Segmentation Of Layout-Based Documents

Bachelor's Thesis

Albert-Ludwigs-Universität Freiburg

Elias Kempf

Department of Computer Science
Chair of Algorithms and Data Structures

October 20, 2021

What problem do we want to solve?

What problem do we want to solve?

- Extracting text blocks from PDFs

What problem do we want to solve?

- Extracting text blocks from PDFs

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

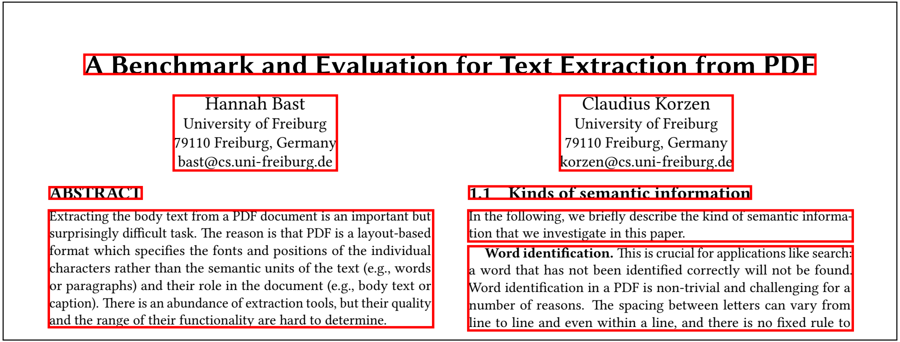
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

What problem do we want to solve?

- Extracting text blocks from PDFs



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

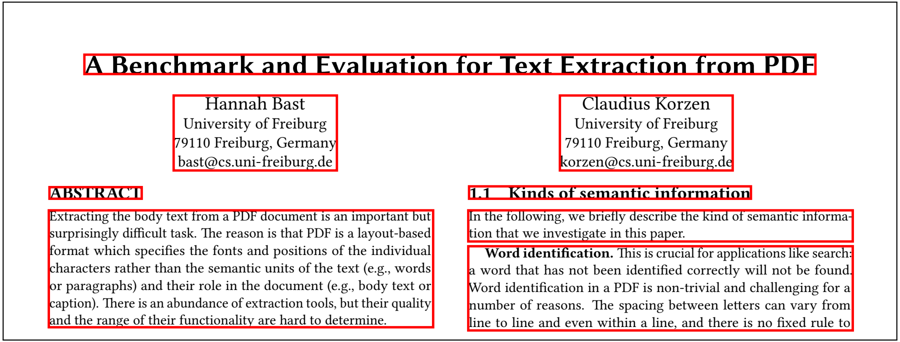
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

What problem do we want to solve?

- Extracting text blocks from PDFs
- Sort extracted text blocks by natural reading order



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

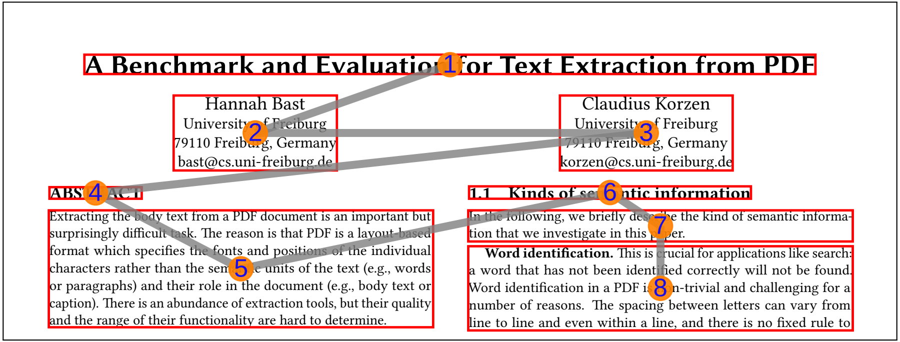
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

What problem do we want to solve?

- Extracting text blocks from PDFs
- Sort extracted text blocks by natural reading order



Why is this problem difficult?

Why is this problem difficult?

- PDF is a layout-based format (meaning it does not store text as words, lines, or paragraphs)

Why is this problem difficult?

- PDF is a layout-based format (meaning it does not store text as words, lines, or paragraphs)
- Only characters, their bounding boxes, and font information is stored

Why is this problem difficult?

- PDF is a layout-based format (meaning it does not store text as words, lines, or paragraphs)
- Only characters, their bounding boxes, and font information is stored
- Usually also no whitespace characters

Why is this problem difficult?

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg

Claudius Korzen
University of Freiburg

Why is this problem difficult?

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg

Claudius Korzen
University of Freiburg

Character	bounding box	font name	font size
„A“	$(75.8, 697.2), (87.9, 708.5)$	Arial	17.2

Why is this problem difficult?

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg

Claudius Korzen
University of Freiburg

Character	bounding box	font name	font size
„A“	$(75.8, 697.2), (87.9, 708.5)$	Arial	17.2
„B“	$(92.2, 697.2), (103.8, 708.4)$	Arial	17.2

Why is this problem difficult?

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg

Claudius Korzen
University of Freiburg

Character	bounding box	font name	font size
„A“	$(75.8, 697.2), (87.9, 708.5)$	Arial	17.2
„B“	$(92.2, 697.2), (103.8, 708.4)$	Arial	17.2
„e“	$(103.8, 697.1), (112.5, 704.8)$	Arial	17.2

Why is this problem difficult?


A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg

Claudius Korzen
University of Freiburg

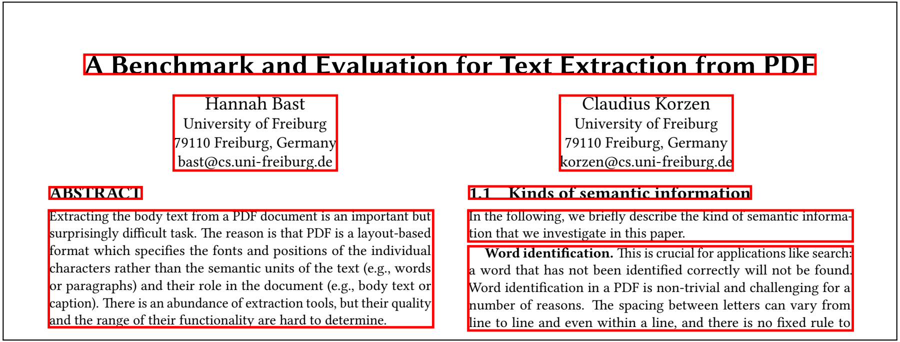
Character	bounding box	font name	font size
„A“	$(75.8, 697.2), (87.9, 708.5)$	Arial	17.2
„B“	$(92.2, 697.2), (103.8, 708.4)$	Arial	17.2
„e“	$(103.8, 697.1), (112.5, 704.8)$	Arial	17.2
	⋮		

Why is this problem difficult?

