

# Significance of Bridging Real-world Documents and NLP Technologies

Tadayoshi Hara

Goran Topić

Yusuke Miyao

Akiko Aizawa

*National Institute of Informatics, Japan*

# Goal of Research

- Analyze any documents with NLP tools

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles.** Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: `DISPUTED`, `TOTALLYDISPUTED`, `DISPUTED=SECTION`, `TOTALLYDISPUTED=SECTION`, `POV`. The tags indicate that an article is disputed, or the neutrality of the article is disputed (`POV`).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (`RFC`) tag on talk pages. According to Wikipedia<sup>4</sup>, `RFC` is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: `CONTROVERSY`, `REQUEST FOR COMMENT (RFC)`, and `RESOLVED` based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	<code>CONTROVERSIAL</code> , <code>TOTALLYDISPUTED</code> , <code>DISPUTED</code> , <code>CALM TALK</code> , <code>POV</code>
Request for Comment	<code>RFC</code>
Resolved	Any tag from above + <code>RESOLVED</code>

Table 1: Subcategory for disputes with corresponding tags. Note that each

ACL 2014 paper  
(XHTML)

# Goal of Research

- Analyze any documents with NLP tools

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles.** Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: `DISPUTED`, `TOTALLYDISPUTED`, `DISPUTED-SECTION`, `TOTALLYDISPUTED-SECTION`, `POV`. The tags indicate that an article is disputed or the neutrality of the article is disputed (`POV`).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (`RFC`) tag on talk pages. According to Wikipedia<sup>4</sup>, `RFC` is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: `CONTROVERSY`, `REQUEST FOR COMMENT (RFC)`, and `RESOLVED` based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	<code>CONTROVERSIAL</code> , <code>TOTALLYDISPUTED</code> , <code>DISPUTED</code> , <code>CALM TALK</code> , <code>POV</code>
Request for Comment	<code>RFC</code>
Resolved	Any tag from above + <code>RESOLVED</code>

Table 1: Subcategory for disputes with corresponding tags. Note that each

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles.** Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: `DISPUTED`, `TOTALLYDISPUTED`, `DISPUTED-SECTION`, `TOTALLYDISPUTED-SECTION`, `POV`. The tags indicate that an article is disputed, or the neutrality of the article is disputed (`POV`).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (`RFC`) tag on talk pages. According to Wikipedia<sup>4</sup>, `RFC` is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: `CONTROVERSY`, `REQUEST FOR COMMENT (RFC)`, and `RESOLVED` based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	<code>CONTROVERSIAL</code> , <code>TOTALLYDISPUTED</code> , <code>DISPUTED</code> , <code>CALM TALK</code> , <code>POV</code>
Request for Comment	<code>RFC</code>
Resolved	Any tag from above + <code>RESOLVED</code>

Table 1: Subcategory for disputes with corresponding tags. Note that each

NLP tools



ACL 2014 paper  
(XHTML)

# Goal of Research

- Analyze any documents with NLP tools

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of 19,071 talk pages and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles** Wikipedia articles are often edited by different editors. If an article is observed to have multiple editors, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: DISPUTED, TOTALLYDISPUTED, DISPUTED-SECTION, TOTALLYDISPUTED-SECTION, POV. The tags indicate that an article is disputed or the neutrality of the article is disputed (POV).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to talk pages themselves. Therefore, in addition to the tags mentioned above, we also consider the "Request for Comment" (RFC) tag on talk pages. According to Wikipedia<sup>4</sup>, RFC is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: CONTROVERSY, REQUEST FOR COMMENT (RFC), and RESOLVED based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	CONTROVERSIAL, TOTALLYDISPUTED, DISPUTED, COMMENT TALK, POV
Request for Comment	RFC
Resolved	Any tag from above + RESOLVED

Table 1: Subcategory for disputes with corresponding tags. Note that each

Real-world text is variously *structuralized*

NLP tools

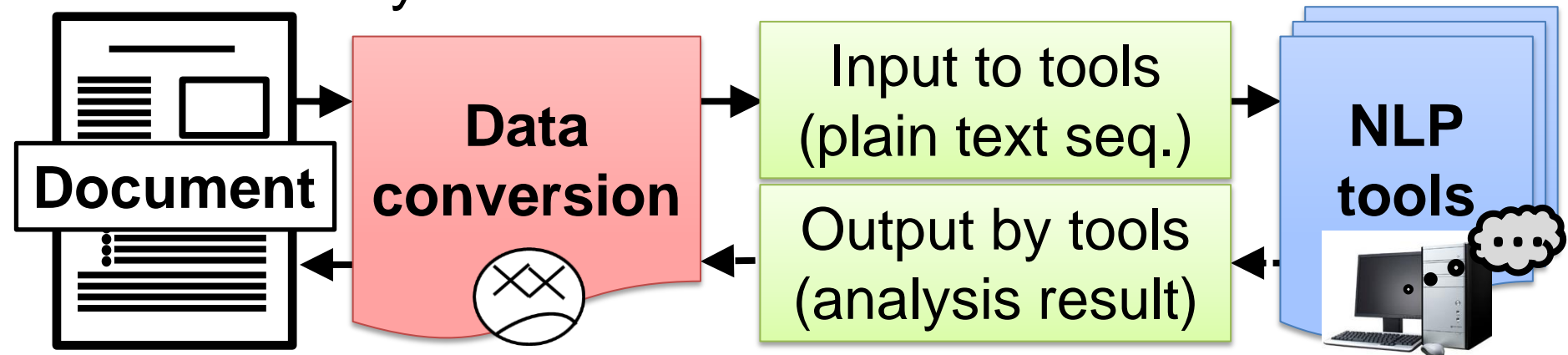


Most NLP tools assume *plain* text as input

Table 1: Subcategory for disputes with corresponding tags. Note that each

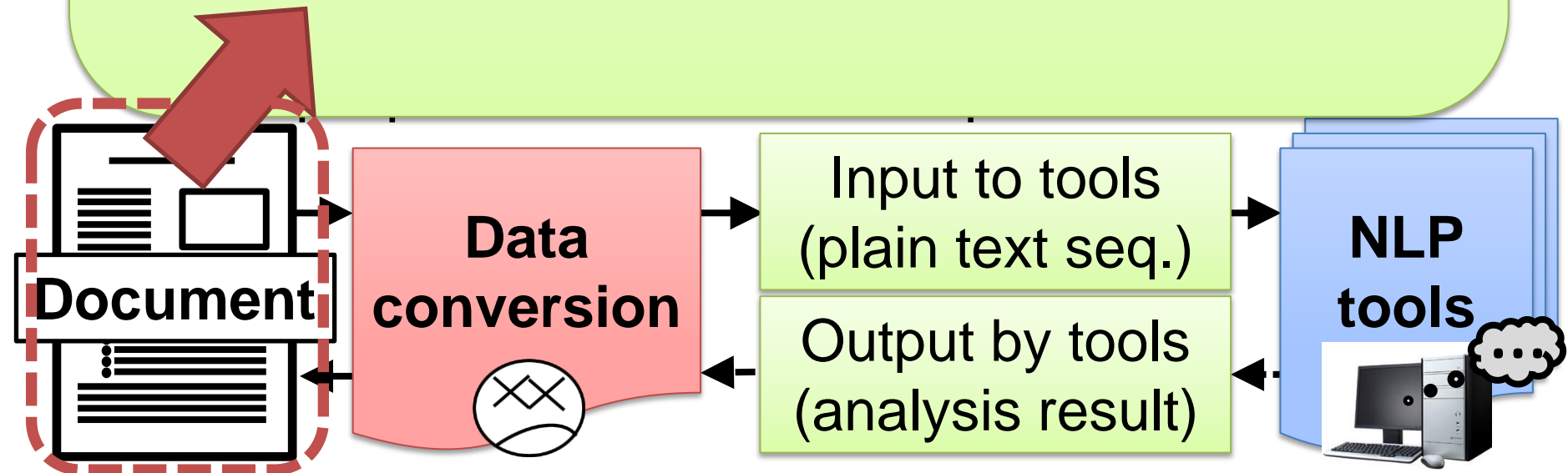
# Goal of Research

- Analyze any documents with NLP tools
  - Real-world text is variously **structuralized**
  - Most NLP tools assume **plain** text as input
- Data conversion is required (up to users)
  - Programming for every target is bothersome
  - Rudely converted text can confuse NLP tools



# Goal of Research

*New UI* is shown. The UI is more useful than XYZ in [3]\*, and ... .  
\*Notice that ... .



