Lab #4: Table Reconstruction

EE447 Mobile Network, Luoyi Fu, Spring 2021

Due: *Sunday, June 13th*

Name: Hongjie Fang Student ID:518030910150 Email: galaxies@sjtu.edu.cn

1 Purpose and Objective

Nowadays, PDF documents have become the mainstream document format because of its unique cross-platform convenience advantage. PDF documents contain a large amount of valuable data information, and the table is one of the important carriers of these data. However, the structure of PDF documents is complex, and it is difficult for us to obtain accurate table information directly from the document format. Therefore, for PDF tables, we need to reconstruct the structure of the table, so as to achieve the extraction of the table.

PDF documents are often used in academic paper writing, therefore in this lab, we mainly focus on the table reconstruction in academic papers. The table reconstruction may help a lot in information extraction from PDFs, especially structured data extraction.

This Lab focuses on the table line reconstruction for the tables without frame lines. In this Lab, we are required to use python to complete table line drawing of a specific table without frame lines and show the results. After that, there are three questions:

1. How to automatically locate the tables in a PDF?
2. What do you think is the most difficult step to extract the table from the PDF? Why?
3. **(Bonus)** How to accurately identify the header of the table, and use natural language processing (NLP) or other methods to understand the information in the table, and then extract the entities and relationships from table to construct a specific knowledge graph?

2 Frameline-drawing Algorithm

In this section, we will explain the algorithm of completing table frameline drawing in given code.

- We first convert the image to gray-scale image, which will make things easier since we do not need to worry about the color of the word in the table. Then, we extract the edges of the table, which is used to detect the partial-drawing lines in the table.
- After that, we erase the partial-drawing lines from the table, and detect the boundary of the table (up-left point and down-right point). Then, we can remove the bounding lines of the table. Then, we erode the image slightly and perform binaryzation for convenience.
- Next, we detect the horizontal lines, if we have sacnned a all-white line, then it may be the boundary between two lines in the table. Therefore we can record it in the list. If we find several continuous "potential boundary", we can combine them into one using the average value of the y coordinates.
- Similarly, we transpose the image and then perform the same operation again to detect the vertical lines.
- Finally, we combine the horizontal lines and vertical lines, and draw these lines in the picture.

We implement the algorithm based on the given code in python. Fig. 1, Fig. 2, Fig. 3, Fig. 4 and Fig. 5 are the results.

Table 2
Summary of structural rock fabrics from metamorphic rocks on Hall Peninsula, Baffin Island.

Fabric		Mean orientation[*]
D_1: E-W crustal shortening		
F_{1a}	isoclinal folds	AP: 159°/64°; FA: 19°–335°
S_{1a}	metamorphic foliation axial planar to F_{1a}	164°/63°
F_{1b}	isoclinal to open folds of S_{1a}	AP: 159°/64°; FA: 19°–335°
S_{1b}	metamorphic foliation axial planar to F_{1b}	164°/63°
$L_{1a,b}$	elongate metamorphic mineral growth	parallel to $F_{1a,b}$ fold hinges, or down-dip
D_2: E-W crustal shortening		
T_2	thick-skinned reverse faults	SSE-striking, NNE-dipping
L_2	mineral stretching and elongate growth	30°–265°
F_2	thick-skinned folds, E-vergent	AP: SSE-striking, NNE-dipping; FA: subhorizontal, trending SSE or NNW
D_3: N-S crustal shortening		
F_3	thick-skinned folds; crenulations	AP: 269°/51°; FA: 37°–269°
S_3	crenulation cleavage axial planar to F_3	269°/51°

Table 2
Summary of structural rock fabrics from metamorphic rocks on Hall Peninsula, Baffin Island.

Fabric		Mean orientation[*]
D_1: E-W crustal shortening		
F_{1a}	isoclinal folds	AP: 159°/64°; FA: 19°–335°
S_{1a}	metamorphic foliation axial planar to F_{1a}	164°/63°
F_{1b}	isoclinal to open folds of S_{1a}	AP: 159°/64°; FA: 19°–335°
S_{1b}	metamorphic foliation axial planar to F_{1b}	164°/63°
$L_{1a,b}$	elongate metamorphic mineral growth	parallel to $F_{1a,b}$ fold hinges, or down-dip
D_2: E-W crustal shortening		
T_2	thick-skinned reverse faults	SSE-striking, NNE-dipping
L_2	mineral stretching and elongate growth	30°–265°
F_2	thick-skinned folds, E-vergent	AP: SSE-striking, NNE-dipping; FA: subhorizontal, trending SSE or NNW
D_3: N-S crustal shortening		
F_3	thick-skinned folds; crenulations	AP: 269°/51°; FA: 37°–269°
S_3	crenulation cleavage axial planar to F_3	269°/51°

