98.83	98.83	
dvv_test_xl	VICTIM	81.76	90.28	85.81	
dvv_test_xl	GUILTY	89.29	90.36	89.82	
dvv_test_xl	NOTGUILTY	92.06	95.08	93.55	

tok2vec

- Uses a **Convolutional Neural Network (CNN)**. Good for local feature extraction. Older architecture. Relatively computationally efficient.

- Learns word and subword embeddings from your training data. It is **essentially blank** when first initialised.

- Your pipeline components **directly train** the weights of the tok2vec layer.

- Should capture the contextual meaning of words to some extent, but over a **smaller distance** than a a transformer.

- Can be trained using the **CPU** and RAM of a modest computer.

transformer

- Uses a **Transformer** architecture. Good for long range dependency modelling and global context understanding. Newer architecture. Computationally expensive.

- Imports an existing, **pre-trained** transformer model from HuggingFace/transformers.

- Your pipeline components **fine tune** the transformer model and train lightweight neural layers built on top of the transformer outputs.

- Should be more aware of the **long range** contextual meaning of words (model is orders of magnitude larger and we know transformers scale better than CNNs)

- Massively more resource intensive than a CNN, in practice requiring a **GPU**. Can also require a lot of VRAM.

```
⚠️ Low number of examples for label 'INFANTICIDE' (12)
⚠️ Low number of examples for label 'FRAUD' (16)
⚠️ Low number of examples for label 'PETTYLARCENY' (17)
⚠️ Low number of examples for label 'EXTORTION' (10)
⚠️ Low number of examples for label 'FORGERY' (20)
⚠️ Low number of examples for label 'COININGOFFENCES' (12)
⚠️ Low number of examples for label 'RAPE' (21)
⚠️ Low number of examples for label 'BIGAMY' (44)
```

- During the debug stage it warned us that it **didn’t have enough examples** of some labels.

- In any case, distinguishing between some of these categories is **borderline expert knowledge**.

Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

THEFT	stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief
RECEIVING	feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to ..

With a transformer, the offence descriptions in our example trial are now annotated **100% correctly**.

EntityRecognizer

“named real-world objects, like persons, companies or locations.”

SpanCategorizer

“a wide variety of labeled spans, including long phrases, non-named entities, or overlapping annotations”

Our offence description task definitely matches the SpanCategorizer specification better. eg:

“stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief”

“feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole” crs_tra_spf 2,314 Training Docs 669 Evaluation Docs (length restricted to 1000 characters)

crs_test		88.37	21.59	34.70
crs_test	GRANDLARCENY	85.37	26.52	40.46
crs_test	PETTYLARCENY	100.00	20.00	33.33
crs_test	THEFT	100.00	22.68	36.97
crs_test	THEFTFROMPLACE	87.50	24.14	37.84
crs_test	BURGLARY	0.00	0.00	0.00
crs_test	MURDER	0.00	0.00	0.00
crs_test	POCKETPICKING	100.00	8.33	15.38
crs_test	RECEIVING	100.00	9.09	16.67
crs_test	SHOPLIFTING	81.82	42.86	56.25
crs_test	HIGHWAYROBBERY	0.00	0.00	0.00
crs_test	ROBBERY	0.00	0.00	0.00
crs_test	ANIMALTHEFT	0.00	0.00	0.00
crs_test	FORGERY	0.00	0.00	0.00
crs_test	HOUSEBREAKING	0.00	0.00	0.00
crs_test	BIGAMY	0.00	0.00	0.00
crs_test	PERJURY	0.00	0.00	0.00
crs_test	FRAUD	0.00	0.00	0.00

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

Nothing!

George Butterfield, Edward Mould and Elizabeth Cook, of St. Ann's Westminster, were indicted for feloniously stealing 27 Saws the Property of several Persons, viz. 2 of John White 's, 5 of William Keys, 4 of William Anderson 's, 6 of Robert Raper 's, 6 of Anthony Sampson 's, and 5 of James Brody 's, the 20th of September last. To which Indictment George Butterfield pleaded Guilty ; but there not being sufficient Evidence against the two others, they were acquitted. [Transportation. See summary.]

