Machine Learning for Document Layout Analysis

Animesh Prasad

Hervé Déjean
Jean-Luc Meunier

June 2018

The Problem: Information Extraction from Marriage Records

A Record

Dataset: Esposalles

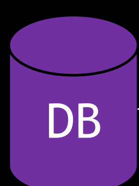
- Marriage license book conserved at the Archives of the Cathedral of Barcelona, written in old Catalan by only one writer in the 17th century

husband> fille de <husband's father> y <husband's mother> ab <wife> fille de <wife's father> y <wife's mother>

<husband> fille de <husband's father> y <husband's mother> ab <wife> viusa <wife's former husband>

The Problem: Information Extraction from Marriage Records

Track 1: Basic: CSV file	Track 2: Complete: CSV file
Antoni, name	Antoni, name, husband
Duran, surname	Duran, surname, husband
pages, occupation	pages, occupation, husband
Regne, location	Regne, location, husband
de, location	de, location, husband
fransa, location	fransa, location, husband
Bara, location	Bara, location, husband
Elisabeth, name	Elisabeth, name, wife
Juana, name	Juana, name, wife
donsella, state	donsella, state, wife
Bernat, name	Bernat, name, wifes_father
Prats, surname	Prats, surname, wifes_father
forner, occupation	forner, occupation, wifes_father
Bara, location	Bara, location, wifes_father
Juana, name	Juana, name, wifes_mother



Dataset: Esposalles

- Gold: 774 Records with Handwritten Transcriptions (~1K unique Tokens)
- Input: HTR performed by CITLab with fairly high accuracy (~5%CER)
- Output: CSV with transcription and predetermined classes
  - CAT: name, surname, occupation, etc
  - PER: husband, wife, wife's father, etc
- Evaluation:
  - Basic (Task1 only)
  - Complete (Task1 and Task2)

Approach: Modeling

Graph-based approach

- One Record : one graph
- Features: n-gram [2-4] (~3K)/ CE (Dim 10)
- Nodes: tokens
- Edges: positional encoding
- Machine learning algorithms
  - (((Node Type)? Edge-Feature)? Graph?) Conditional Random Fields
  - ((CE?)CRF?)LSTM

Graph Conditional Random Field

A CRF is an (undirected) graph:

The x_i are the nodes

The y_i are the node labels

An edge between two nodes indicates that their labels have a dependency

$$\sum_{i \epsilon V} W_{y_i}^T . node\_feature(x_i) \;+\; \sum_{(i,j) \epsilon E} W_{y_i, y_j}^T . edge\_feature(x_i, x_j)$$

Graph Conditional Random Field

A CRF is an (undirected) graph:

- The x_i are the nodes
- The y_i are the node labels
- An edge between two nodes indicates that their labels have a dependency

$$\sum_{i \epsilon V} W_{y_i}^T . node\_feature(x_i) + \sum_{(i,j) \epsilon E} W_{y_i, y_j}^T . edge\_feature(x_i, x_j)$$

Graph Conditional Random Field

A CRF is an (undirected) graph:

- The x_i are the nodes
- The y_i are the node labels
- An edge between two nodes indicates that their labels have a dependency

$$\sum_{i \epsilon V} W_{y_i}^T . node\_feature(x_i) + \sum_{(i,j) \epsilon E} W_{y_i, y_j}^T . edge\_feature(x_i, x_j)$$

- No Feature (always 1)
- Positional one-hot

Dev-Results (10Fold-CV)

Train/Test	Task 1							Task 2								
Class	Name	Surname	State	Location	Occupation	*Other*#	F_1	Husband	Husband's Father	Husband's Mother	Other Person	Wife	Wife's Father	Wife's Mother	*Other*#	F_1
CCRF																
True/True	0.9831	0.9534	0.9668	0.9546	0.9826	0.9851	**0.9844**	0.8089	0.5733	0.2959	0.6704	0.1992	0.6621	0.1119	0.9569	**0.7333**
True/Noisy	0.9921	0.9742	0.9794	0.9703	0.9854	0.9887	**0.9759**	0.8367	0.653	0.4575	0.7234	0.2942	0.7014	0.3533	0.9581	**0.7797**
GCRF																
True/True	0.9384	0.8523	0.8979	0.835	0.9721	0.9534	**0.9218**	0.8759	0.8668	0.912	0.9244	0.7586	0.8739	0.6793	0.9536	**0.9072**
True/Noisy	0.9554	0.895	0.9174	0.8536	0.9816	0.9568	**0.9345**	0.9116	0.8894	0.9363	0.9479	0.8289	0.92	0.815	0.9557	**0.9260**
EFGCRF																
True/True	0.9876	0.9611	0.9699	0.9638	0.9844	0.9868	**0.9798**	0.9334	0.922	0.9403	0.9607	0.9032	0.945	0.9008	0.9595	**0.9430**
True/Noisy	0.9937	0.9758	0.982	0.9731	0.9864	0.9896	**0.9859**	0.9469	0.9341	0.9521	0.981	0.9352	0.957	0.9562	0.9623	**0.9529**
BLSTM																
True/True	0.9914	0.9667	0.9795	0.9757	0.987	0.9893	**0.9848**	0.9714	0.9864	0.9841	0.9874	0.9857	0.9832	0.9672	0.9873	**0.9849**
True/Noisy	0.9984	0.9935	0.996	0.9905	0.9916	0.9964	**0.9954**	0.9912	0.9951	0.9958	0.9966	0.9944	0.994	0.9957	0.9959	**0.9952**