Fig. 1. The original table and the reconstructed table (1)

Table 2
$^{40}Ar/^{39}Ar$ closure temperature calculations for samples mentioned in the text[a]

Sample	Location	Mineral	Composition	Diff radius, a (μm)	Act. energy, E (cal/mol)	Do/a^2	dT/dt (°C/Ma)	EdT/dt ($\times 10^{-9}$)	Approximate T_c (°C)	Ages, Ma (+/−)		%^{39}Ar (plateau age)
										Integrated	Plateau	
W. San. Felsic Dyke	Santoy	Biotite	Ann65	60	45 000	2139	30	42.81	290	1705 (7)	1757 (7)	82.4
3088	Santoy	Biotite	Ann80	1000	42 000	7.7	30	39.95	335	1711 (17)	1713 (6)	58.8
9222-20	Santoy	Biotite	Ann55	100	47 000	770	30	44.71	330	1709 (6)	1732 (6)	88.4
9222-62	Santoy	Biotite	Ann40	90	50 000	951	30	47.56	365	1733 (8)	1733 (8)	100.0
9222-73	Santoy	Hornblende	Ferroan pargasite	90	64 100	296	2.5	4.065	480	1715 (11)	1716 (11)	99.2
9222-9	Santoy	Hornblende	Ferroan pargasite	90	64 100	295	2.5	4.065	480	1713 (9)	1716 (9)	98.7
9222-56	Santoy	Hornblende	Ferroan pargasite	140	64 100	122	2.5	4.065	495	1711 (18)	1717 (19)	95.6
9222-41	Santoy	Hornblende	Ferroan pargasite	170	64 100	83	2.5	4.065	505	1737 (19)	1741 (19)	97.8
8822-1099	Brownell	Biotite	Ann55	90	47 000	951	4	4.471	300	1713 (8)	1727 (8)	91.9
kbl-1	Brownell	Biotite	Ann55	140	47 000	393	4	4.471	310	1827 (15)	1759 (6)	92.9
kbl-9	Brownell	Biotite	Ann55	125	47 000	493	4	4.471	310	1719 (7)	1745 (7)	73.5
kbl-2	Brownell	Biotite	Ann45	80	49 000	1203	4	4.661	320	1738 (7)	1743 (7)	96.7
kbl-8	Brownell	Biotite	Ann45	70	49 000	1571	4	4.661	315	1724 (8)	1756 (7)	90.4

[a] Diffusion coefficients for biotite from Harrison et al. (1985) and for hornblende from Harrison (1981). Abbreviations, T_c, closure temperature; Ann, annite component in biotite [Fe/(Fe+Mg)100]; D_o, diffusion coefficient; a, average grain diffusion radius; dT/dt, cooling rate in °C/s; act, activation.

Table 2
$^{40}Ar/^{39}Ar$ closure temperature calculations for samples mentioned in the text[a]

Sample	Location	Mineral	Composition	Diff radius, a (μm)	Act. energy, E (cal/mol)	Do/a^2	dT/dt (°C/Ma)	EdT/dt ($\times 10^{-9}$)	Approximate T_c (°C)	Ages, Ma (+/−)		%^{39}Ar (plateau age)
										Integrated	Plateau	
W. San. Felsic Dyke	Santoy	Biotite	Ann65	60	45 000	2139	30	42.81	290	1705 (7)	1757 (7)	82.4
3088	Santoy	Biotite	Ann80	1000	42 000	7.7	30	39.95	335	1711 (17)	1713 (6)	58.8
9222-20	Santoy	Biotite	Ann55	100	47 000	770	30	44.71	330	1709 (6)	1732 (6)	88.4
9222-62	Santoy	Biotite	Ann40	90	50 000	951	30	47.56	365	1733 (8)	1733 (8)	100.0
9222-73	Santoy	Hornblende	Ferroan pargasite	90	64 100	296	2.5	4.065	480	1715 (11)	1716 (11)	99.2
9222-9	Santoy	Hornblende	Ferroan pargasite	90	64 100	295	2.5	4.065	480	1713 (9)	1716 (9)	98.7
9222-56	Santoy	Hornblende	Ferroan pargasite	140	64 100	122	2.5	4.065	495	1711 (18)	1717 (19)	95.6
9222-41	Santoy	Hornblende	Ferroan pargasite	170	64 100	83	2.5	4.065	505	1737 (19)	1741 (19)	97.8
8822-1099	Brownell	Biotite	Ann55	90	47 000	951	4	4.471	300	1713 (8)	1727 (8)	91.9
kbl-1	Brownell	Biotite	Ann55	140	47 000	393	4	4.471	310	1827 (15)	1759 (6)	92.9
kbl-9	Brownell	Biotite	Ann55	125	47 000	493	4	4.471	310	1719 (7)	1745 (7)	73.5
kbl-2	Brownell	Biotite	Ann45	80	49 000	1203	4	4.661	320	1738 (7)	1743 (7)	96.7
kbl-8	Brownell	Biotite	Ann45	70	49 000	1571	4	4.661	315	1724 (8)	1756 (7)	90.4