- Reading order is also often difficult to detect

Why is this problem difficult?

- Reading order is also often difficult to detect
- Especially, in documents featuring a two-column layout:



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Problem definition

For each page of a given PDF, . . .

- Our input is a list of characters. Each character comes with its bounding box and its font information.
- Our output is a list of text blocks sorted by reading order.

Approach outline

Approach outline

- We seperate our approach into two main steps

Approach outline

- We seperate our approach into two main steps
- First, we use page segmentation to detect text blocks

Approach outline

- We seperate our approach into two main steps
- First, we use page segmentation to detect text blocks
- Second, we order the detected text blocks using a similar but more informed approach

Page segmentation

- Page segmentation is the process of reassembling the characters of a layout-based document into semantic units like words, lines, or text blocks

Page segmentation

- Page segmentation is the process of reassembling the characters of a layout-based document into semantic units like words, lines, or text blocks
- We only focus on reassembling characters into text blocks

Page segmentation

- Page segmentation is the process of reassembling the characters of a layout-based document into semantic units like words, lines, or text blocks
- We only focus on reassembling characters into text blocks
- We perform our segmentation using an XY-cut algorithm

XY-cut algorithm

- An XY-cut algorithm can be used to group the characters of a page

XY-cut algorithm

- An XY-cut algorithm can be used to group the characters of a page
- It does so by applying vertical cuts (through the **X**-axis) and horizontal cuts (through the **Y**-axis) to the page

XY-cut algorithm

- An XY-cut algorithm can be used to group the characters of a page
- It does so by applying vertical cuts (through the **X**-axis) and horizontal cuts (through the **Y**-axis) to the page
- Diagonal cuts and cuts through characters are not allowed

XY-cut algorithm

- An XY-cut algorithm can be used to group the characters of a page
- It does so by applying vertical cuts (through the **X**-axis) and horizontal cuts (through the **Y**-axis) to the page
- Diagonal cuts and cuts through characters are not allowed
- Cuts can also be used to detect reading order (more on that later)

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

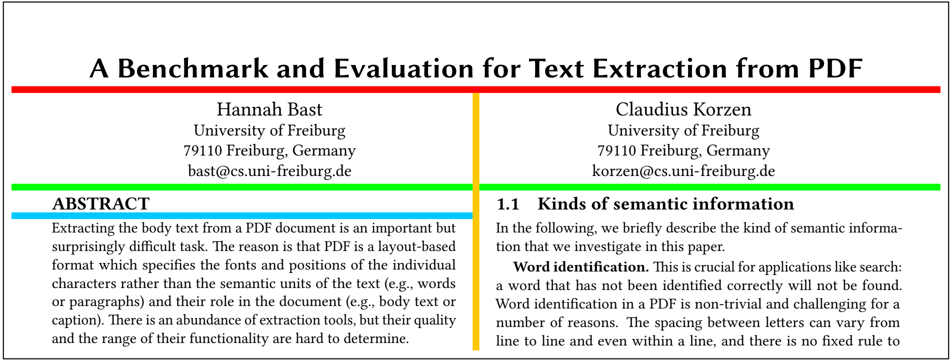
Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

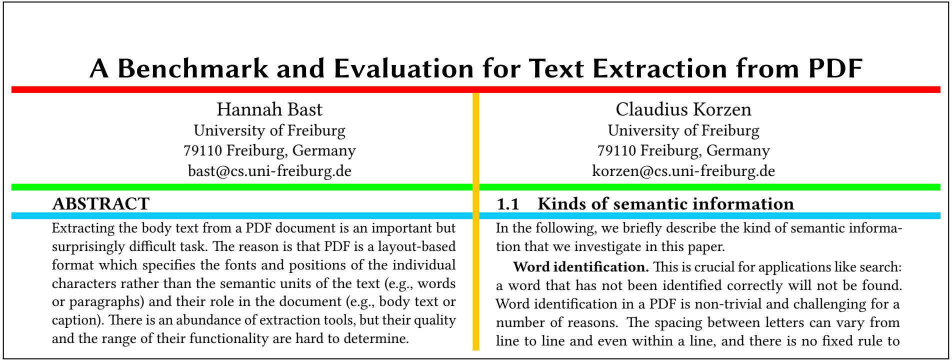
Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

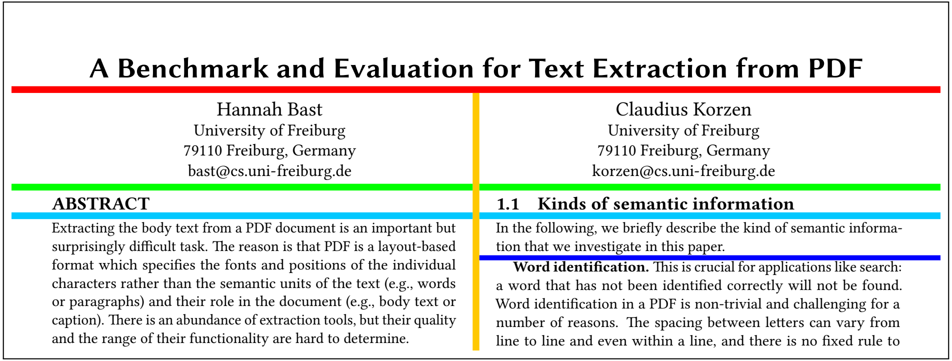
Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Segmenting a page

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Reading order detection

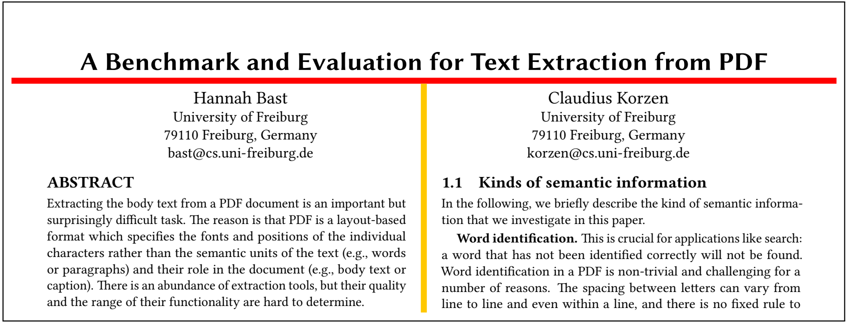
- Each cut divides a page into two parts