# Goal of Research

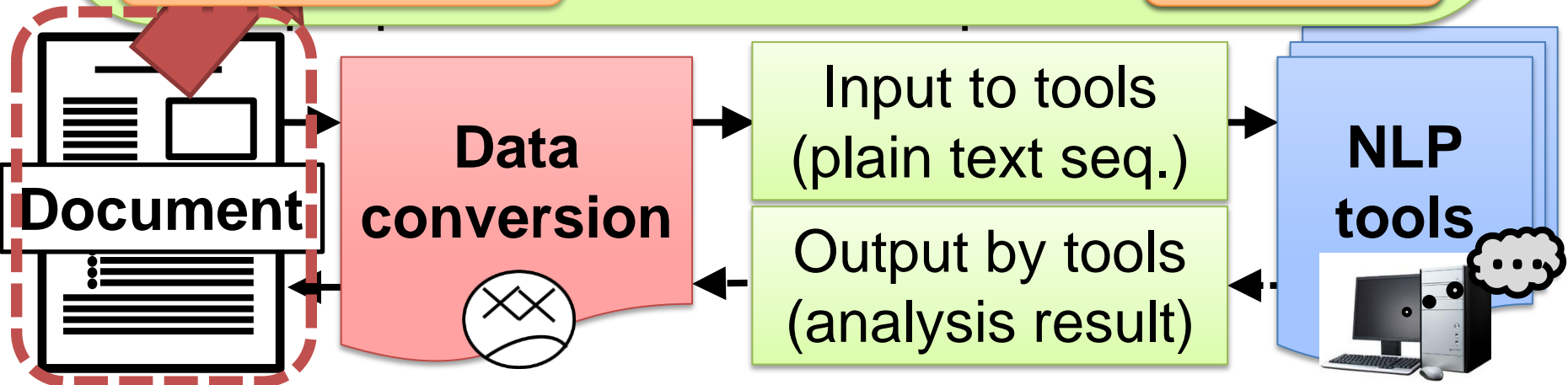
Text style (bold, italic, ... )

Citation link

***New UI*** is shown. The  
UI is more useful than  
XYZ in [3]\*, and ... .  
\*Notice that ... .

Index marker

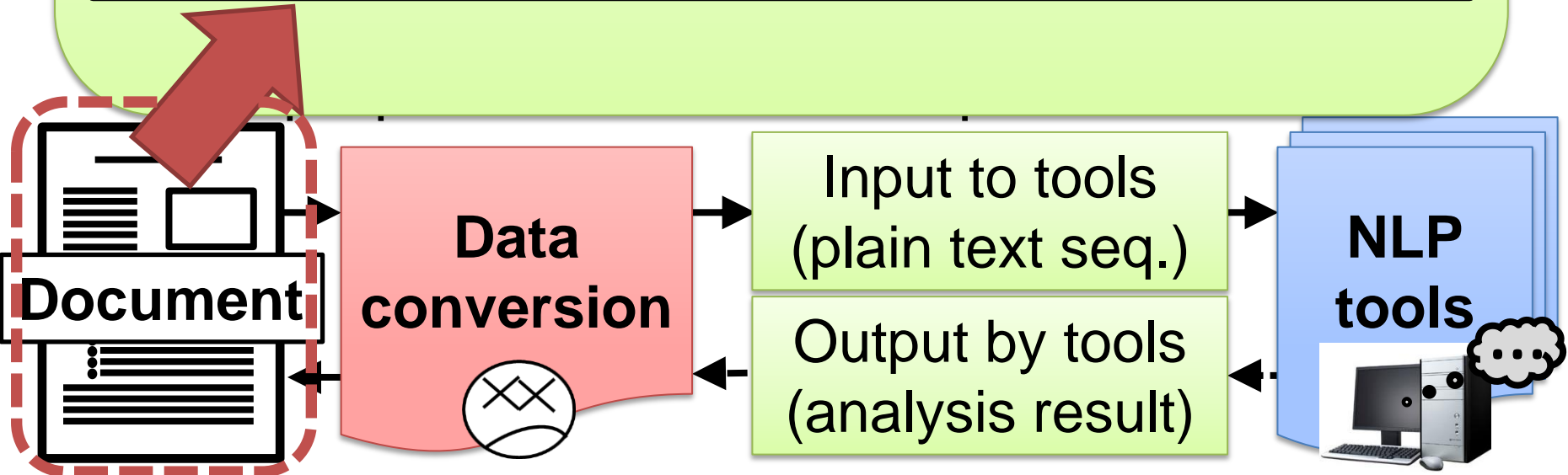
Footnote



# Goal of Research

Tagged text

**<text>**New UI**</text>** is shown. The UI is more useful than XYZ**<index>##</index>** in **<cite>[...]</cite>****<note>**Notice that ...**</note>**, and ...