[a] Diffusion coefficients for biotite from Harrison et al. (1985) and for hornblende from Harrison (1981). Abbreviations, T_c, closure temperature; Ann, annite component in biotite [Fe/(Fe+Mg)100]; D_o, diffusion coefficient; a, average grain diffusion radius; dT/dt, cooling rate in °C/s; act, activation.

Fig. 2. The original table and the reconstructed table (2)

Table 3
Single zircon Pb-evaporation results and interpreted $^{207}Pb/^{206}Pb$ ages for selected Santoy Lake area granitoids[a]

Sample	Zircon	$^{207}Pb/^{206}Pb$	(±)	$^{208}Pb/^{206}Pb$	Age (Ma)	(±)	Interpreted age (Ma)
3088	1	0.105714	962	0.022901	1727	17	n/a
	2	0.110973	621	0.125159	1815	10	
	3	0.126693	683	0.568054	2053	10	
	4	0.129915	8285	0.249933	2097	117	
West Santoy	1	0.115195	449	0.084297	1883	7	1886 ± 5*
Diorite	2	0.116044	1550	0.088469	1896	24	
	3	0.116430	883	0.087384	1902	18	
	4	0.114711	604	0.081209	1875	10	
	5	0.115404	539	0.087640	1886	9	
West Santoy	1	0.115123	303	0.081256	1882	5	1882 ± 4*
Felsite	2	0.115082	531	0.083569	1881	8	
West Santoy	1	0.114394	491	0.025535	1870	8	1870 ± 7*
Felsic dyke	2	0.114355	1253	0.026106	1870	19	
Zone 6	1	0.113696	348	0.080107	1859	6	1857 ± 3*
Felsite	2	0.114044	368	0.064292	1865	6	
	3	0.113040	318	0.078513	1849	5	

[a] * Weighted mean zircon age

Table 3
Single zircon Pb-evaporation results and interpreted $^{207}Pb/^{206}Pb$ ages for selected Santoy Lake area granitoids[a]

Sample	Zircon	$^{207}Pb/^{206}Pb$	(±)	$^{208}Pb/^{206}Pb$	Age (Ma)	(±)	Interpreted age (Ma)
3088	1	0.105714	962	0.022901	1727	17	n/a
	2	0.110973	621	0.125159	1815	10	
	3	0.126693	683	0.568054	2053	10	
	4	0.129915	8285	0.249933	2097	117	
West Santoy	1	0.115195	449	0.084297	1883	7	1886 ± 5*
Diorite	2	0.116044	1550	0.088469	1896	24	
	3	0.116430	883	0.087384	1902	18	
	4	0.114711	604	0.081209	1875	10	
	5	0.115404	539	0.087640	1886	9	
West Santoy	1	0.115123	303	0.081256	1882	5	1882 ± 4*
Felsite	2	0.115082	531	0.083569	1881	8	
West Santoy	1	0.114394	491	0.025535	1870	8	1870 ± 7*
Felsic dyke	2	0.114355	1253	0.026106	1870	19	
Zone 6	1	0.113696	348	0.080107	1859	6	1857 ± 3*
Felsite	2	0.114044	368	0.064292	1865	6	
	3	0.113040	318	0.078513	1849	5	

[a] * Weighted mean zircon age

Fig. 3. The original table and the reconstructed table (3)

TABLE II—1964 Peak Period Trans-Hudson Model Calibration Facility Comparisons

	Actual	Assignment Model Only	Assignment and Modal Split Models	Assignment Modal Split and Trip Interchange Models
Grand Total	159 438	159 398	159 394	158 659
Auto	45 909	45 909	46 199	46 230
GWB	23 560	23 675	22 847	23 091
LT	11 159	11 677	12 618	12 466
HT	3 214	3 269	3 474	3 586
SIB	2 770	2 540	2 534	2 672
TZ	5 205	4 749	4 726	4 416
Bus	58 845	58 833	58 142	57 195
GWB	11 268	10 270	10 860	10 765
PABT	47 577	48 563	47 282	46 430
Rail	54 684	54 658	55 056	55 235
PS	7 593	7 449	7 543	7 504
HT	26 060	25 652	25 282	25 199
PUP	13 153	12 885	13 521	13 954
CNJ	7 878	8 671	8 712	8 578