TABLE I

CLASS-WISE F_1 SCORES FOR BOTH SEMANTIC CATEGORIES (ONLY BEST PERFORMING MODELS OR MODELS WITH INTERESTING OBSERVATIONS ARE SHOWN, # CLASSES ARE NOT CONSIDERED FOR IHHER EVALUATION)

Dev-Results: Reading Between the Lines

CRF, GCRF, EFGCRF

LSTM

NTEFGCRF

Multitasking (in general on all model types)

CRF-LSTM

CE-(LSTM/CNN)-(CRF?)-LSTM

PRIVATE method: Naver Labs 2018-06-25

Authors: **Naver Labs**

expires on 2018-07-25

edit delete

method: CITlab ARGUS (with OOV) 2017-07-09

Authors: **Tobias Strauß, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning, Roger Labahn**

Description: **The training data is divided into a training set (2790 line images) and a validation set (280 line images). Several normalization methods such as contrast, size, slant and skew normalization are applied. These preprocessed line images serve as input for the optical model, a recurrent neural network (layer from input to output: conv, conv, lstm (256 cells), conv, lstm (512 cells)) trained by CTC (150 epochs of 5000 noisy line images each). To enlarge input variety, the line images we use data argumentation on line images.**
The output of the optical model are probabilities for each character at each position in the image collected in a matrix. The various output matrices for one record (which represent the lines) are glued together to one single matrix. We define regular expressions to extract the required information from this matrix. This is done in two steps: First, we segment the matrix into regions of interest: regions containing information about the husband, the husbands parents, the wife or the wife's parents. These regions are matched against a valid combination of dictionary items in a second step. For the name fields additional OOV words are allowed if the dictionary items do not fit.

method: CITlab ARGUS (with OOV, net2) 2017-07-10

Authors: **Tobias Strauß, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning, Roger Labahn**

Description: **The training data is divided into a training set (2790 line images) and a validation set (280 line images). Several normalization methods such as contrast, size, slant and skew normalization are applied. These preprocessed line images serve as input for the optical model, a recurrent neural network (layer from input to output: conv, conv, blstm (512), conv, blstm (512 cells), blstm (512 cells)) trained by CTC (150 epochs of 5000 noisy line images each). To enlarge input variety, the line images we use data argumentation on line images.**
The output of the optical model are probabilities for each character at each position in the image collected in a matrix. The various output matrices for one record (which represent the lines) are glued together to one single matrix. We define regular expressions to extract the required information from this matrix. This is done in two steps: First, we segment the matrix into regions of interest: regions containing information about the husband, the husbands parents, the wife or the wife's parents. These regions are matched against a valid combination of dictionary items in a second step. For the name fields additional OOV words are allowed if the dictionary items do not fit.

Blind Test Results

Ranking Table

Description Paper Source Code

Date	Method	Basic Score	Complete Score	Name	Surname	Location	Occupation	State
2018-06-25	Naver Labs	95.46%	95.03%	97.01%	92.73%	95.03%	96.43%	96.41%
2017-07-09	CITlab ARGUS (with OOV)	91.94%	91.58%	95.14%	85.78%	88.43%	93.08%	97.54%
2017-07-10	CITlab ARGUS (with OOV, net2)	91.63%	91.19%	95.09%	85.84%	87.32%	92.96%	97.19%
2017-07-09	CITlab ARGUS (without OOV)	89.54%	89.17%	94.37%	76.54%	87.65%	92.66%	97.43%
2017-07-01	Baseline HMM	80.28%	63.11%	81.06%	60.15%	78.90%	90.23%	93.79%