Reading order detection

- Each cut divides a page into two parts
- We order them respecting our top-to-bottom left-to-right writing system

Reading order detection

- Each cut divides a page into two parts
- We order them respecting our top-to-bottom left-to-right writing system



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Reading order detection

- Detecting text blocks also yields a preliminary reading order

Reading order detection

- Detecting text blocks also yields a preliminary reading order
- Problem: cuts are not specifically chosen for reading order detection

Reading order detection

- Detecting text blocks also yields a preliminary reading order
- Problem: cuts are not specifically chosen for reading order detection

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Reading order detection

- Detecting text blocks also yields a preliminary reading order
- Problem: cuts are not specifically chosen for reading order detection

A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

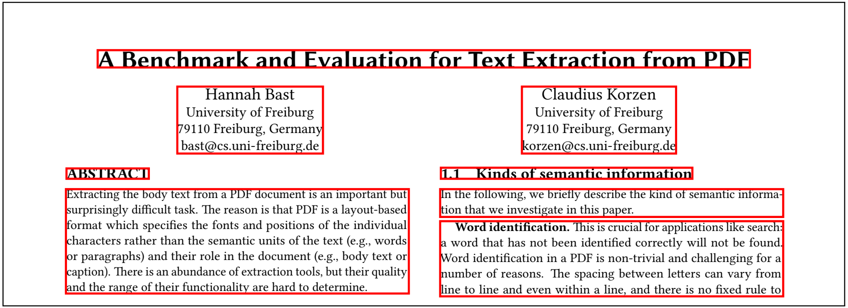
Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Reading order detection

- Solution: detect reading order **after** detecting text blocks

Reading order detection

- Solution: detect reading order **after** detecting text blocks
- This allows us to make use of information about the text blocks themselves:



A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

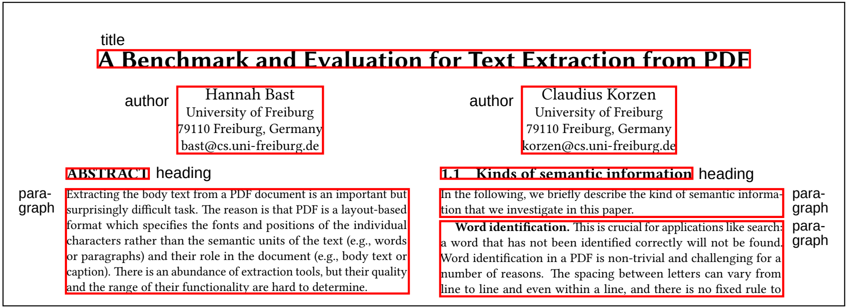
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

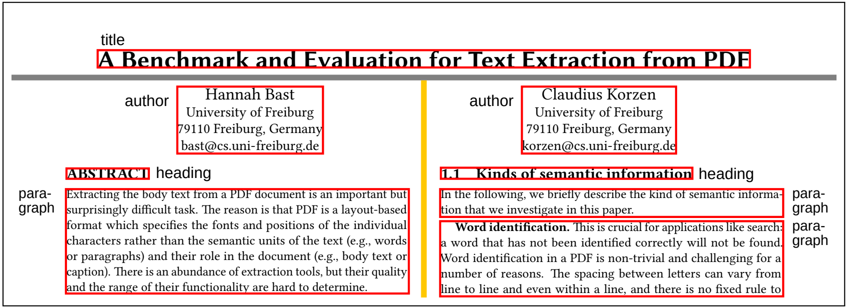
Reading order detection

- Solution: detect reading order **after** detecting text blocks
- This allows us to make use of information about the text blocks themselves:



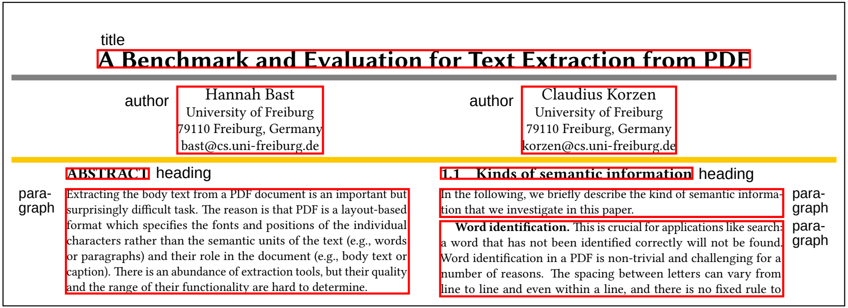
Reading order detection

- Solution: detect reading order **after** detecting text blocks
- This allows us to make use of information about the text blocks themselves:



Reading order detection

- Solution: detect reading order **after** detecting text blocks
- This allows us to make use of information about the text blocks themselves:



Reading order detection

- We predict the semantic roles of detected text blocks using a machine-learning model developed by Korzen

Reading order detection

- We predict the semantic roles of detected text blocks using a machine-learning model developed by Korzen
- We then use the XY-cut algorithm again but now choose cuts using a machine-learning model which also considers semantic roles

Reading order detection

- We predict the semantic roles of detected text blocks using a machine-learning model developed by Korzen
- We then use the XY-cut algorithm again but now choose cuts using a machine-learning model which also considers semantic roles
- This way, we can correct potential mistakes in the preliminary reading order

Evaluation setup

Evaluation setup

- We evaluate our approach on 1,750 randomly selected articles from `arXiv.org`

Evaluation setup

- We evaluate our approach on 1,750 randomly selected articles from `arXiv.org`
- We evaluate both text block detection and reading order detection

Evaluation setup

- We evaluate our approach on 1,750 randomly selected articles from `arXiv.org`
- We evaluate both text block detection and reading order detection
- We compute the expected text blocks and their natural reading order from TeX data

Evaluation setup

- We evaluate our approach on 1,750 randomly selected articles from `arXiv.org`
- We evaluate both text block detection and reading order detection
- We compute the expected text blocks and their natural reading order from TEX data
- We compare text blocks using their bounding boxes

Evaluation metrics

- Text block detection:

Evaluation metrics

- Text block detection:
  - $B^{=}_{G}$:= percentage of expected blocks that were detected