TABLE II—1964 Peak Period Trans-Hudson Model Calibration Facility Comparisons

	Actual	Assignment Model Only	Assignment and Modal Split Models	Assignment Modal Split and Trip Interchange Models
Grand Total	159 438	159 398	159 394	158 659
Auto	45 909	45 909	46 199	46 230
GWB	23 560	23 675	22 847	23 091
LT	11 159	11 677	12 618	12 466
HT	3 214	3 269	3 474	3 586
SIB	2 770	2 540	2 534	2 672
TZ	5 205	4 749	4 726	4 416
Bus	58 845	58 833	58 142	57 195
GWB	11 268	10 270	10 860	10 765
PABT	47 577	48 563	47 282	46 430
Rail	54 684	54 658	55 056	55 235
PS	7 593	7 449	7 543	7 504
HT	26 060	25 652	25 282	25 199
PUP	13 153	12 885	13 521	13 954
CNJ	7 878	8 671	8 712	8 578

Fig. 4. The original table and the reconstructed table (4)

Table 1
Electron microprobe analyzed representative mineral compositions used in *P–T* Pseudosection modeling (SG-158A) of schists/phyllites from Mahakoshal Belt.

Sample	SG-158A (Andalusite - bearing schists/phyllites)										SG-159C (Andalusite - corundum - quartz bearing vein)				
Mineral	Biotite (1–2)		Muscovite (3–5)			Margarite (6–7)		Chlorite (8–9)		Andalusite	Muscovite	Chlorite	Diaspore	Corundum	Andalusite
Dataset	1/1.	2/1.	3/1.	4/1.	5/1.	6/1.	7/1.	8/1.	9/1.	10/1.	1/1.	2/1.	3/1.	4/1.	5/1.
SiO_2	34.47	36.47	47.78	47.35	47.42	30.56	31.41	23.30	23.75	36.62	47.78	26.22	0.15	0.12	35.86
TiO_2	2.10	2.11	0.21	0.26	0.40	0.00	0.00	0.16	0.06	0.02	0.21	0.10	0.03	0.00	0.01
Al_2O_3	21.16	19.16	36.57	35.58	37.69	50.49	49.67	21.77	21.88	62.17	36.57	20.46	85.11	96.83	63.81
Cr_2O_3	0.00	0.00	0.00	0.00	0.02	0.00	0.05	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.03
FeO^*	20.12	21.11	0.94	1.01	0.97	0.73	0.39	31.00	30.58	0.22	0.94	29.25	0.34	0.00	0.28
MnO	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.04	0.08	0.00	0.00	0.09	0.02	0.00	0.00
MgO	5.13	4.13	0.36	0.44	0.42	0.16	0.11	9.34	9.53	0.04	0.36	10.46	0.00	0.00	0.04
CaO	0.10	0.90	0.00	0.02	0.02	10.38	10.46	0.11	0.09	0.01	0.00	0.21	0.03	0.00	0.01
Na_2O	0.11	0.12	1.10	1.03	0.92	2.14	1.98	0.05	0.04	0.01	1.10	0.04	0.01	0.00	0.05
K_2O	9.70	9.71	9.44	9.45	9.40	0.02	0.09	0.03	0.02	0.00	9.44	0.43	0.02	0.00	0.02
Total	92.91	93.72	96.39	95.14	96.50	94.47	94.32	85.88	86.10	99.12	96.39	87.32	85.72	97.07	99.17
No. of (O)	11	11	11	11	11	11	11	14	14	5	11	14	–	3	5
Si	2.70	2.84	3.11	3.13	3.06	2.04	2.10	2.60	2.63	1.00	3.11	2.84	–	0.00	1.00
Ti	0.12	0.12	0.01	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.01	–	0.00	0.00
Al	1.95	1.76	2.81	2.77	2.87	3.98	3.92	2.86	2.86	2.00	2.81	2.61	–	2.00	2.00
Cr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00
Fe^{+2}	1.32	1.38	0.05	0.06	0.05	0.04	0.02	2.89	2.83	0.01	0.05	2.65	–	0.00	0.00
Mn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	–	0.00	0.00
Mg	0.60	0.48	0.03	0.04	0.04	0.02	0.01	1.55	1.58	0.00	0.03	1.69	–	0.00	0.00
Ca	0.01	0.08	0.00	0.00	0.00	0.74	0.75	0.01	0.01	0.00	0.00	0.02	–	0.00	0.00
Na	0.02	0.02	0.14	0.13	0.12	0.28	0.26	0.01	0.01	0.00	0.14	0.01	–	0.00	0.00
K	0.97	0.97	0.78	0.80	0.77	0.00	0.01	0.00	0.00	0.00	0.78	0.06	–	0.00	0.00
Total	7.69	7.64	6.94	6.94	6.93	7.10	7.07	9.96	9.94	3.00	6.94	9.89	–	2.00	3.00
X_{Fe}	0.69	0.74	–	–	–	–	–	–	–	–	–	–	–	–	–

[*] FeO indicates total iron

Table 1
Electron microprobe analyzed representative mineral compositions used in *P–T* Pseudosection modeling (SG-158A) of schists/phyllites from Mahakoshal Belt.