Takeaways

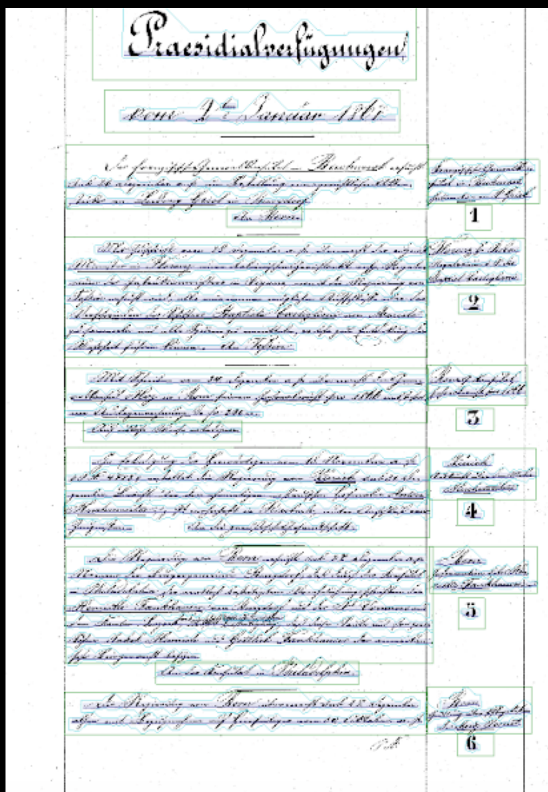
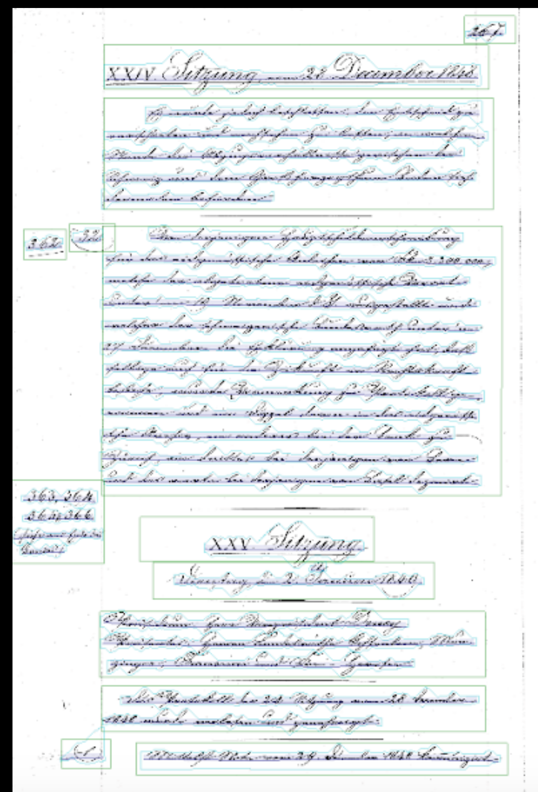
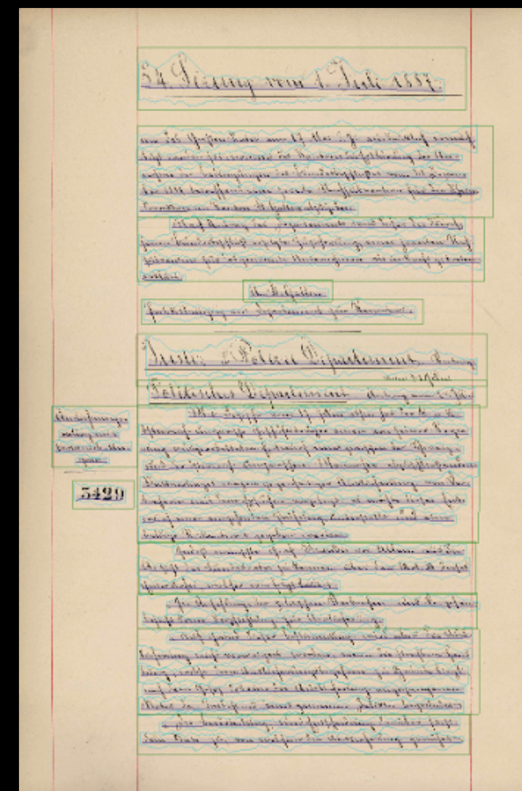
- Fine grain semantic class is difficult to classify for graphical models. Increasing the model complexity in graphical models while keeping the surface features from the tokens same increases the performance

- Due to small CER on the HTR system the models trained with the noisy data as well perform near the gold data standards.

- Structured surface level strings with anchors are easy to parse using ML models as compared to RegEx match and end-to-end models as well. (Near perfect Result 95% given 5%CER; ~3% gain)

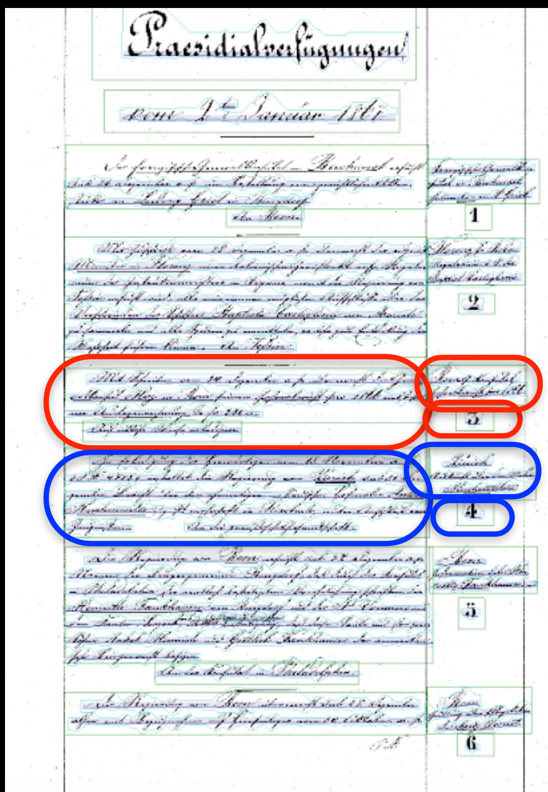
Moving On….