Evaluation metrics

- Text block detection:
  - $B_G^{=}$:= percentage of expected blocks that were detected
  - $B_A^{=}$:= percentage of detected blocks that were expected

Evaluation metrics

- Text block detection:
  - $B_G^=$:= percentage of expected blocks that were detected
  - $B_A^=$:= percentage of detected blocks that were expected
  - B_G^+ := percentage of expected blocks that were split too much

Evaluation metrics

- Text block detection:
  - $B_G^{=}$:= percentage of expected blocks that were detected
  - $B_A^{=}$:= percentage of detected blocks that were expected
  - B_G^{+} := percentage of expected blocks that were split too much
  - B_A^{-} := percentage of detected blocks that were split too less

Evaluation metrics

- Text block detection:
  - $B_G^{=}$:= percentage of expected blocks that were detected
  - $B_A^{=}$:= percentage of detected blocks that were expected
  - B_G^{+} := percentage of expected blocks that were split too much
  - B_A^{-} := percentage of detected blocks that were split too less

- Reading order detection:

Evaluation metrics

- Text block detection:
  - $B_G^{=}$:= percentage of expected blocks that were detected
  - $B_A^{=}$:= percentage of detected blocks that were expected
  - B_G^{+} := percentage of expected blocks that were split too much
  - B_A^{-} := percentage of detected blocks that were split too less

- Reading order detection:
  - τ_n := the normalized Kendall-τ-correlation between expected and detected reading order

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$7,\ 5,\ 6,\ 9$$

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$7,\ 5,\ 6,\ 9$$

Concordant pairs: 0 Discordant pairs: 0

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$\overbrace{7,\ 5},\ 6,\ 9$$

Concordant pairs: 0

Discordant pairs: 1

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$\overbrace{\color{red}{7}, 5, \color{red}{6}}, 9$$

Concordant pairs: 0

Discordant pairs: 2

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$\overbrace{\color{red}{7}, 5, 6, \color{red}{9}}$$

Concordant pairs: 1

Discordant pairs: 2

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$7, \overbrace{5, 6}, 9$$

Concordant pairs: 2

Discordant pairs: 2

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$7,\ \overbrace{\color{red}{5},\ 6,\ \color{red}{9}}$$

Concordant pairs: 3

Discordant pairs: 2

Kendall-τ-correlation

- τ can be used to compare the order of a sequence of numbers to an ascending order
- To compute τ, we count concordant and discordant pairs in a given sequence
- Let's look at an example:

$$7,\ 5,\ \overbrace{6,\ 9}$$

Concordant pairs: 4

Discordant pairs: 2

Kendall-τ-correlation

- Assuming the sequence does not contain duplicates, we can define τ as

$$\tau = \frac{\#con - \#dis}{\#con + \#dis}$$

Kendall-τ-correlation

- Assuming the sequence does not contain duplicates, we can define τ as

$$\tau = \frac{\#con - \#dis}{\#con + \#dis}$$

- For our example, we obtained $\#con = 4$ and $\#dis = 2$ yielding

$$\tau = \frac{4-2}{4+2} = \frac{2}{6} = \frac{1}{3}$$

Kendall-τ-correlation

- Assuming the sequence does not contain duplicates, we can define τ as

$$\tau = \frac{\#con - \#dis}{\#con + \#dis}$$

- For our example, we obtained $\#con = 4$ and $\#dis = 2$ yielding

$$\tau = \frac{4-2}{4+2} = \frac{2}{6} = \frac{1}{3}$$

- τ takes values between -1 and 1, so we normalize it using

$$\tau_n = \frac{\tau + 1}{2}$$

Evaluation results

- Average metric values on our evaluation dataset:

$B_G^{=}$	$B_A^{=}$	B_G^{+}	B_A^{-}	τ_n	τ_n^f
51.4%	46.7%	12.9%	14.7%	0.873	0.994

Input format

```
{
    "glyphs": [{
        "char": "A",
        "font size": "11pt",
        "bounding box": [1, 4, 2.5, 6]
     },
     {
        "char": "4",
        "font size": "11pt",
        "bounding box": [1, 1, 3, 3]
     }]
}
```

Some important remarks

- The example omitted some important aspects:
  - How do we compute potential cuts algorithmically?
    $\Longrightarrow$ using projection profiles of bounding boxes
  - How do we decide which cut to choose?
    $\Longrightarrow$ based on cut size
  - When do we stop cutting?
    $\Longrightarrow$ after cut size falls below a certain size threshold

Text block postprocessing

Raw detected block

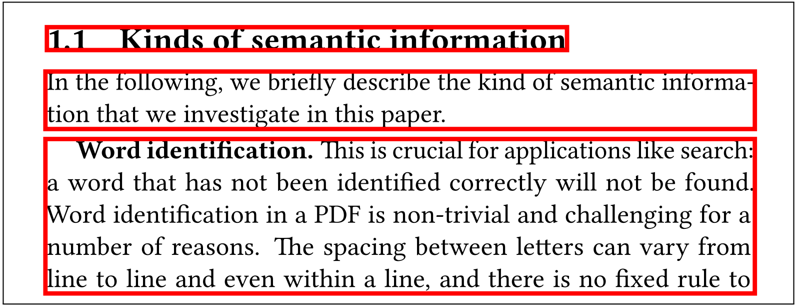
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Text block postprocessing

Splits applied in postprocessing



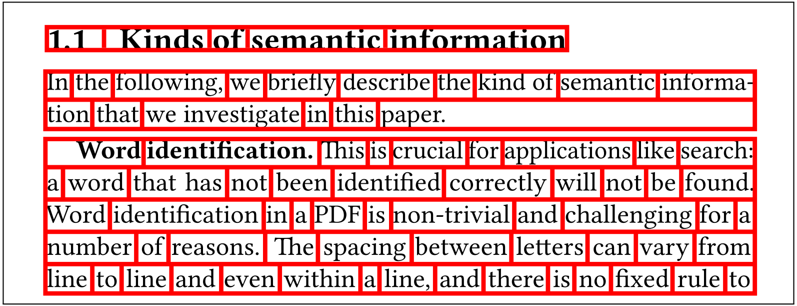
1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

Text block postprocessing

Split down into words

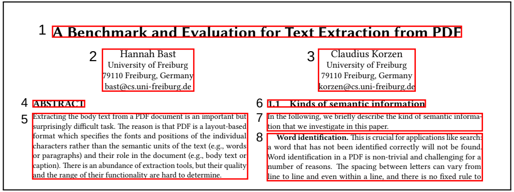


1.1 Kinds of semantic information

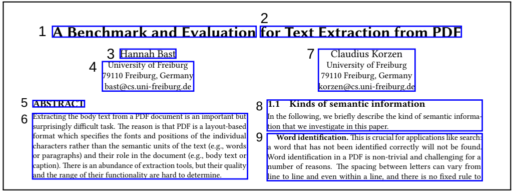
In the following, we briefly describe the kind of semantic information that we investigate in this paper.

Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

$B_G^{=}$ and $B_A^{=}$

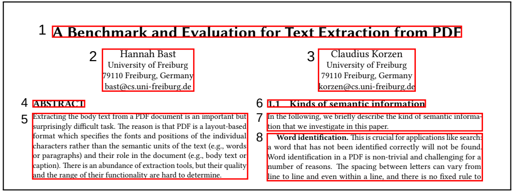


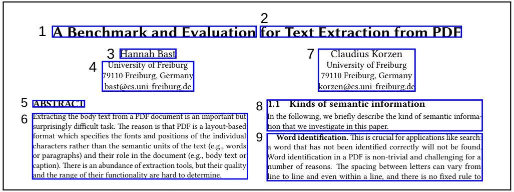
1 A Benchmark and Evaluation for Text Extraction from PDF
2 Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de
3 Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de
4 ABSTRACT
5 Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.
6 1.1 Kinds of semantic information
7 In the following, we briefly describe the kind of semantic information that we investigate in this paper.
8 Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to



1 A Benchmark and Evaluation
2 for Text Extraction from PDF
3 Hannah Bast
4 University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de
7 Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de
5 ABSTRACT
6 Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.
8 1.1 Kinds of semantic information
In the following, we briefly describe the kind of semantic information that we investigate in this paper.
9 Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

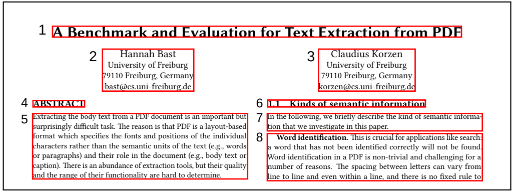
$B_G^{=}$ and $B_A^{=}$

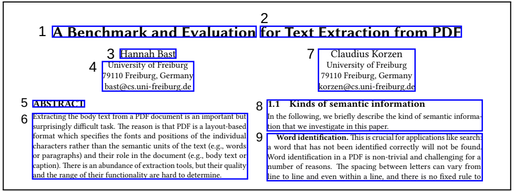




- $\#exp = 8$

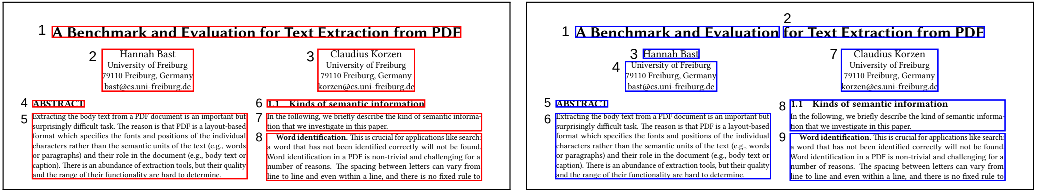
$B_G^{=}$ and $B_A^{=}$





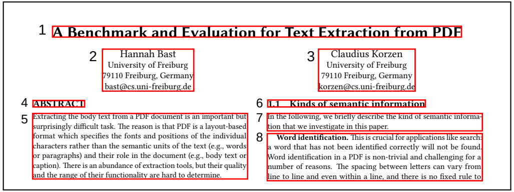
- $\#exp = 8, \#det = 9$

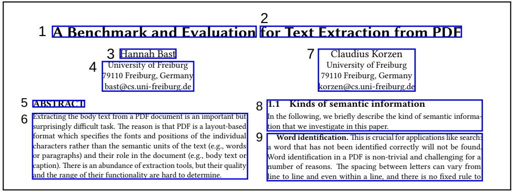
$B_G^=$ and $B_A^=$



- $\#exp = 8, \#det = 9, \#cor = 4$

$B_G^{=}$ and $B_A^{=}$



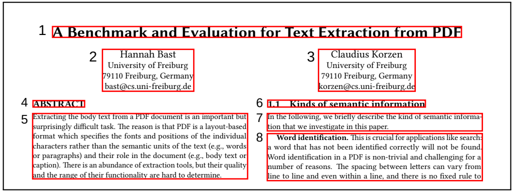


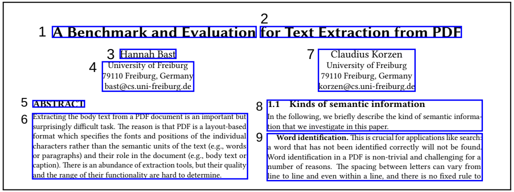
- $\#exp = 8, \#det = 9, \#cor = 4$

$$B_G^{=} = \frac{\#cor}{\#exp} = \frac{4}{8} = 0.5 \mathrel{\widehat{=}} 50\%$$

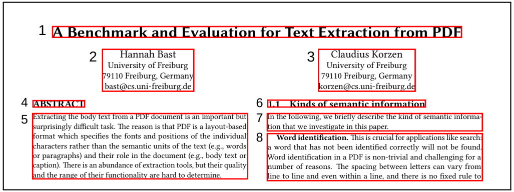
$$B_A^{=} = \frac{\#cor}{\#det} = \frac{4}{9} = 0.\overline{4} \mathrel{\widehat{=}} 44.\overline{4}\%$$

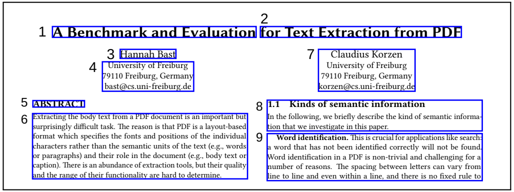
B_G^+ and B_A^-


1 A Benchmark and Evaluation for Text Extraction from PDF
2 Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de
3 Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de
4 ABSTRACT
5 Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.
6 1.1 Kinds of semantic information
7 In the following, we briefly describe the kind of semantic information that we investigate in this paper.
8 Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to


1 A Benchmark and Evaluation
2 for Text Extraction from PDF
3 Hannah Bast
4 University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de
7 Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de
5 ABSTRACT
6 Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.
8 1.1 Kinds of semantic information
In the following, we briefly describe the kind of semantic information that we investigate in this paper.
9 Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