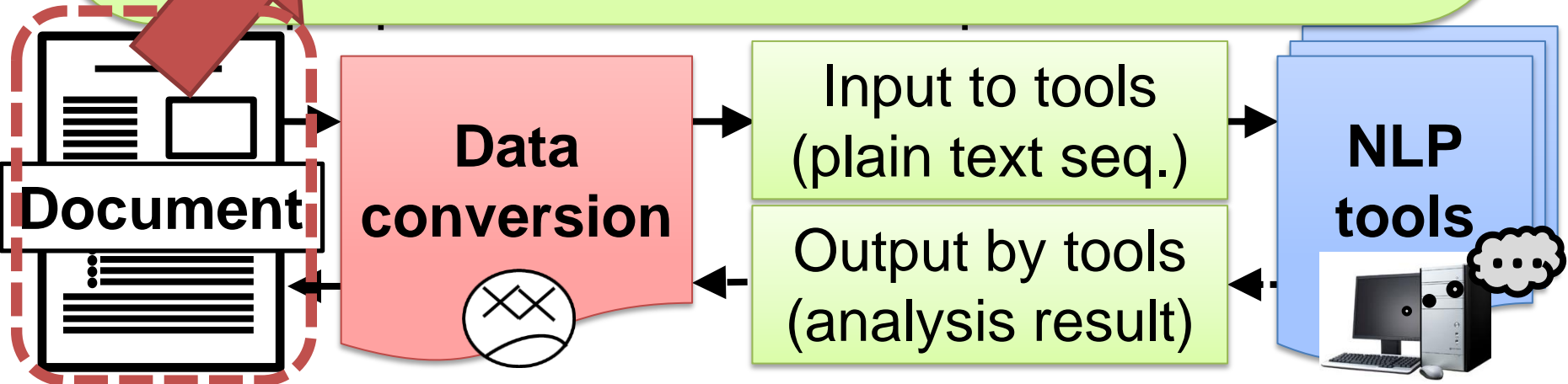




# Goal of Research

Plain (?) text

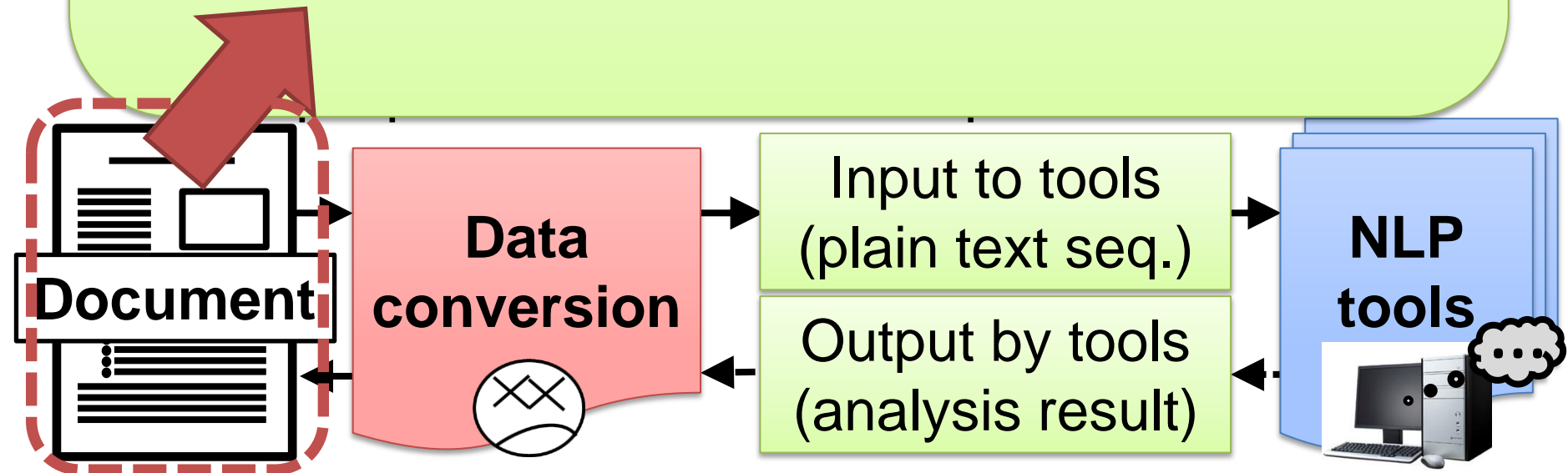
New UI is shown. The UI is  
more useful than XYZ ##  
in [...] Notice that ...  
. , and ...



# Goal of Research

Plain (?) text

New UI is shown. The UI is more useful than XYZ## in [...] Notice that ... ., and ...



# Goal of Research

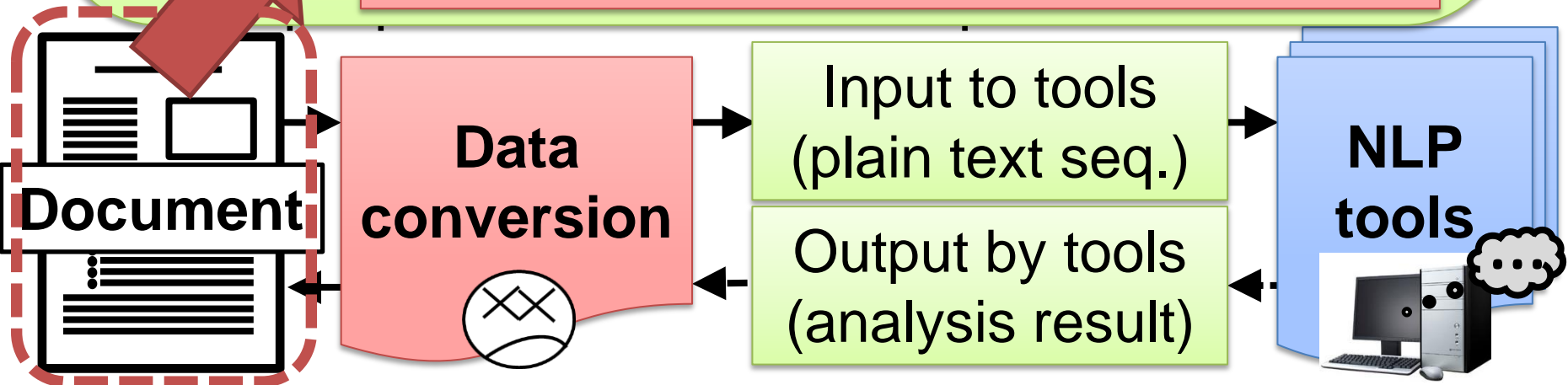
Plain (?) text

New UI is shown. The UI is more useful than XYZ## in [...] Notice that ... , and ...

Non-target fragments

Embedded sentences

Non-natural language (NL) structures



# Goal of Research

Plain (?) text

New UI is shown. The UI is more useful than XYZ## in [...] Notice that ... , and ...

Non-target fragments

Embedded sentences

Non-natural language (NL) structures



Data conversion



Input to tools  
(plain text seq.)

Output by tools  
(analysis result)

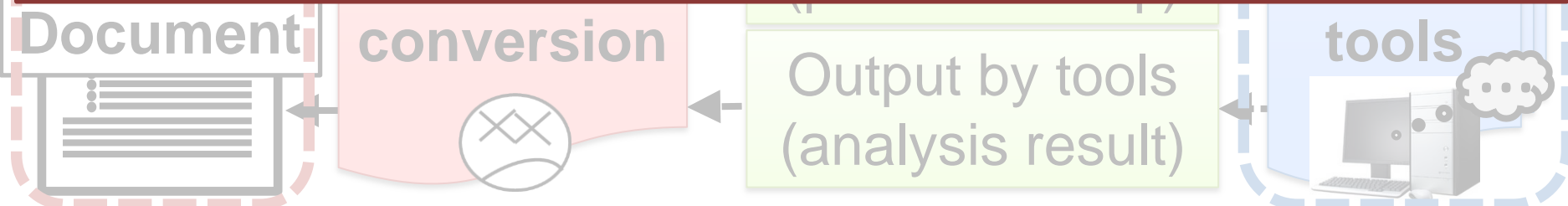
NLP tools



# Goal of Research

- Proper data conversion should be important task for applying NLP tools to real-world documents  
→ Not extensively tackled so far

Promote discussion by demonstrating significance of proper data conversion



# Outline

- **Related work & Our objective**
- **Our framework**
  - Extracting plain text sequences from XML-tagged text based on manual tag classification
- **Experimental results**
  - Extracting plain text sequences from documents
  - Applying parsers to obtained sequences
- **Discussion**
  - Significance of bridging real-world documents and NLP technologies

# Related Work on Unified Methodology for Data Conversion

- (Not extensively tackled so far)
- Some NLP tools provide conversion scripts (e.g. parser with POS-tagger<sup>\*1,2</sup>)
  - Even the scripts assume plain-text input
- Some frameworks enable us to apply various NLP tools to various documents (e.g. UIMA<sup>\*3,4,5,6</sup>/ GATE<sup>\*7</sup>)
  - Tools should be incorporated beforehand

---

<sup>\*1</sup> C&C (Clark et al., 2007), <sup>\*2</sup> Enju (Ninomiya et al., 2007), <sup>\*3</sup> Ferruci et al.(2006)

<sup>\*4</sup> RASP4UIMA (Andersen et al., 2008), <sup>\*5</sup> U-compare (Kano et al., 2011)