Sample	SG-158A (Andalusite - bearing schists/phyllites)										SG-159C (Andalusite - corundum - quartz bearing vein)				
Mineral	Biotite (1–2)		Muscovite (3–5)			Margarite (6–7)		Chlorite (8–9)		Andalusite	Muscovite	Chlorite	Diaspore	Corundum	Andalusite
Dataset	1/1.	2/1.	3/1.	4/1.	5/1.	6/1.	7/1.	8/1.	9/1.	10/1.	1/1.	2/1.	3/1.	4/1.	5/1.
SiO_2	34.47	36.47	47.78	47.35	47.42	30.56	31.41	23.30	23.75	36.62	47.78	26.22	0.15	0.12	35.86
TiO_2	2.10	2.11	0.21	0.26	0.40	0.00	0.00	0.16	0.06	0.02	0.21	0.10	0.03	0.00	0.01
Al_2O_3	21.16	19.16	36.57	35.58	37.69	50.49	49.67	21.77	21.88	62.17	36.57	20.46	85.11	96.83	63.81
Cr_2O_3	0.00	0.00	0.00	0.00	0.02	0.00	0.05	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.03
FeO^*	20.12	21.11	0.94	1.01	0.97	0.73	0.39	31.00	30.58	0.22	0.94	29.25	0.34	0.00	0.28
MnO	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.04	0.08	0.00	0.00	0.09	0.02	0.00	0.00
MgO	5.13	4.13	0.36	0.44	0.42	0.16	0.11	9.34	9.53	0.04	0.36	10.46	0.00	0.00	0.04
CaO	0.10	0.90	0.00	0.02	0.02	10.38	10.46	0.11	0.09	0.01	0.00	0.21	0.03	0.00	0.01
Na_2O	0.11	0.12	1.10	1.03	0.92	2.14	1.98	0.05	0.04	0.01	1.10	0.04	0.01	0.00	0.05
K_2O	9.70	9.71	9.44	9.45	9.40	0.02	0.09	0.03	0.02	0.00	9.44	0.43	0.02	0.00	0.02
Total	92.91	93.72	96.39	95.14	96.50	94.47	94.32	85.88	86.10	99.12	96.39	87.32	85.72	97.07	99.17
No. of (O)	11	11	11	11	11	11	11	14	14	5	11	14	–	3	5
Si	2.70	2.84	3.11	3.13	3.06	2.04	2.10	2.60	2.63	1.00	3.11	2.84	–	0.00	1.00
Ti	0.12	0.12	0.01	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.01	–	0.00	0.00
Al	1.95	1.76	2.81	2.77	2.87	3.98	3.92	2.86	2.86	2.00	2.81	2.61	–	2.00	2.00
Cr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00
Fe^{+2}	1.32	1.38	0.05	0.06	0.05	0.04	0.02	2.89	2.83	0.01	0.05	2.65	–	0.00	0.00
Mn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	–	0.00	0.00
Mg	0.60	0.48	0.03	0.04	0.04	0.02	0.01	1.55	1.58	0.00	0.03	1.69	–	0.00	0.00
Ca	0.01	0.08	0.00	0.00	0.00	0.74	0.75	0.01	0.01	0.00	0.00	0.02	–	0.00	0.00
Na	0.02	0.02	0.14	0.13	0.12	0.28	0.26	0.01	0.01	0.00	0.14	0.01	–	0.00	0.00
K	0.97	0.97	0.78	0.80	0.77	0.00	0.01	0.00	0.00	0.00	0.78	0.06	–	0.00	0.00
Total	7.69	7.64	6.94	6.94	6.93	7.10	7.07	9.96	9.94	3.00	6.94	9.89	–	2.00	3.00
X_{Fe}	0.69	0.74	–	–	–	–	–	–	–	–	–	–	–	–	–

[*] FeO indicates total iron

Fig. 5. The original table and the reconstructed table (5)