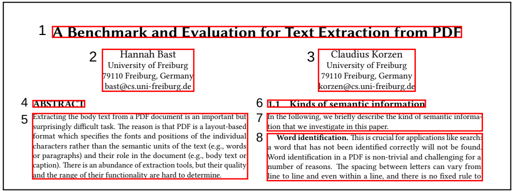
B_G^+ and B_A^-

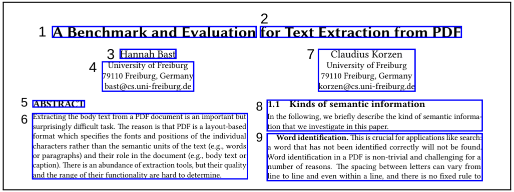




- $\#exp = 8$

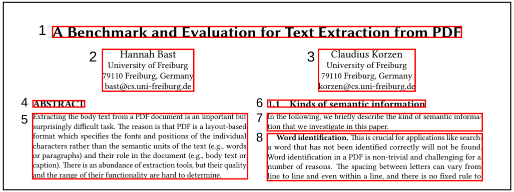
B_G^+ and B_A^-

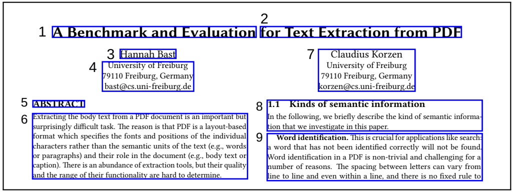




- $\#exp = 8, \#det = 9$

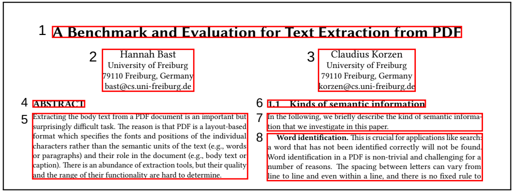
B_G^+ and B_A^-

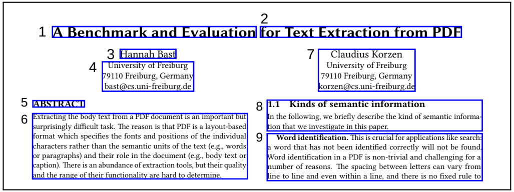




- $\#exp = 8, \#det = 9, \#stm = 2$

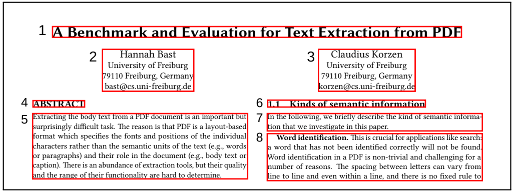
B_G^+ and B_A^-

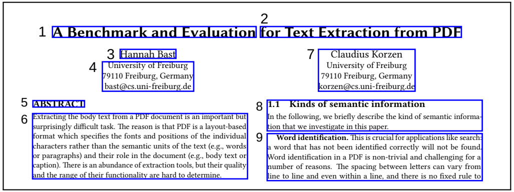




- $\#exp = 8, \#det = 9, \#stm = 2, \#stl = 1$

B_G^+ and B_A^-



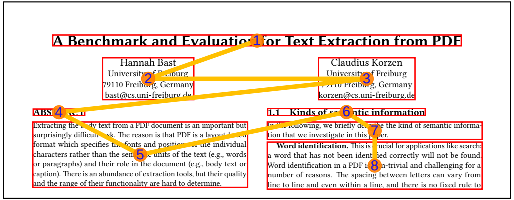


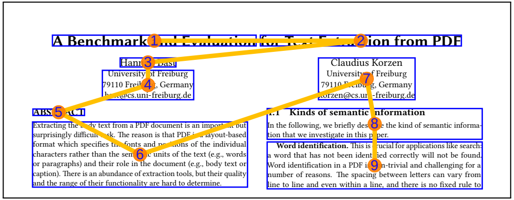
- $\#exp = 8, \#det = 9, \#stm = 2, \#stl = 1$

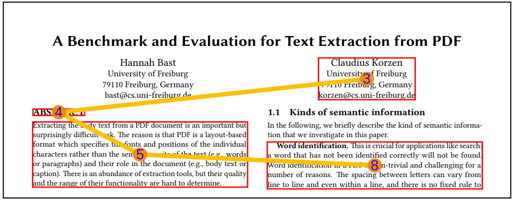
$$B_G^+ = \frac{\#stm}{\#exp} = \frac{2}{8} = 0.25 \mathrel{\widehat{=}} 25\%$$

$$B_A^- = \frac{\#stl}{\#det} = \frac{1}{9} = 0.\overline{1} \mathrel{\widehat{=}} 11.\overline{1}\%$$

τ_n







A Benchmark and Evaluation for Text Extraction from PDF

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@cs.uni-freiburg.de

Claudius Korzen
University of Freiburg
79110 Freiburg, Germany
korzen@cs.uni-freiburg.de

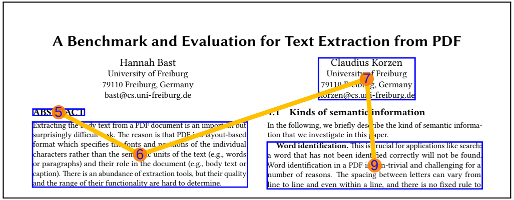
ABSTRACT

Extracting the body text from a PDF document is an important but surprisingly difficult task. The reason is that PDF is a layout-based format which specifies the fonts and positions of the individual characters rather than the semantic units of the text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There is an abundance of extraction tools, but their quality and the range of their functionality are hard to determine.

1.1 Kinds of semantic information

In the following, we briefly describe the kind of semantic information that we investigate in this paper.