<sup>\*6</sup> Kachako (Kano, 2012), <sup>\*7</sup> Cunningham et al.(2013)

# Objective & Approach

## Objective:

Show significance of proper data conversion

## Approach:

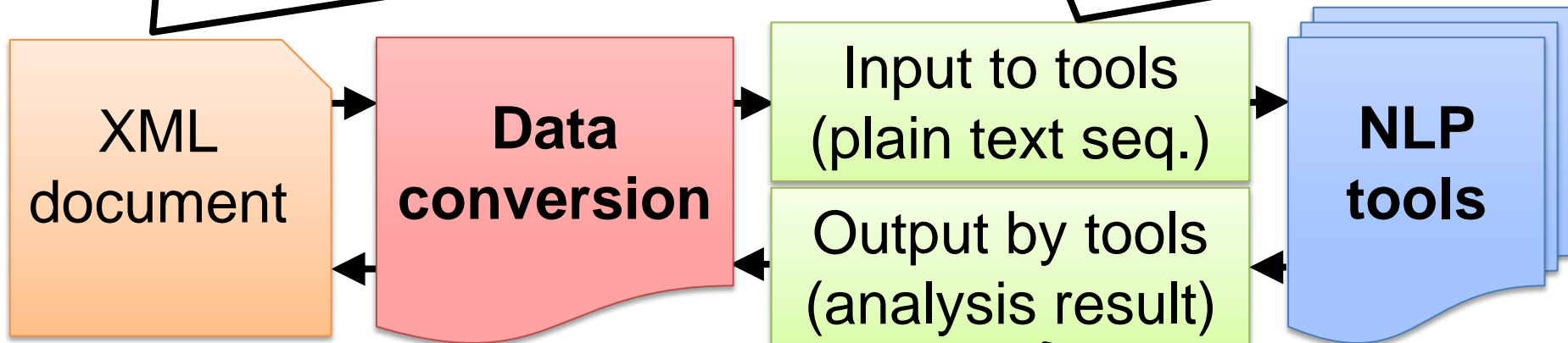
1. Focus on XML documents
  - XML-tags provide structures beyond plain text
2. Propose framework for applying NLP tools to documents without modifying the tools
  - Exemplify impact through experiments



# Overview of Our Framework

<p> In our case, we use the CTT (Concur Task Tree) <cite>[<bibref bibrefs="paterno-ctte-2001"/>]</cite>.</p>

In our case, we use the CTT (Concur Task Tree) [1].

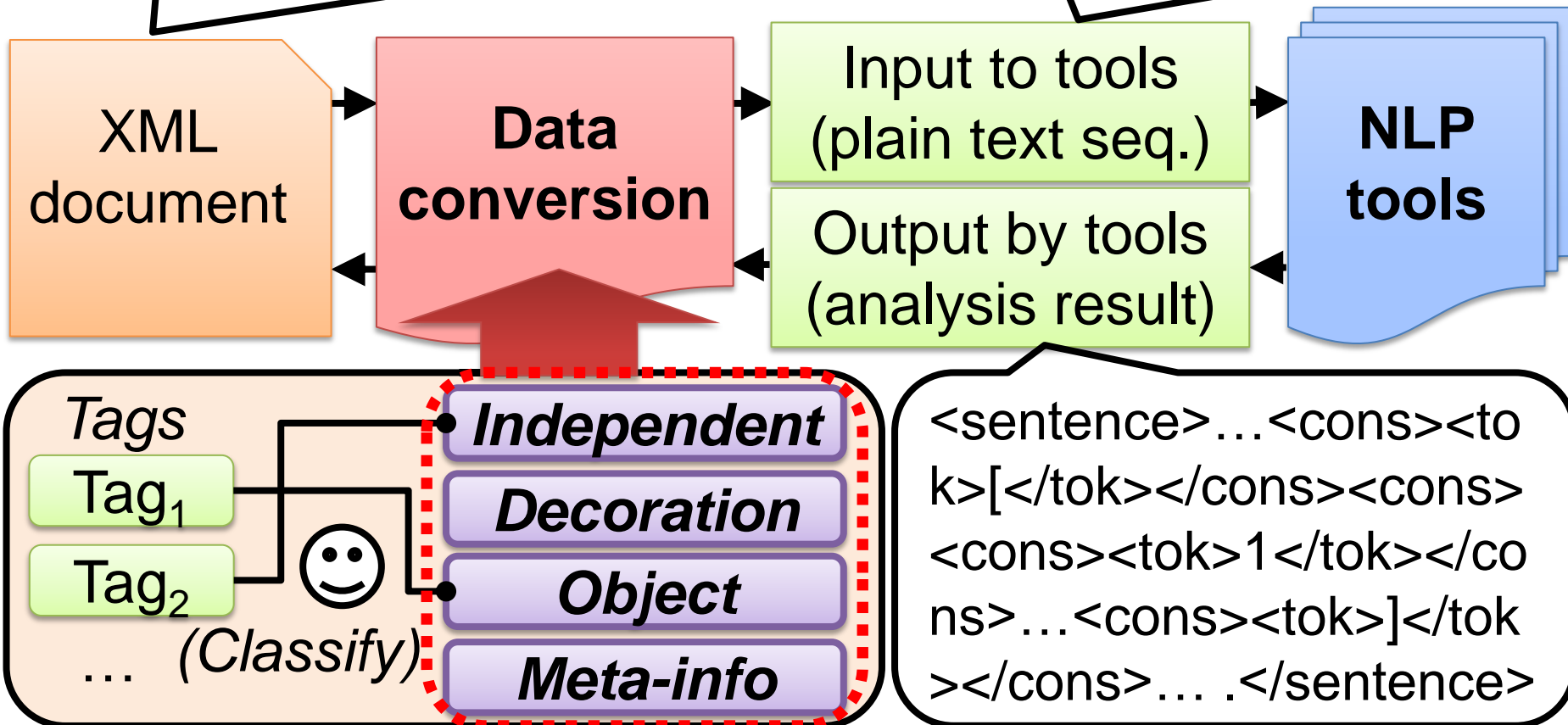


<sentence>...<cons><tok>[</tok></cons><cons><cons><tok>1</tok></cons>...<cons><tok>]</tok></cons>...</sentence>

# Overview of Our Framework

<p> In our case, we use the CTT (Concur Task Tree) <cite>[<bibref bibref s="paterno-ctte-2001"/>]</cite>.</p>

In our case, we use the CTT (Concur Task Tree) [1].



**Classify into 4 functional types → auto-conversion**

# Four Types of Textual Functions

Set display style of enclosed region

***Decoration***

Describe some settings or additional info. (not displayed)

***Meta-info***

`<text>New UI</text>` is shown. The UI is more useful than XYZ `<indexmark>...</indexmark>` in `<cite>[...]</cite>` `<note>Notice that ... </note>` and ...

***Object***

Enclose object consisting of non-NL construction

***Independent***

Enclose syntactically independent region

# Intuition for Applying NLP Tools

Tags seem to be ignorable

**Decoration**

Tagged regions do not seem targets of analysis

**Meta-info**

`<text>New UI</text>` is shown. The UI is more useful than XYZ `<indexmark>...</indexmark>` in `<cite>[...]</cite>` `<note>Notice that ... </note>` and ...

**Object**

Inside regions do not seem analyzable

**Independent**

It seems better to analyze separately

# Strategies for Conversion (1/3): Conversion into Input for NLP Tools

Remove tag

Remove tag & region

***Decoration***

***Meta-info***

`<text>New UI</text>` is shown. The UI is more useful than XYZ `<indexmark>...</indexmark>` in `<cite>[...]</cite>` `<note>Notice that ... </note>` and ...