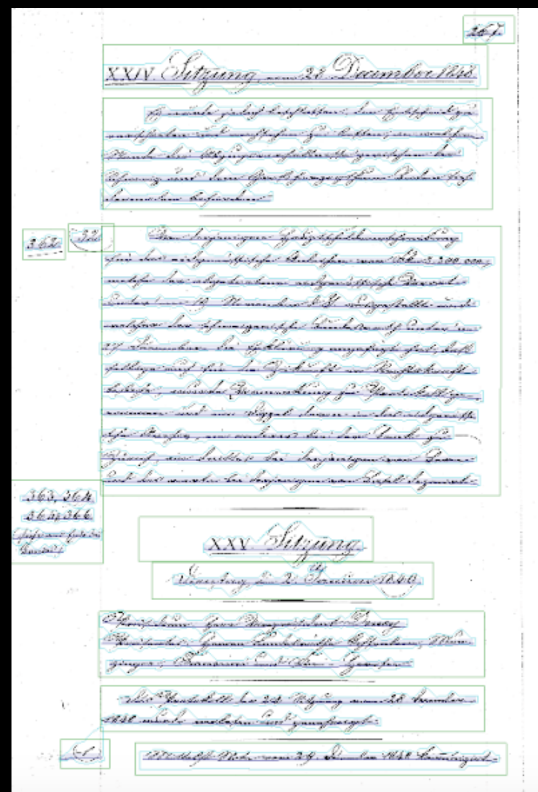
The Problem: BAR “Bundesratsprotokolle”

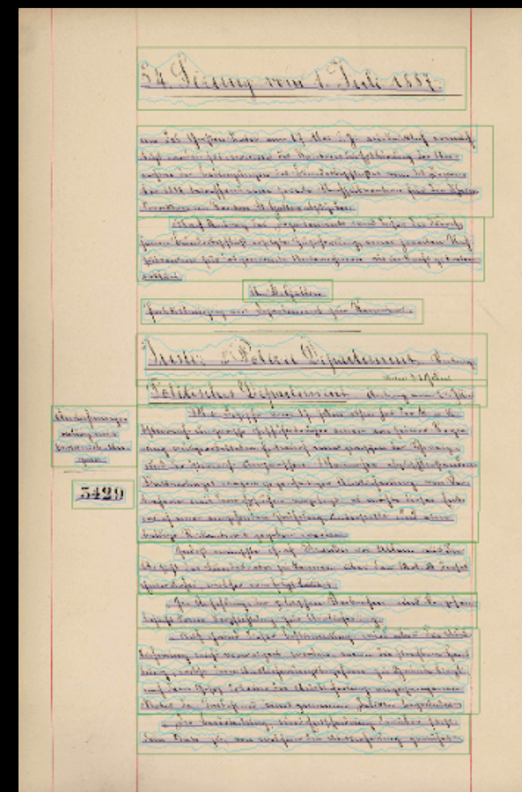




Some Pages

The Problem: BAR “Bundesratsprotokolle”