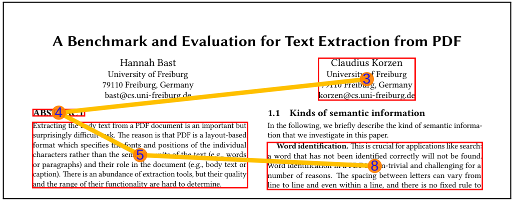
Word identification. This is crucial for applications like search: a word that has not been identified correctly will not be found. Word identification in a PDF is non-trivial and challenging for a number of reasons. The spacing between letters can vary from line to line and even within a line, and there is no fixed rule to

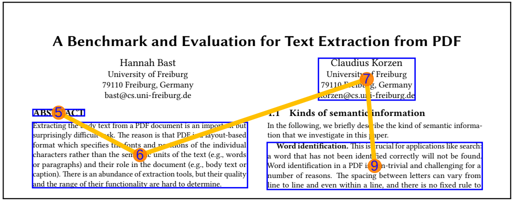


A Benchmark and Evaluation for Text Extraction from PDF

- Detected sequence:

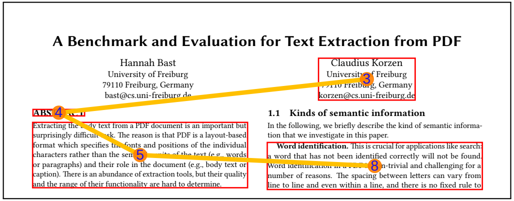
τ_n

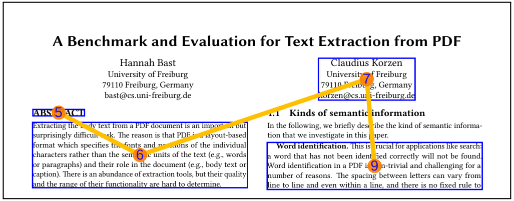




- Detected sequence: 7

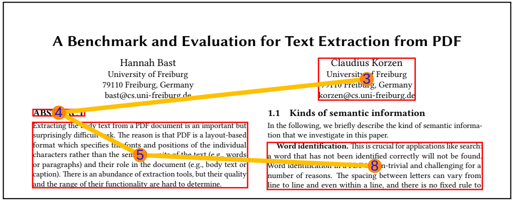
τ_n

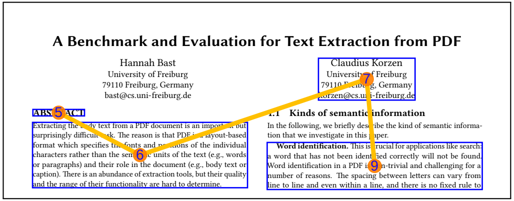




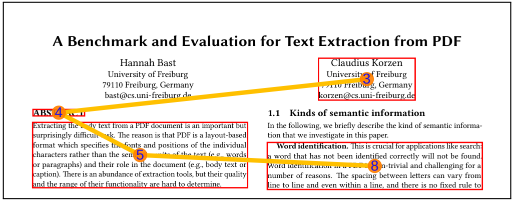
- Detected sequence: 7, 5

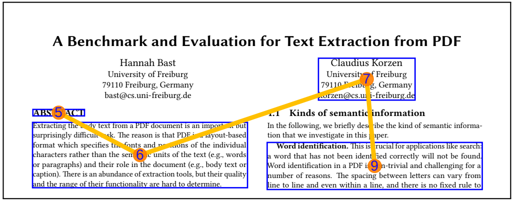
τ_n





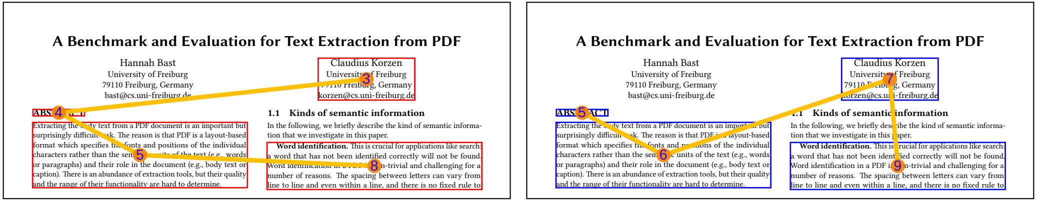
- Detected sequence: 7, 5, 6





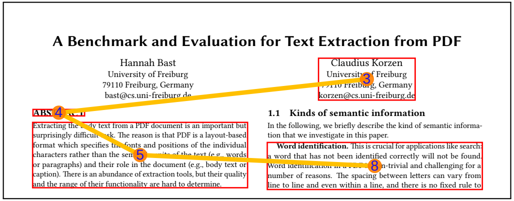
- Detected sequence: 7, 5, 6, 9

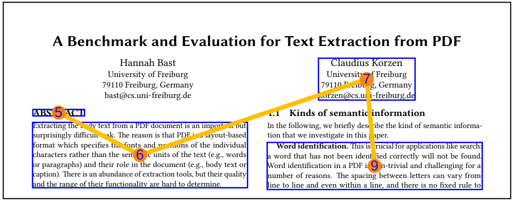
τ_n



- Detected sequence: 7, 5, 6, 9 $\#con = 4$

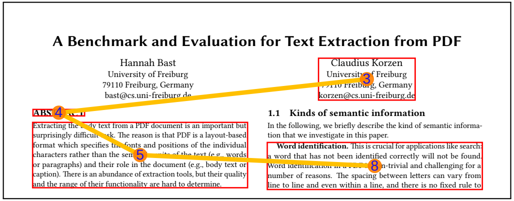
τ_n

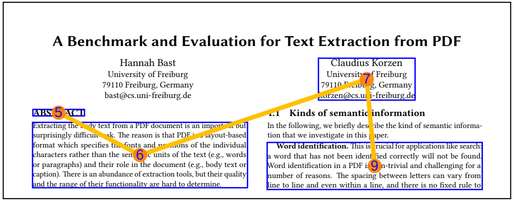




- Detected sequence: 7, 5, 6, 9 $\#con = 4, \#dis = 2$

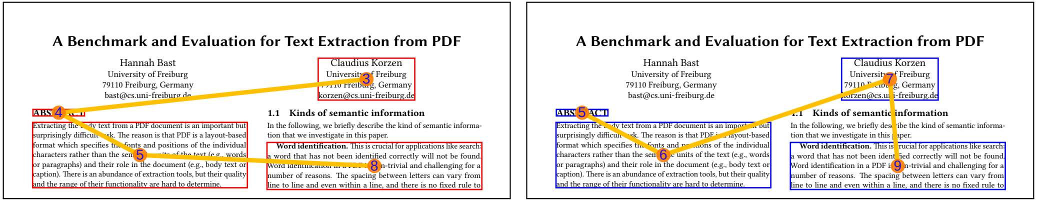
τ_n





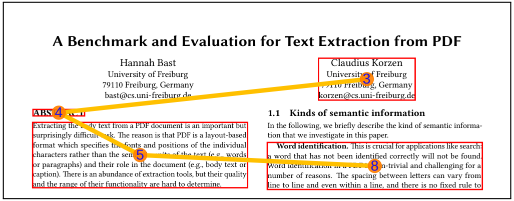
- Detected sequence: 7, 5, 6, 9 $\quad \#con = 4, \#dis = 2 \quad \tau = \frac{1}{3}$