***Object***

***Independent***

Replace region with dummy word

Separate region

# Strategies for Conversion (1/3): Conversion into Input for NLP Tools

Retain offset information

Remove tag & region

text

**Meta-info**

New UI is shown. The UI is more useful than XYZ `<indexmark>...</indexmark>` in `<cite>[...]</cite>` `<note>Notice that ... </note>` and ...

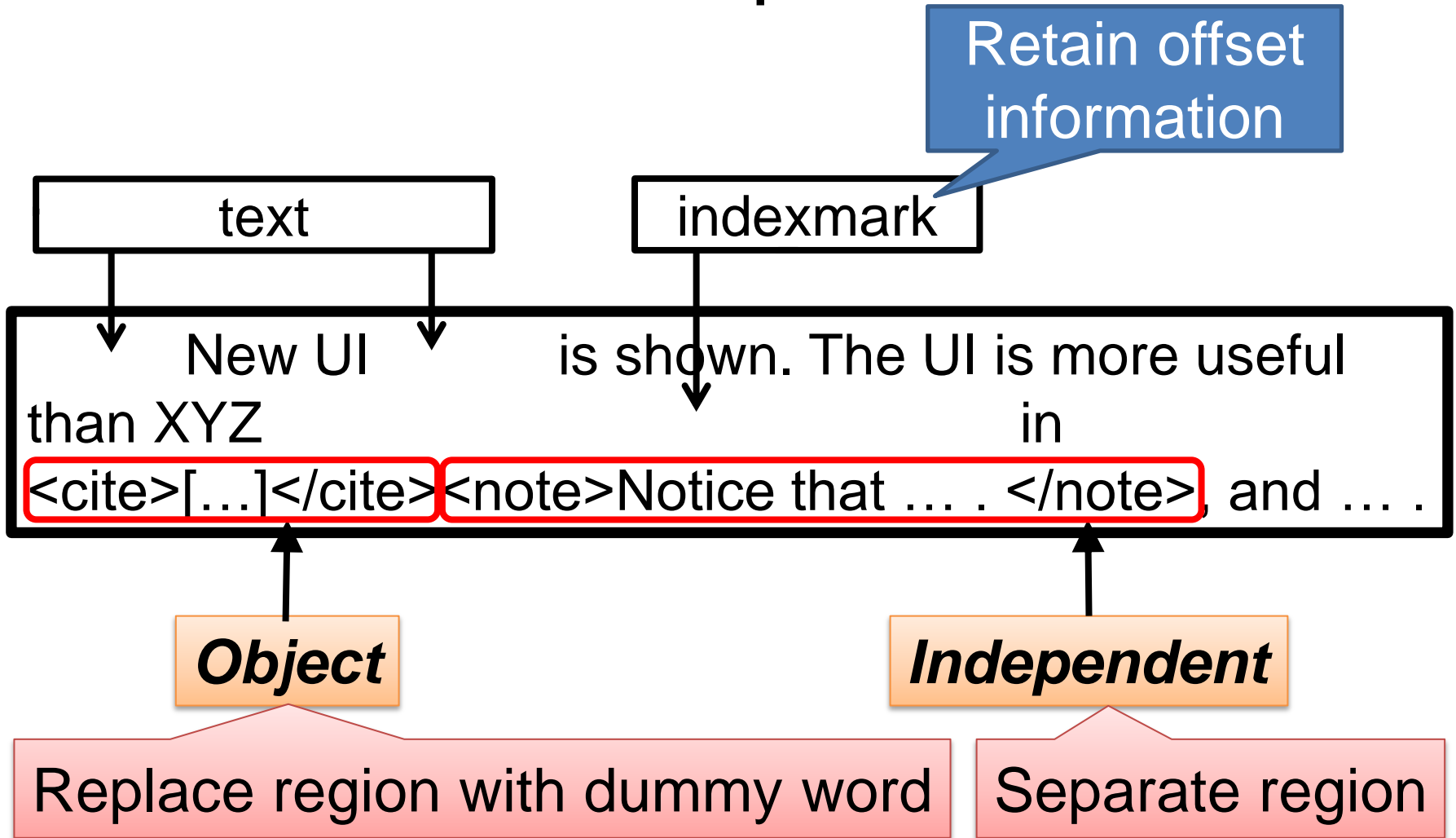
**Object**

**Independent**

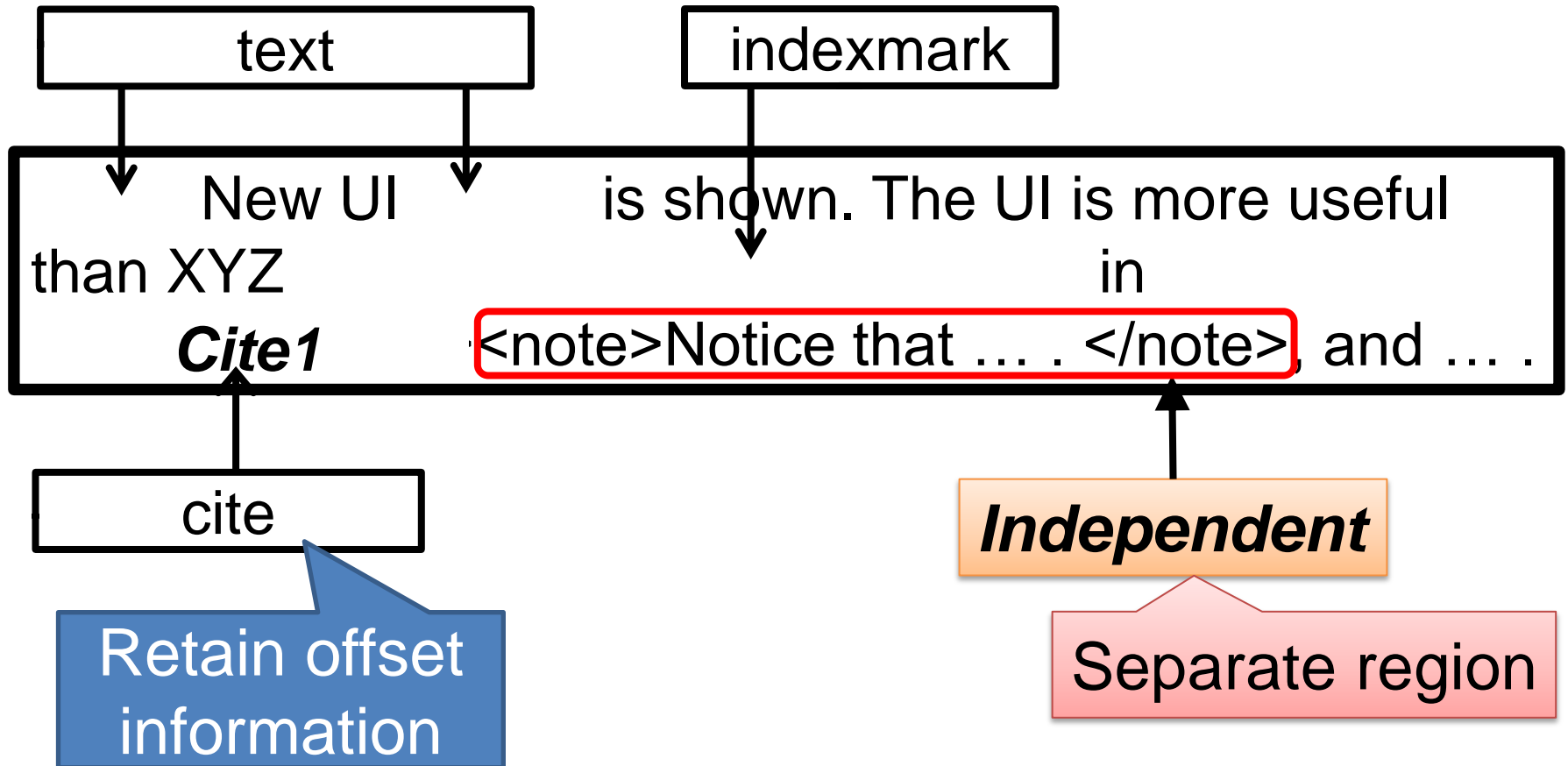
Replace region with dummy word

Separate region

# Strategies for Conversion (1/3): Conversion into Input for NLP Tools

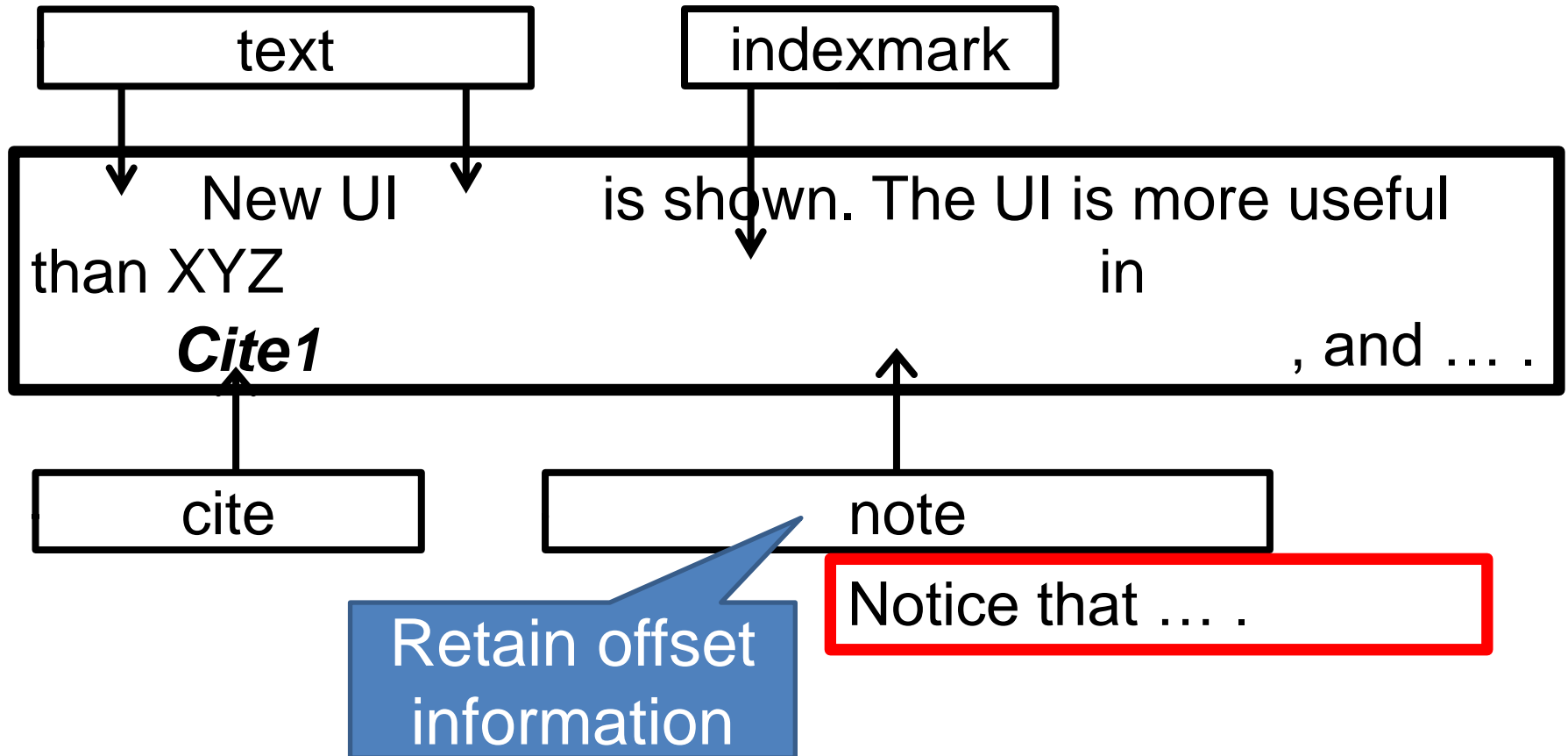