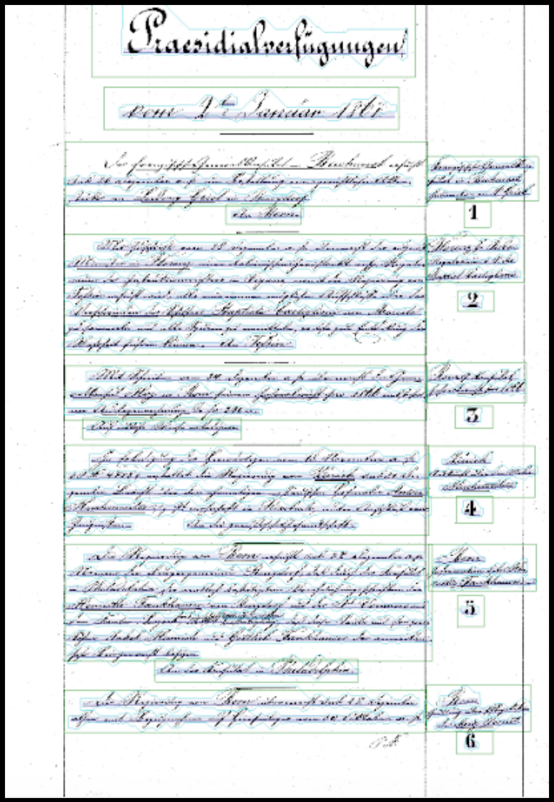


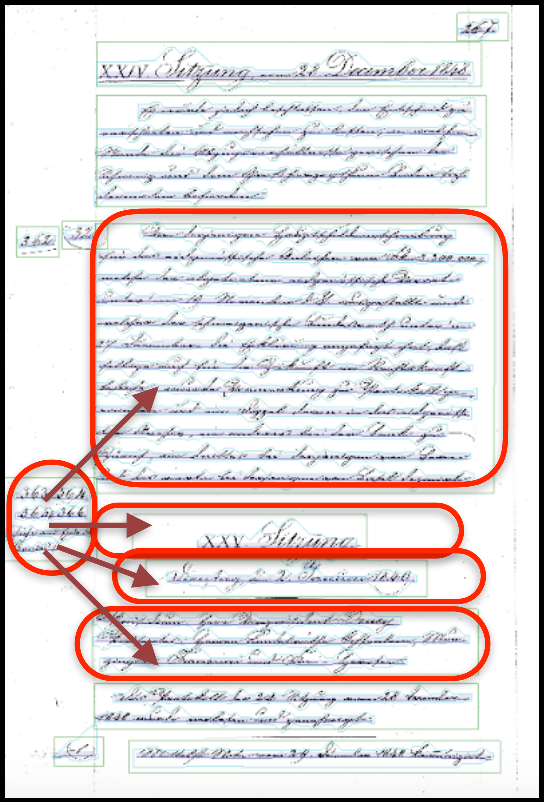


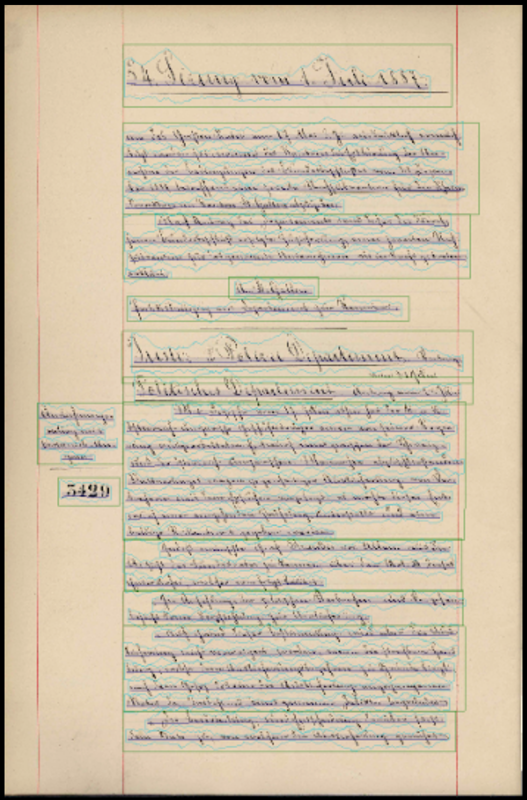


Some Pages

The Problem: BAR “Bundesratsprotokolle”