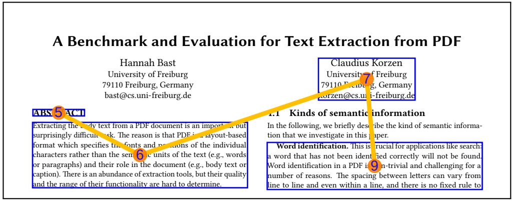
τ_n



- Detected sequence: 7, 5, 6, 9 $\#con = 4, \#dis = 2$ $\tau = \frac{1}{3}$
- We obtain

$$\tau_n = \frac{\tau + 1}{2}$$

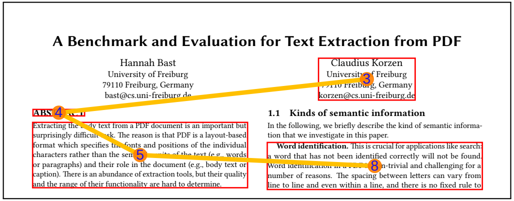


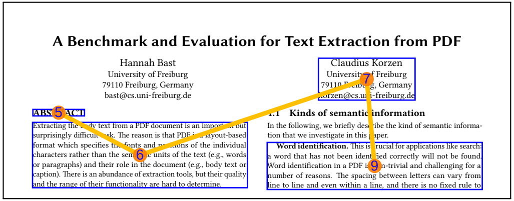


- Detected sequence: 7, 5, 6, 9 $\#con = 4, \#dis = 2$ $\tau = \frac{1}{3}$
- We obtain

$$\tau_n = \frac{\tau + 1}{2} = \frac{1/3 + 1}{2}$$

τ_n

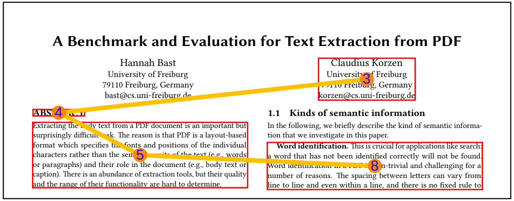


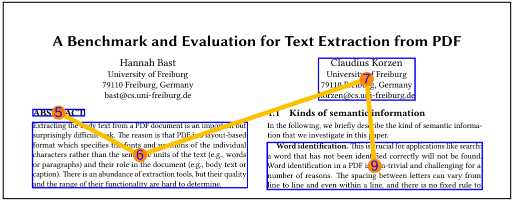


- Detected sequence: 7, 5, 6, 9 $\quad \#con = 4, \#dis = 2 \quad \tau = \frac{1}{3}$
- We obtain

$$\tau_n = \frac{\tau + 1}{2} = \frac{1/3 + 1}{2} = \frac{4/3}{2}$$

τ_n

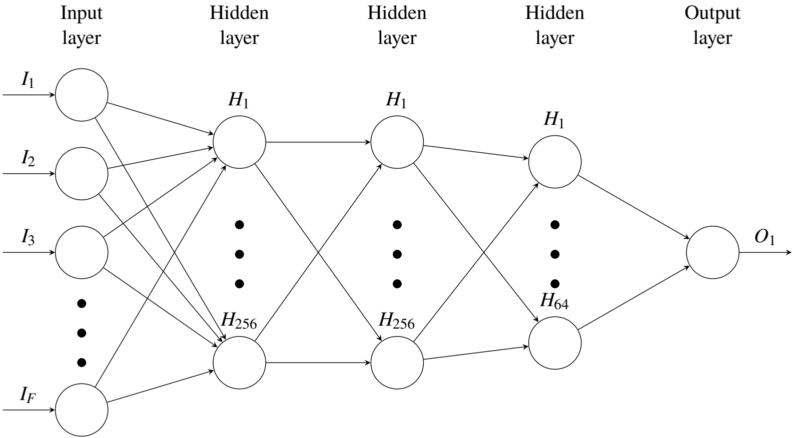




- Detected sequence: 7, 5, 6, 9 $\quad \#con = 4, \#dis = 2 \quad \tau = \frac{1}{3}$
- We obtain

$$\tau_n = \frac{\tau + 1}{2} = \frac{1/3 + 1}{2} = \frac{4/3}{2} = \frac{2}{3}$$

Model architecture



Model input/output

- The input is a matrix whose rows correspond to feature representations of cuts

- The output of the model is a score vector
- We then choose the cut with highest score

Training data format

# name	page num	width,height	subpage	depth	dir
example.pdf	42	612,796	0,0,360,640	2	**X**
# cut	left/upper semantic roles		right/lower semantic roles		label
$([530,550],\mathbf{Y})$	heading		paragraph		1
$([270,290],\mathbf{Y})$	paragraph		paragraph		0
$([170,175],\mathbf{X})$	-		-		0
example.pdf	43	612,796	0,0,612,796	1	-
$([720,740],\mathbf{Y})$	marginal		heading,paragraph		0
$([680,685],\mathbf{Y})$	-		-		0
$([420,430],\mathbf{Y})$	table,paragraph		formula, caption		0
$([296,316],\mathbf{X})$	heading,table,caption		marginal,paragraph,formula		1

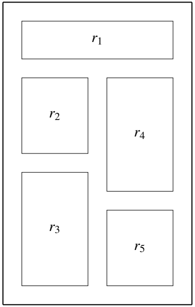
PdfAct comparison

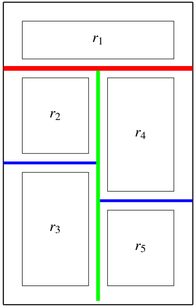
	$B_G^{=}$	$B_A^{=}$	B_G^{+}	B_A^{-}	τ_n	τ_n^f
Thesis	51.4%	46.7%	12.9%	14.7%	0.873	0.994
PdfAct	66.5%	54.3%	10.1%	7.5%	0.859	0.985

Full reading order results

strategy	τ_n	τ_n^f
Largest cut	0.872	0.993
Weighted-largest cut	0.863	0.983
Parameter cut	0.865	0.984
LogisticRegressor	0.873	0.994
BatchClassifier	0.872	0.992
Transformer	0.860	0.978
PdfAct	0.859	0.985

XY-cut limitations for reading order





- Wang et al.'s LayoutReader shows a way to overcome these limitations