# Strategies for Conversion (1/3): Conversion into Input for NLP Tools

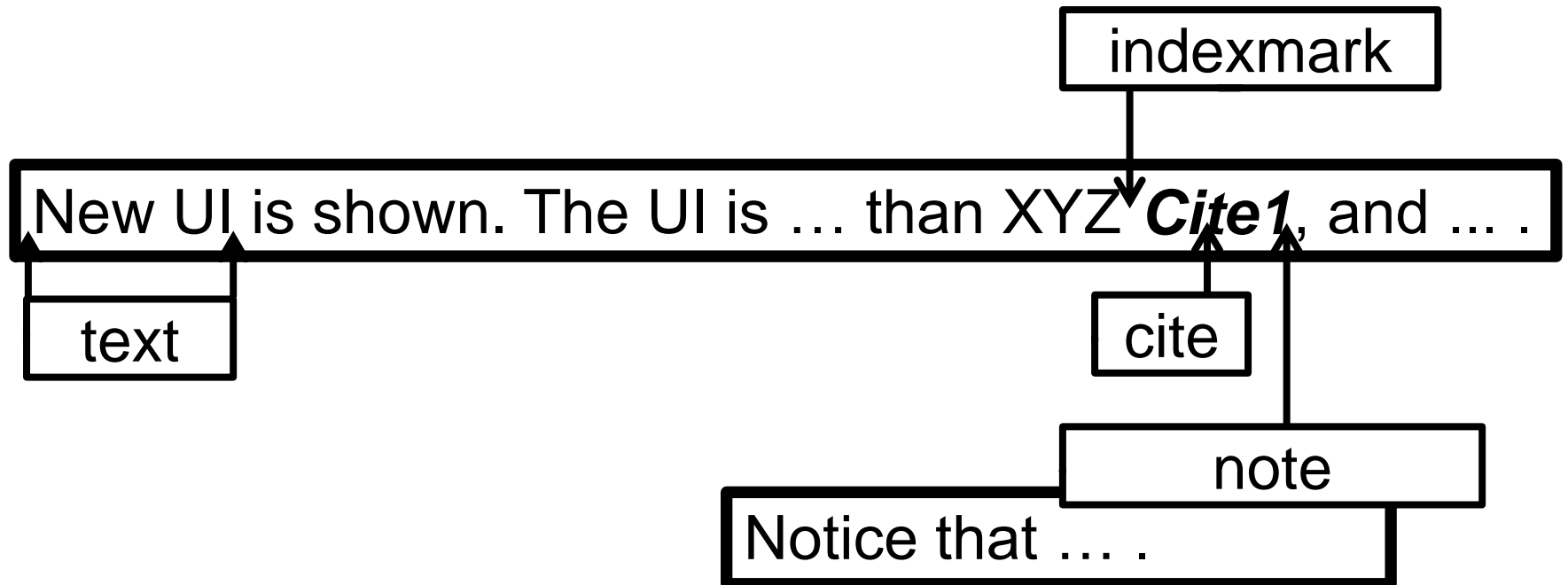




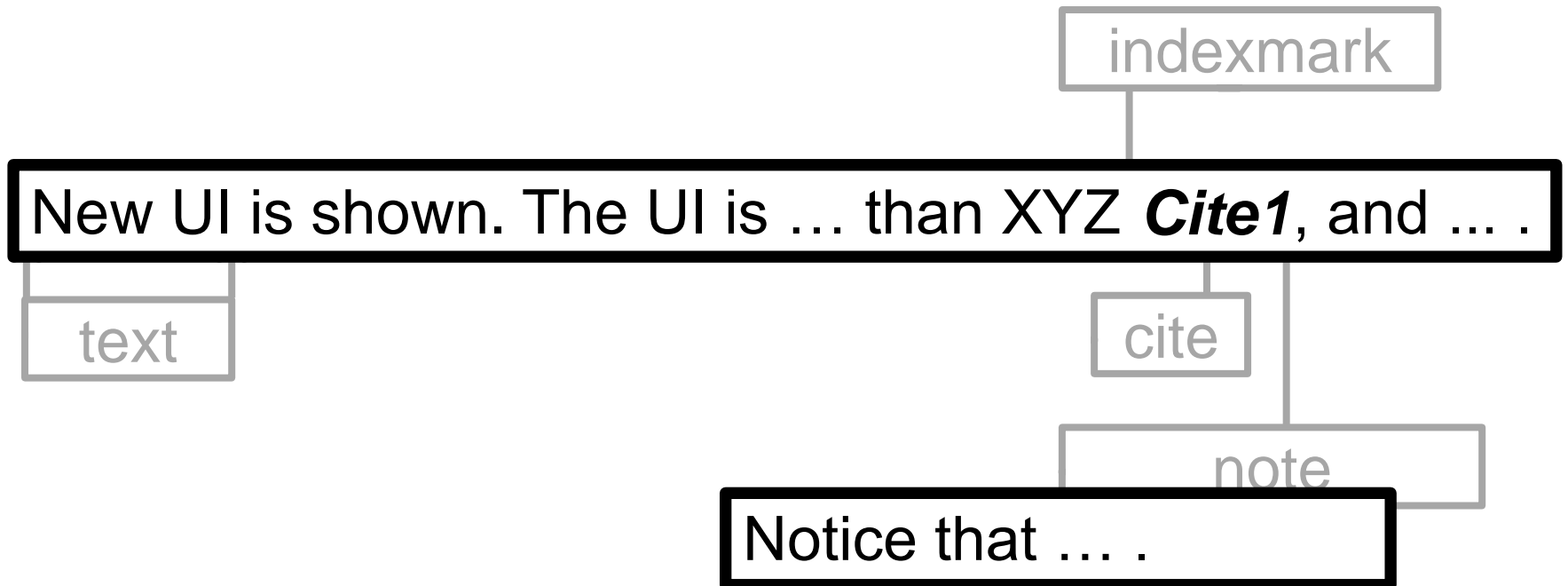
# Strategies for Conversion (1/3): Conversion into Input for NLP Tools



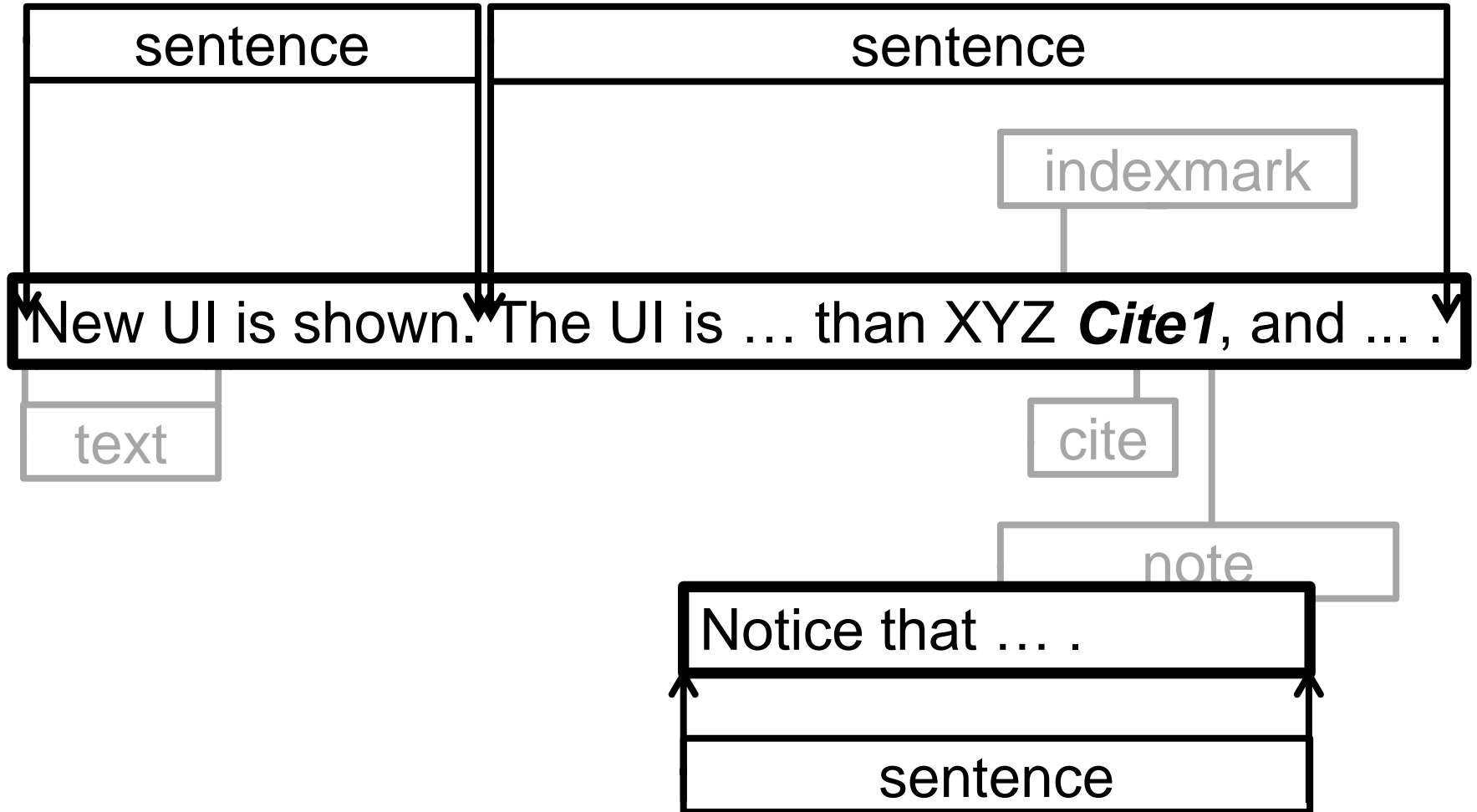
# Strategies for Conversion (1/3): Conversion into Input for NLP Tools