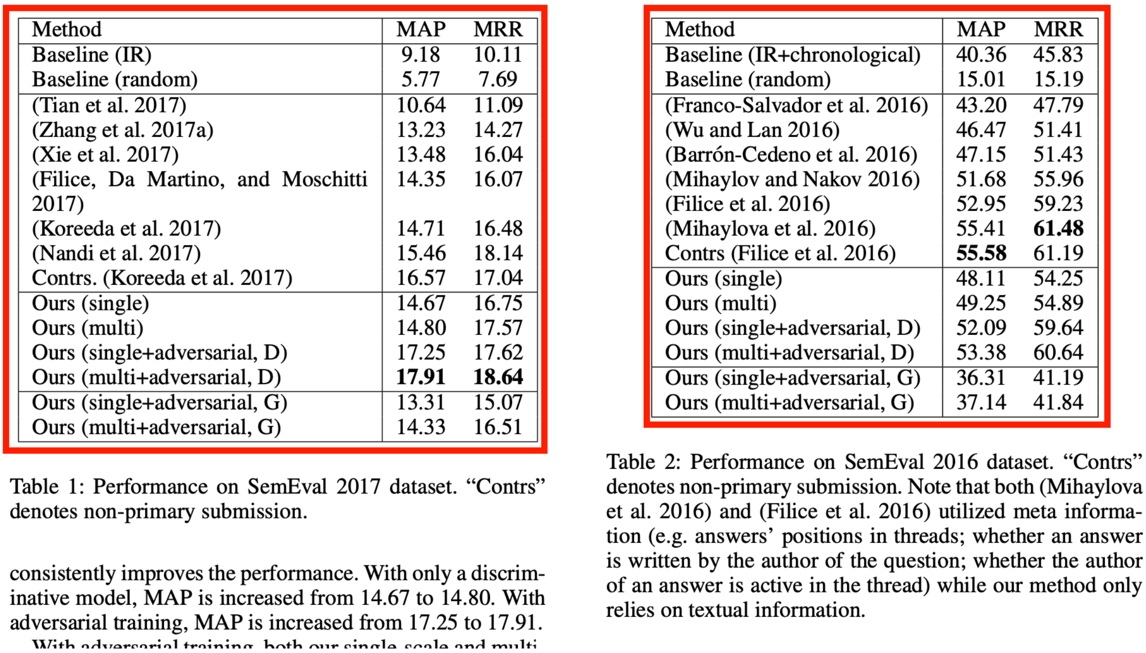
Analyses From results we can observe that, our algorithm can detect the table boundary accurately, which completes table reconstruction task.

3 Extracting Tables from PDFs

We use a open-source project pdftotree* to get the hierarchical tree of context objects such as text blocks, figures, tables, etc. For tables, it will extract their bounding boxes. Therefore, we can use the bounding boxes to extract tables from PDF files. Here is the specific pipeline:

1. Use pdftotree* to get the hierarchical tree of the context objects.
2. Find table objects in the tree extracted in step 1, and get the bounding boxes of the tables from the extracted tree.
3. Extract tables according to the bounding boxes.

Here is an example of the bounding boxes we extracted from a PDF paper†.



Method	MAP	MRR
Baseline (IR)	9.18	10.11
Baseline (random)	5.77	7.69
(Tian et al. 2017)	10.64	11.09
(Zhang et al. 2017a)	13.23	14.27
(Xie et al. 2017)	13.48	16.04
(Filice, Da Martino, and Moschitti 2017)	14.35	16.07
(Koreeda et al. 2017)	14.71	16.48
(Nandi et al. 2017)	15.46	18.14
Contrs. (Koreeda et al. 2017)	16.57	17.04
Ours (single)	14.67	16.75
Ours (multi)	14.80	17.57
Ours (single+adversarial, D)	17.25	17.62
Ours (multi+adversarial, D)	**17.91**	**18.64**
Ours (single+adversarial, G)	13.31	15.07
Ours (multi+adversarial, G)	14.33	16.51

Table 1: Performance on SemEval 2017 dataset. "Contrs" denotes non-primary submission.

consistently improves the performance. With only a discriminative model, MAP is increased from 14.67 to 14.80. With adversarial training, MAP is increased from 17.25 to 17.91.

Method	MAP	MRR
Baseline (IR+chronological)	40.36	45.83
Baseline (random)	15.01	15.19
(Franco-Salvador et al. 2016)	43.20	47.79
(Wu and Lan 2016)	46.47	51.41
(Barrón-Cedeno et al. 2016)	47.15	51.43
(Mihaylov and Nakov 2016)	51.68	55.96
(Filice et al. 2016)	52.95	59.23
(Mihaylova et al. 2016)	55.41	**61.48**
Contrs (Filice et al. 2016)	**55.58**	61.19
Ours (single)	48.11	54.25
Ours (multi)	49.25	54.89
Ours (single+adversarial, D)	52.09	59.64
Ours (multi+adversarial, D)	53.38	60.64
Ours (single+adversarial, G)	36.31	41.19
Ours (multi+adversarial, G)	37.14	41.84

Table 2: Performance on SemEval 2016 dataset. "Contrs" denotes non-primary submission. Note that both (Mihaylova et al. 2016) and (Filice et al. 2016) utilized meta information (e.g. answers' positions in threads; whether an answer is written by the author of the question; whether the author of an answer is active in the thread) while our method only relies on textual information.