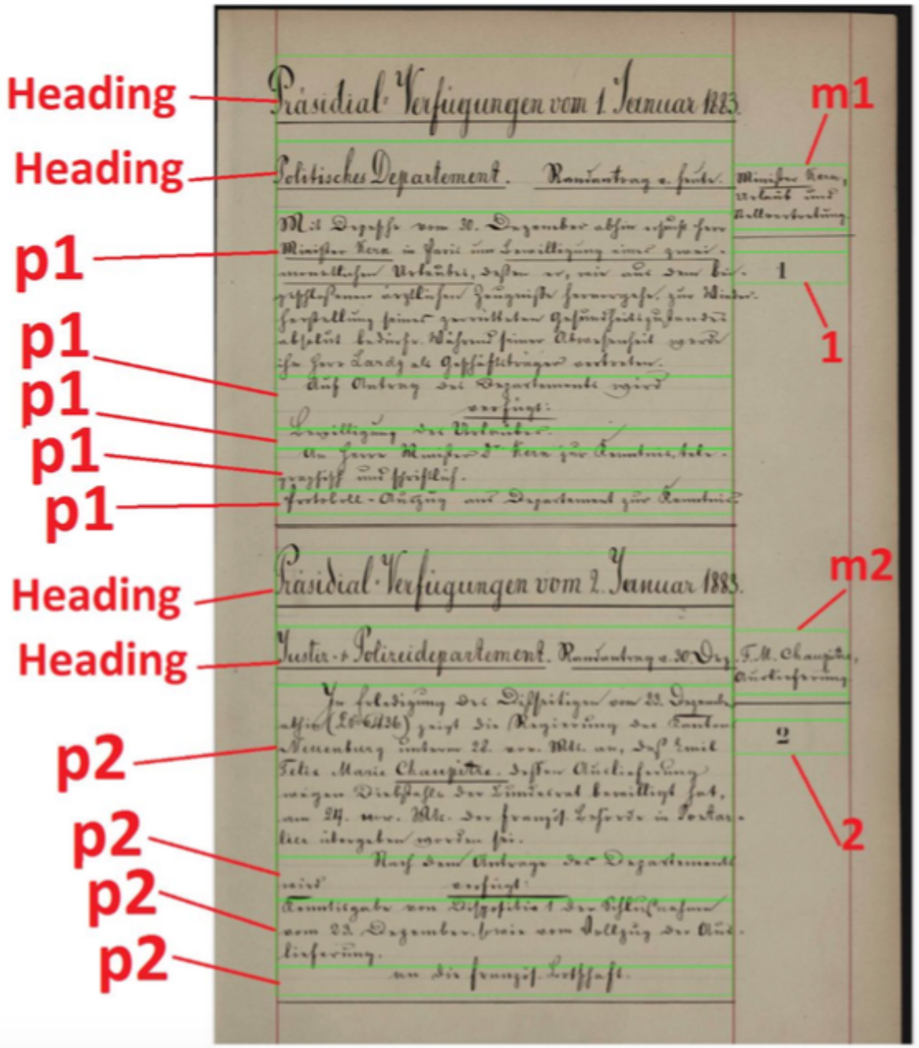


Some Pages

The Problem: BAR “Bundesratsprotokolle”

Heading and Resolution number, marginalia, paragraphs

(page 1 of doc 13685 in collection 4583)



Why a problem?

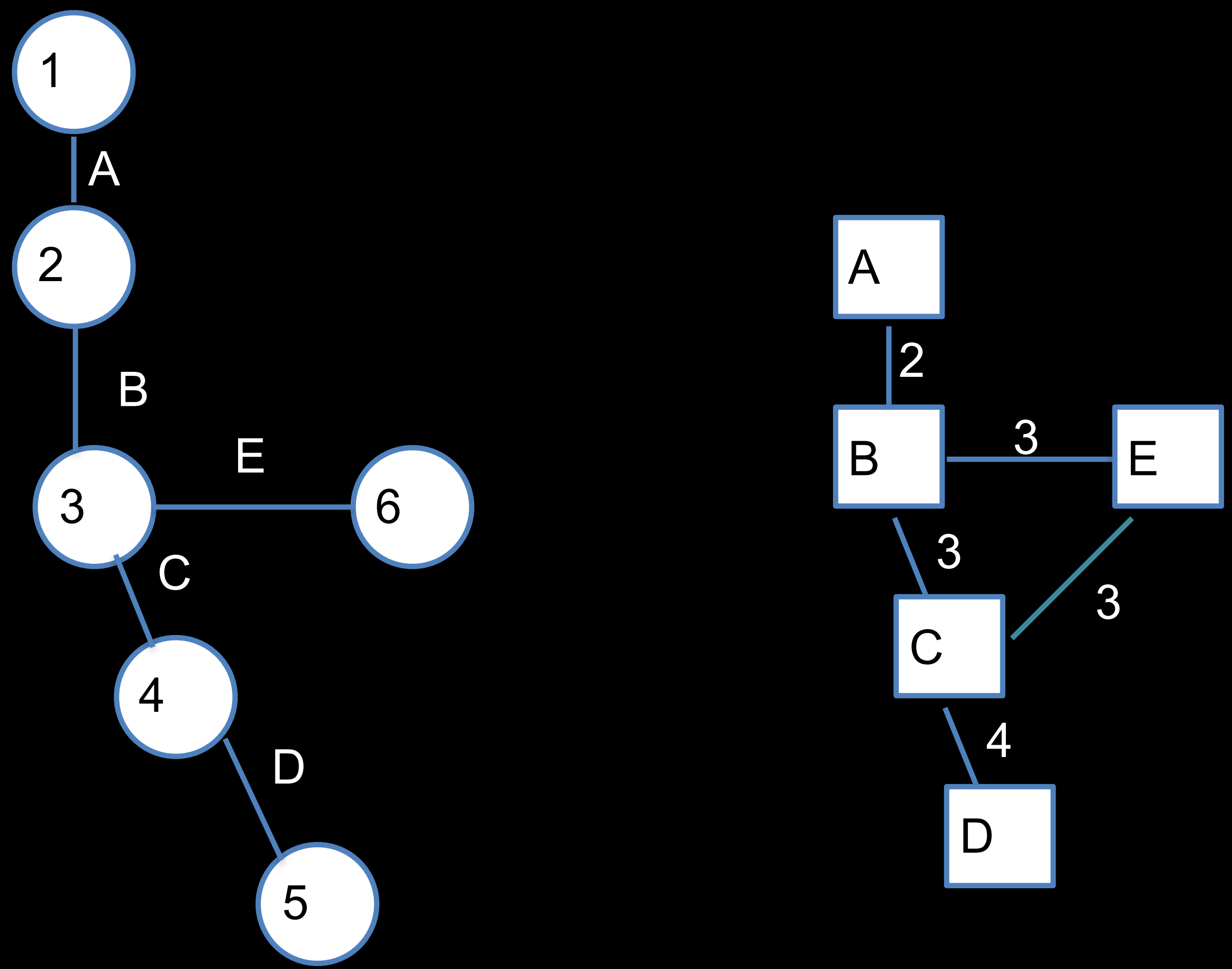
For a collection of Documents like BAR

- Horizontal line is not strictly the delimiter (unlike ABP - Table Understanding)