GRANDLARCENY feloniously stealing

2. William Caddy Francis, was indicted for stealing 4 lb. weight of Brass, value 3 s. the Goods of Thomas Ackland, August 27. Guilty 10 d. [Transportation. See summary.]

GRANDLARCENY stealing

EntityRecognizer

- **Transition based**. Maintains a state machine as it parses the document sequentially and predicts the likelihood of the next token being the start of an entity if it is currently outside one, or the end of the entity if it is currently inside one.

- The transition based approach means that the Entity Recognizer predicts **binary outcomes**. A single EntityRecognizer instance cannot predict overlapping entities. The SpaCy implementation also does not give confidence scores, although in principle it could give a confidence score for the transitions.

- Because it predicts **transitions** in the document it can seem surprisingly context aware.

- It also copes well when the **length of an entity can vary**.

SpanCategorizer

- **Suggester function** + **Labeler model**.

- Default Suggester function is completely naive and just suggests **all possible spans in the document of the preset lengths**. This is a huge problem for spans of wildly differing lengths, like our offence descriptions.

- Labeler model looks at the suggested spans in isolation and predicts the likelihood of each span belonging to each label. **It is only as context aware as the token embeddings within the span**.

- There is an experimental trainable Suggester function called **SpanFinder**. It tries to learn the tokens which tend to start and end spans. In practice it still suggests spans which are too short for our purposes.

- EntityRecognizer + SpanCategorizer could be a useful combination.

dcr_t2v_rel		2,243 Training Docs	655 Evaluation Docs		(length restricted to 1,000 characters)
	dcr_test_lg	93.42	99.13	**96.19**	
	dcr_test_mu	72.50	95.08	**82.27**	
dcr_lg_t2v_rel		3,150 Training Docs	908 Evaluation Docs		(no maximum length)
	dcr_test_lg	91.37	99.83	**95.41**	
	dcr_test_mu	66.30	100.00	**79.74**	
dcr_mu_t2v_rel		145 Training Docs	50 Evaluation Docs		(multiple defendants only)
	dcr_test_lg	91.24	100.00	**95.42**	
	dcr_test_mu	65.59	100.00	**79.22**	
dcr_tra_rel		2,243 Training Docs	655 Evaluation Docs		(length restricted to 1,000 characters)
	dcr_test_lg	93.86	98.78	**96.26**	
	dcr_test_mu	72.73	91.80	**81.16**	
dcr_md_tra_rel		2,502 Training Docs	740 Evaluation Docs		(length restricted to 10,000 characters)
	dcr_test_lg	92.23	99.48	**95.72**	
	dcr_test_mu	68.24	95.08	**79.45**	
dcr_mu_tra_rel		145 Training Docs	50 Evaluation Docs		(multiple defendants only)
	dcr_test_lg	94.75	97.56	**96.13**	
	dcr_test_mu	72.73	91.80	**81.16**	

Morgan Ellis, of St. Giles's in the Fields, was indicted for feloniously stealing a Pair of Sheets, value 5 s. the Goods of William Fowler, the 7th of May last. The Prosecutor depos'd, The Prisoner was his Lodger, and went away, and examining his Lodging after he was gone, the Sheets were missing; but there not being sufficient Proof that the Prisoner stole the Sheets, he was acquitted.

Morgan Ellis >> **DEFOFF 0.998** >> feloniously stealing a Pair of Sheets

21. Humphry Belmosset *, was indicted for assaulting Ann Metcalf on the Highway, and robbing her of a Necklace, and five Shillings. Acquitted. * Belmosset (by the Name of Benjamin Belmosset) was capitally Convicted in December, 1730.