Fig. 6. The bounding boxes of the table, which is extracted using pdftotree

Then, we can easily extract the images from the PDF file, and use the algorithms introduced in Section 2 to draw the framelines. The results are shown in Fig. 7 in the next page. Therefore, we have successfully build a pipeline to extract tables from PDFs and reconstruct the framelines of the tables.

4 Answers to Questions

1. **Q**: How to automatically locate the tables in a PDF?

   A: The table is very structured data. Therefore, we propose the following pipeline to automatically locate the tables in a PDF.

*https://github.com/HazyResearch/pdftotree

†Yang, Xiao, et al. "Adversarial training for community question answer selection based on multi-scale matching." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

Method	MAP	MRR
Baseline (IR)	9.18	10.11
Baseline (random)	5.77	7.69
(Tian et al. 2017)	10.64	11.09
(Zhang et al. 2017a)	13.23	14.27
(Xie et al. 2017)	13.48	16.04
(Filice, Da Martino, and Moschitti 2017)	14.35	16.07
(Koreeda et al. 2017)	14.71	16.48
(Nandi et al. 2017)	15.46	18.14
Contrs. (Koreeda et al. 2017)	16.57	17.04
Ours (single)	14.67	16.75
Ours (multi)	14.80	17.57
Ours (single+adversarial, D)	17.25	17.62
Ours (multi+adversarial, D)	**17.91**	**18.64**
Ours (single+adversarial, G)	13.31	15.07
Ours (multi+adversarial, G)	14.33	16.51

Method	MAP	MRR
Baseline (IR+chronological)	40.36	45.83
Baseline (random)	15.01	15.19
(Franco-Salvador et al. 2016)	43.20	47.79
(Wu and Lan 2016)	46.47	51.41
(Barrón-Cedeno et al. 2016)	47.15	51.43
(Mihaylov and Nakov 2016)	51.68	55.96
(Filice et al. 2016)	52.95	59.23
(Mihaylova et al. 2016)	55.41	**61.48**
Contrs (Filice et al. 2016)	**55.58**	61.19
Ours (single)	48.11	54.25
Ours (multi)	49.25	54.89
Ours (single+adversarial, D)	52.09	59.64
Ours (multi+adversarial, D)	53.38	60.64
Ours (single+adversarial, G)	36.31	41.19
Ours (multi+adversarial, G)	37.14	41.84

Fig. 7. The original table and the reconstructed table (4)

(a) Parse each page from PDF and get the coordinates of characters and text lines for text detection;

(b) Preprocess the texts to remove all the blank spaces and special characters;

(c) Perform K-means clustring technique, and store the IDs and centroids of the clusters.

(d) Find the appropriate clusters, that is different from simple text cluster, then the cluster area should be a table.

(e) Check all clusters and then we can automatically locate all the tables in a PDF.

In contemperary papers, tables all have a header, which should have be the destination of a hyper-link somewhere. Therefore, for this kind of papers, we just need to find all the hyper-links, and use the methods we proposed above to check whether it is a table.

In our pipeline, we use a handy tool pdftotree* to automatically find the tables in the PDFs using pretrained machine learning models.

2. **Q**: What do you think is the most difficult step to extract the table from the PDF? Why?

   A: Our pipeline only completes the prepositive works, such as extract tables from PDF file and perform table reconstruction by drawing framelines. The process is not very difficult, since we can use many handy tools to efficient parse a large batch of PDF files, extract tables according to bounding boxes coordinates, and draw framelines using digital image processing knowledges. **I think the most difficult step of the table reconstruction task is to further parse the reconstructed paper into structured data.** We know that the table may have various structures, such as merging several cells into one, *etc.* Under such circumstances, our previous algorithm may divide the combined block into several blocks, and use several lines to represent a block. For example, our frameline-drawing algorithm may misidentify table structure as shown in Fig. 8 in the next page.

   So how to identify the combined blocks and other special structure of the table, and then extract the information in the table to a structured data remains challenging. Therefore, I think this is the most difficult step to extract the table from the PDF.

3. **(Bonus) Q**: How to accurately identify the header of the table, and use natural language processing (NLP) or other methods to understand the information in the table, and then extract the entities and relationships from table to construct a specific knowledge graph?