It requires a knowledge of what is a Segment

- Model the problem as learning that knowledge
- Can utilise the semantic tags here

Input: Line Dual Graph



Edge Conv Nets (ECN)

DAS 2018 Idea: learn graph convolutions which depends on edge features

- A convolution computes a scalar for each edge
- i.e a parametrized adjacency matrix: E_{ij}

$$h_i^{l+1} = \sigma\Big([\ \sum_{j \in N_i} E_{ij}^l W^l h_j,\ \ W^l h_i\ ]\Big)$$

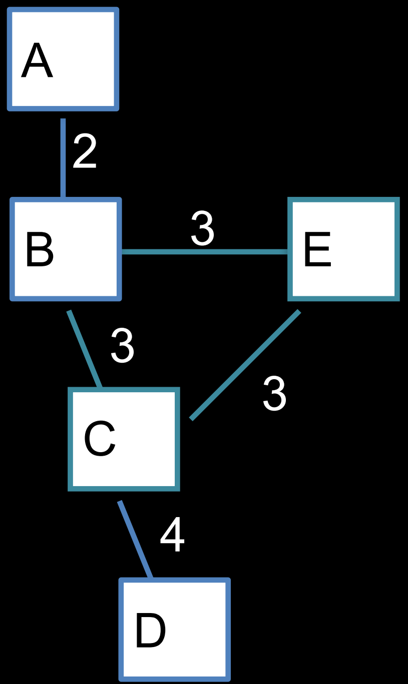
Stack/Adding them

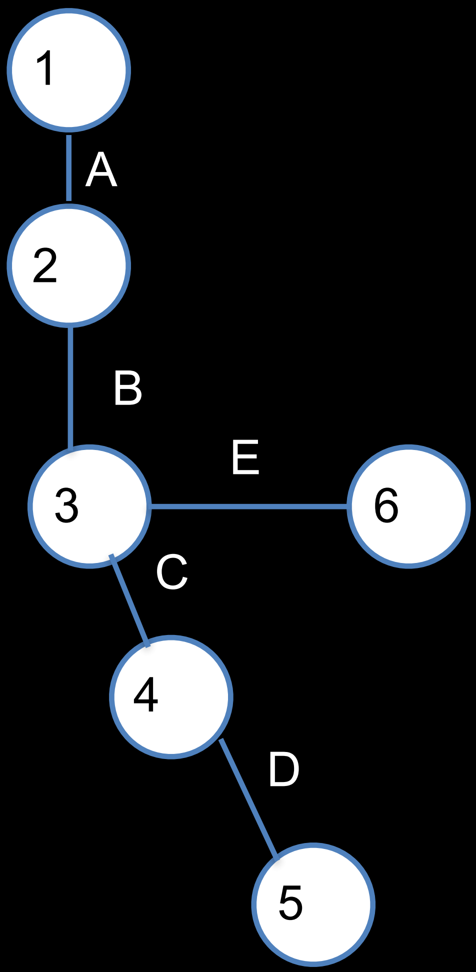
several convolutions per layer

Similar Idea in Graph Attention Network, ICLR'18

Modeling: Retrieving Segments as Clusters

- Score [0,1]   Use score to form cluster
  - Many pure/heuristic approaches
    - Threshold Based selection (T>0.5)





Results

Edge Classification (with textual features, positional features, relative positional and other geographical features)

Feature	Model	Accuracy (3Fold-CV)	
TestUnit		TextRegion	TextLine
F	CRF*(==1)	0.89	0.91
	ECN(>0.5)	0.88	0.90
	SVM(==1)	0.70	0.73
	LR(>0.5)	0.81	0.82
F+S	CRF(==1)	0.92	0.92
	ECN(>0.5)	0.90	0.91
	SVM(==1)	0.72	0.74
	LR(>0.5)	0.82	0.83

Results

Segmentation @ F1 (CRF)

Seg Type/F1	0.6	0.7	0.8	0.9	1.0
with 'IGNORE'					
TextUnit (Micro)	0.82	0.74	0.65	0.58	0.47
Resolution(Macro)	0.83	0.75	0.67	0.58	0.49
without 'IGNORE'					
TextUnit (Micro)	0.79	0.72	0.65	0.56	0.45
Resolution(Macro)	0.76	0.69	0.63	0.52	0.44