Humphry Belmosset >> **DEFOFF 0.999** >> assaulting Ann Metcalf ..

It can certainly identify that a **DEFOFF** relationship should comprise a person and an offence description.

dcr_t2v_rel 2,243 Training Docs 655 Evaluation Docs (length restricted to 1,000 characters)

dcr_test_lg 93.42 99.13 **96.19**

dcr_lg_t2v_rel 3,150 Training Docs 908 Evaluation Docs (no maximum length)

dcr_test_lg 91.37 99.83 **95.41** **- 0.78**

dcr_tra_rel 2,243 Training Docs 655 Evaluation Docs (length restricted to 1,000 characters)

dcr_test_lg 93.86 98.78 **96.26** **+ 0.07**

dcr_md_tra_rel 2,502 Training Docs 740 Evaluation Docs (length restricted to 10,000 characters)

dcr_test_lg 92.23 99.48 **95.72** **- 0.54**

dcr_t2v_rel	2,243 Training Docs	655 Evaluation Docs		(length restricted to 1,000 characters)
dcr_test_mu	**72.50**	95.08	**82.27**	
dcr_lg_t2v_rel	3,150 Training Docs	908 Evaluation Docs		(no maximum length)
dcr_test_mu	**66.30**	100.00	**79.74**	
dcr_tra_rel	2,243 Training Docs	655 Evaluation Docs		(length restricted to 1,000 characters)
dcr_test_mu	**72.73**	91.80	**81.16**	
dcr_md_tra_rel	2,502 Training Docs	740 Evaluation Docs		(length restricted to 10,000 characters)
dcr_test_mu	**68.24**	95.08	**79.45**	

DEFENDANT >> OFFENCE

Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole.

The connecting words are **not** included in the tensor which is evaluated. The relation_extractor is only as context aware as the token embeddings **within the entities themselves**.

What we would really like would be a relation extractor which evaluated the words **between** the entities.

dcr_mu_t2v_rel	145 Training Docs	50 Evaluation Docs		(multiple defendants)
dcr_test_lg	91.24	100.00	**95.42**	
dcr_test_mu	**65.59**	100.00	**79.22**	
dcr_mu_tra_rel	145 Training Docs	50 Evaluation Docs		(multiple defendants)
dcr_test_lg	94.75	97.56	**96.13**	
dcr_test_mu	**72.73**	91.80	**81.16**	
dcr_mx_tra_rel	410 Training Docs	130 Evaluation Docs		(multiple defendants, larger corpus)
dcr_test_lg	93.83	92.84	**93.33**	
dcr_test_mu	**75.34**	90.16	**82.09**	

57. Delina Poole otherwise Totley, was indicted for stealing a Cotton-Gown, two Shirts, a silk Petticoat and a Linnen Handkerchief, the Goods of Daniel Smith, May 20. And 58. Ester Wyat, for feloniously receiving two Shirts, a Handkerchief and the Petticoat, knowing them to be stole. Pool Guilty 10 d. Wyat, Acquitted.

Delina Poole	**>>**	**DEFOFF**	>>	stealing a Cotton-Gown, two Shirts ..
Ester Wyat	**>>**	**DEFOFF**	>>	feloniously receiving two Shirts ..
Delina Poole	**>>**	**NO_REL**	>>	feloniously receiving two Shirts ..
Ester Wyat	**>>**	**NO_REL**	>>	stealing a Cotton-Gown, two Shirts ..

dcr_mn_tra_rel 410 Training Docs 130 Evaluation Docs (multiples, larger corpus, negation)

dcr_test_mn **60.07** **69.51** **64.44**

Mary Hughes: Guilty. Grace Hughes: Guilty. Edward Williams: Not Guilty.

Mary Hughes	>>	**DEFVER**	>>	Guilty
Grace Hughes	>>	**DEFVER**	>>	Guilty
Edward Williams	>>	**DEFVER**	>>	Acquitted