*https://github.com/HazyResearch/pdftotree

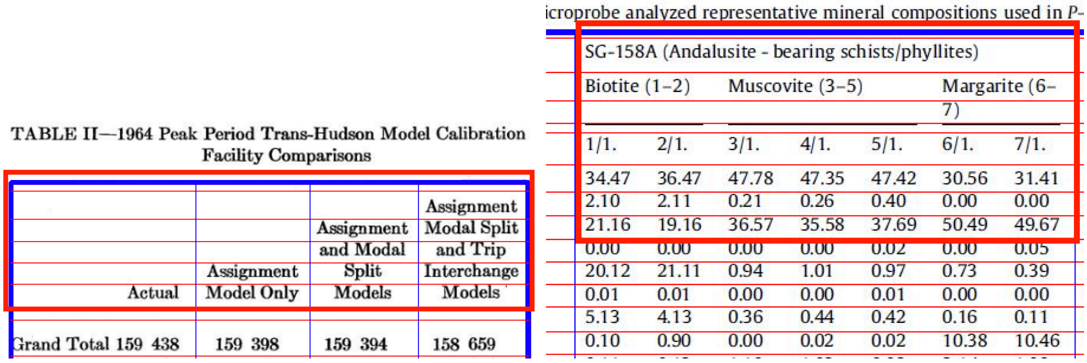


TABLE II—1964 Peak Period Trans-Hudson Model Calibration Facility Comparisons

	Actual	Assignment Model Only	Assignment and Modal Split Models	Assignment Modal Split and Trip Interchange Models
Grand Total	159 438	159 398	159 394	158 659

icroprobe analyzed representative mineral compositions used in P-

SG-158A (Andalusite - bearing schists/phyllites)						
Biotite (1–2)		Muscovite (3–5)			Margarite (6–7)	
1/1.	2/1.	3/1.	4/1.	5/1.	6/1.	7/1.
34.47	36.47	47.78	47.35	47.42	30.56	31.41
2.10	2.11	0.21	0.26	0.40	0.00	0.00
21.16	19.16	36.57	35.58	37.69	50.49	49.67
0.00	0.00	0.00	0.00	0.02	0.00	0.05
20.12	21.11	0.94	1.01	0.97	0.73	0.39
0.01	0.01	0.00	0.00	0.01	0.00	0.00
5.13	4.13	0.36	0.44	0.42	0.16	0.11
0.10	0.90	0.00	0.02	0.02	10.38	10.46

Fig. 8. Examples of the misidentified situations (the red boxes)

A: **Identify the header**: As mentioned above, for PDFs of the contemperary academic papers, the table header must be hyper-linked by somewhere in the document. Therefore, we can simply check all the destination of all hyperlinks in the PDF file, then we can find all table headers. For old scanned documents of PDF format, we may check the bold letters to find the word such as "Table" and "Tab" to find the tables. Actually, I think for those kind of documents, it is more like a computer vision task to recognize table headers from a given image. I think the classic computer vision network like ResNets* can have a great performance on detecting task like this.

Extract information in the table: We can use OCR to extract texts in the pictures. According to our discussion in problem 2, the main challenge is how to get a structured table based on these informations. I would like to regard the task as a **image segmentation** task in computer vision field. We can regard each block in the table as a region and use image segmentation networks to learn how to divide the table into structured data.

Build a knowledge graph: In this part, we may use the NLP models like Bert † to extract the connection between the row name and the column name. Then, we can regard the data as the label of the edge between row name and column name in the knowledge graph. Therefore, we can extract the embeddings from the tables to construct the knowledge graph.

5 Conclusion

In this lab, we explore the table reconstruction tasks of PDF files. I make clear explanations to the given code and modify the code a bit to satisfy the task settings. We have shown our reconstruction results and they are quite satisfiable: they can detect the table boundary accurately and reconstruct the table using several horizontal and vertical lines. We also propose a handy pipeline to extract tables from PDF files, and reconstruct it by drawing framelines.

We also explore further in the questions of how to extract information from the PDF file. We have proposed three ideas concerning three aspects:

- Take the "identify the header" task as a computer vision task and use models like ResNets* to identify the header efficiently and effectively in scanned PDF documents. Check the destinations of hyper-links to identify the headers in today's PDf documents.

*He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

†Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

- Take the "Extract information in the table" as an image segmentation task. Recognize the texts using OCR techniques and extract the precise structure of the table using image segmentation models. Therefore, we can extract more accurate information from the table. Actually, the pdftotree[§] tool we used is based on the same idea.
- Take "Build a knowledge graph" as an natural language processing task. Use models like Bert[†] to extract embeddings from row name and column name. Then, use the data as the label of the edge between row name and the column name in the knowledge graph. Finally, extract the information you want from the built knowledge graph.

In conclusion, I gain knowledges from this lab and I think the lab benifits me a lot.

The full implementation codes of the lab is available in my github repository. For other questions about the lab, please feel free to send an e-mail[‡] to me.

[§] https://github.com/HazyResearch/pdftotree

[‡] mailto:galaxies@sjtu.edu.cn