*Comparison with other EC models reflect direct correlation with edge classification

**Comparison with complete different approach (1/0 segmentation ~59% Micro F1)

Devdcäteisüeg voem 29. Januar 1855.

ine vom 8. Januar 2. 7. N. J. erletigt. nach Vormerlung

Aa aa gewiesen.

h.

der Dcuin candeill über maicht unterm 25. 1. ein
vom Pedlelten dor Btrnommmndungn ihm zugewiesenes Aus-
Kurftseehren der HH. Dek und oldo2s betreffeit einen
and laids in Anweille und gibt zugleich edie gewünschte Aus-
Kunst indem ir heineheus, aauf die Ainstatthaftigkeit
aufneersam macht, daß Schweizer sich mit solchen Begehren-
an die Lndesbehöele wenden, währendt in Konsulat zur
Wahenesung ihrer Interessen auf dem Pflaze bestehe

n Aunat

An den Konsut Mutheitung der Beilage

MadolteLonct
Daus etiacud.
Aassmift

N.

3.) Zuschrift vem 11. 25. 1. erstattert des Konsut in
Diga Bericht über dors Ergebus der nach Auftrag vom 30. 5.
s75f pv4d eingezrogenen Eptundigeungen betreffent die
Nachlaßverhältüisse des in Anitan verstorrbenen Rmdolf Frsse-
vonlikon - Dyukau, im Wesentlichem dahin gehed, daß
das erhaanteneDetizthum Banum zurLetung der darauf
haftenten Sehalden genigen Türfte

A. nrane

Riga Nrenl
Rndolf Loptor.
Nacolaßratältniß.

E.

Di Regierung von Shniyzersucht unterm No. 3.
um Einziehung von Erfändigungen übber einen Peit 1855
Verschollenen C.Mlcsues-Henyon rtte, der sich noch
im Juni 1855 zu Weiten un eheimpreussischenFreis
Masegbanden aufgehalten habe

Andie Gesandtschaftan Salin z. 6

kog-
c.Militior Aunna-
Mumudigung
829.

Bo.Wri
frer Bezunnahnme auf die Eisladung vom
Ingeho- Motinan.

Zoom To Fit 1 of 23 23

0

Devdcäteisüeg vnem 29. Januar 1855

ine vom 8. Januar 2. 7. N. J. erletigt. nach Vormerlung

Aa aa gewiesen

h

der Dcuin candeill über maicht unterm 25. 1. ein MadolteLonct

vom Pedlelten dor Btrnommmndungn ihm zugewiesenes Aus- Daus etiacud

Kurftseehren der HH. Dek und oldo2s betreffeit einen Aassmift

and laids in Anweille und giht zugleich edie gewünschte Aus-

Kunst indem ir heineheus, aauf die Ainstatthaftigkeit N 17

aufneersam macht, daß Schweizer sich mit solchen Begehren-

an die Lndesbehöele wenden, währendt in Konsulat zur

Wahenesung ihrer Interessen auf dem Pflaze bestehe

n Aunat

13 An den Konsut Mutheitung der Beilage

3.) Zuschrift vem 11. 25. 1. erstattert des Konsut in Riga Nrenl

Diga Bericht über dors Ergebus der nach Auftrag vom 30. 5. Rndolf Lontor

s75f pv4d eingezrogenen Eptundigeungen betreffent die NacolaßratältniB

Nachlaßverhältüisse des in Anitan verstorrbenen Rmdolf Frsse-

vonlikon- Dyukau, im Wesentlichem dahin gehed, daß E 29

das erhaanteneDetizthum Banum zurLetung der darauf

haftenten Sehalden genigen Türfte

25 A. nrane

Di Regiegung von Shniyzersucht unterm No. 3. kog-

um Einziehung von Erfändigungen übber einen Peit 1855 c.Militior Auma-

Verschollenen C. Mlcsues-Henyon rtte, der sich noch Mumudigung

im Juni 1855 zu Weiten un eheimpreussischenFreis

Masegbanden aufgehalten habe 829. 39

35 Andie Gesandtschaftan Salin z. 6

40 frer Bezunnahme auf die Eisladung vom Bo Wri

Iugeho- Motinan

Predicted:
0: [0],
1: [1, 2],
2: [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17],
3: [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29],
4: [35, 36, 30, 31],
5: [32, 33, 34, 37, 38, 39],
6: [40, 41, 42, 43, 44, 45, 46]

True:
'bar_576': [1, 2],
'bar_577': [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17],
'bar_578': [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29],
'bar_579': [30, 31, 32, 33, 34, 35, 36, 37, 38, 39],
'bar_580': [40, 41, 42, 43, 44, 45, 46],
'bar_IGNORE': [0]

Takeaway

Edge Classification followed by clustering can result into segmentation using document agnostic clustering

Thank you

https://es.naverlabs.com/animesh-prasad