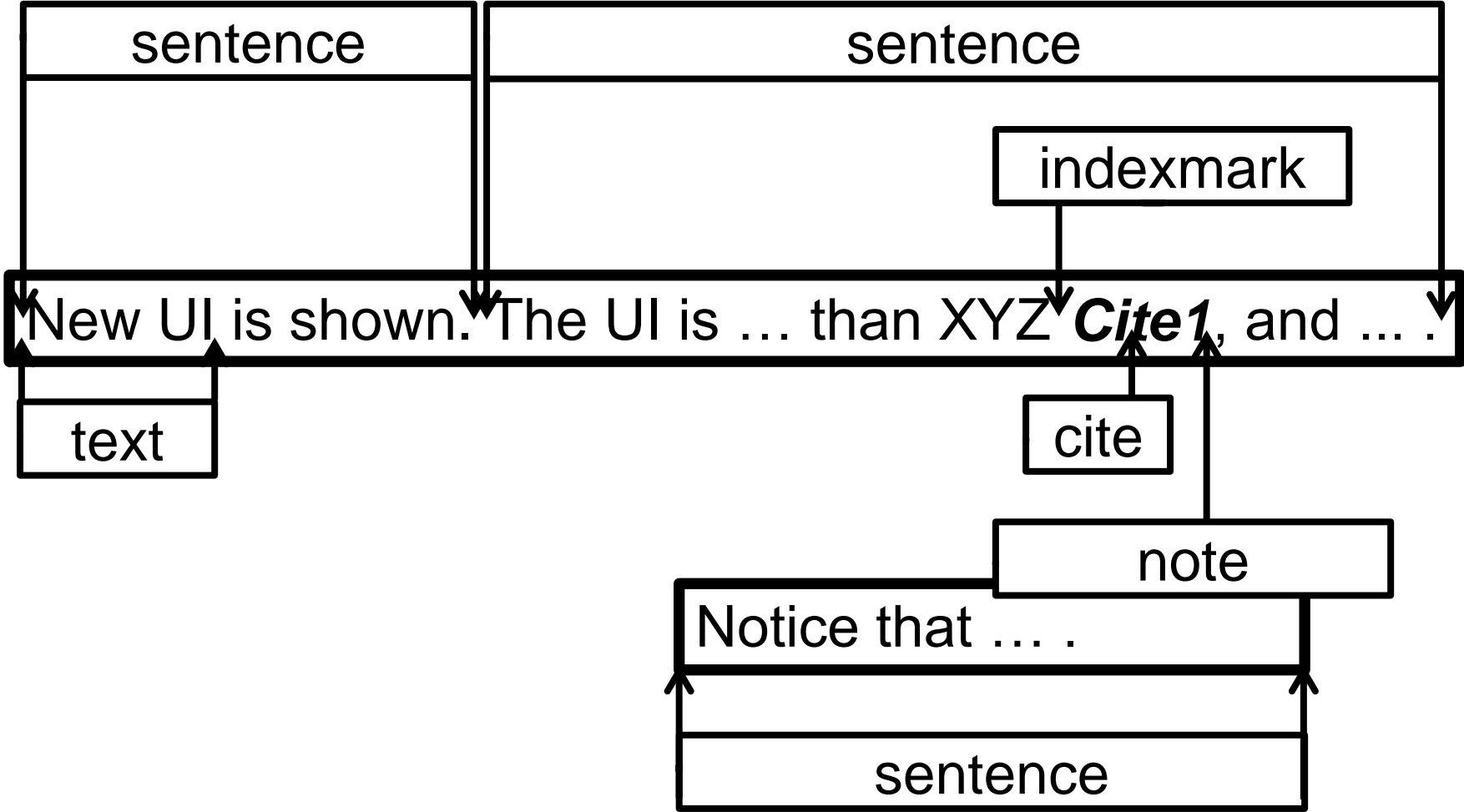
# Strategies for Conversion (1/3): Conversion into Input for NLP Tools



# Strategies for Conversion (2/3): Apply NLP Tools (Sentence Splitter)



# Strategies for Conversion (3/3): Recover Original Tags



# Strategies for Conversion (3/3): Recover Original Tags

```
<sentence><text>New UI</text> is shown.</sentence>  
<sentence>The UI is more useful than XYZ<indexmark  
>...</indexmark> in <cite>[...]</cite><note><sentence>  
Notice that ... . </sentence></note>, and ... .</sentenc  
e>
```

# Goal of Research (Revisit)

- Analyze any (XML) doc. with NLP tools

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles** Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: `DISPUTED`, `TOTALLYDISPUTED`, `DISPUTED-SECTION`, `TOTALLYDISPUTED-SECTION`, `POV`. The tags indicate that an article is disputed or the neutrality of the article is disputed (`POV`).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (`RFC`) tag on talk pages. According to Wikipedia<sup>4</sup>, `RFC` is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: `CONTROVERSIAL`, `REQUEST FOR COMMENT (RFC)`, and `RESOLVED` based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	<code>CONTROVERSIAL</code> , <code>TOTALLYDISPUTED</code> , <code>DISPUTED</code> , <code>CALM TALK</code> , <code>POV</code>
Request for Comment	<code>RFC</code>
Resolved	Any tag from above + <code>RESOLVED</code>

Table 1: Subcategory for disputes with corresponding tags. Note that each

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles.** Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: `DISPUTED`, `TOTALLYDISPUTED`, `DISPUTED-SECTION`, `TOTALLYDISPUTED-SECTION`, `POV`. The tags indicate that an article is disputed, or the neutrality of the article is disputed (`POV`).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes.** Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (`RFC`) tag on talk pages. According to Wikipedia<sup>4</sup>, `RFC` is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: `CONTROVERSIAL`, `REQUEST FOR COMMENT (RFC)`, and `RESOLVED` based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	<code>CONTROVERSIAL</code> , <code>TOTALLYDISPUTED</code> , <code>DISPUTED</code> , <code>CALM TALK</code> , <code>POV</code>
Request for Comment	<code>RFC</code>
Resolved	Any tag from above + <code>RESOLVED</code>

Table 1: Subcategory for disputes with corresponding tags. Note that each

NLP tools



# Goal of Research (Revisit)

- Analyze any (XML) doc. with NLP tools

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles**  
 Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: DISPUTED, TOTALLYDISPUTED, DISPUTED-SECTION, and POV. The tags indicate that an article is disputed, or the neutrality of the article is disputed (POV).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes**  
 Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the "Request for Comment" (RFC) tag on talk pages. According to Wikipedia, RFC is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: CONTROVERSIAL, REQUEST FOR COMMENT (RFC), and RESOLVED based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	CONTROVERSIAL, TOTALLYDISPUTED, DISPUTED, CALM TALK, POV
Request for Comment	RFC
Resolved	Any tag from above + RESOLVED

Table 1: Subcategory for disputes with corresponding tags. Note that each

## 2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

**Step 1: Get Talk Pages of Disputed Articles**  
 Wikipedia articles are edited by different editors. If an article is observed to have disputes on its *talk page*, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: DISPUTED, TOTALLYDISPUTED, DISPUTED-SECTION, and POV. The tags indicate that an article is disputed, or the neutrality of the article is disputed (POV).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

**Step 2: Get Discussions with Disputes**  
 Dispute tags can also be added to *talk pages* themselves. Therefore, in addition to the tags mentioned above, we also consider the "Request for Comment" (RFC) tag on talk pages. According to Wikipedia, RFC is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: CONTROVERSIAL, REQUEST FOR COMMENT (RFC), and RESOLVED based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Dispute Subcategory	Wikipedia Tags on Talk pages
Controversy	CONTROVERSIAL, TOTALLYDISPUTED, DISPUTED, CALM TALK, POV
Request for Comment	RFC
Resolved	Any tag from above + RESOLVED

Table 1: Subcategory for disputes with corresponding tags. Note that each

Our framework

NLP tools





# Summary of Tag Classification & Conversion Strategies

Types	Criteria	Conversion strategies
<b>Independent</b>	Enclose syntactically independent region	Separate (A) from (B) → NLP to (A) & (B) → recover (A) to (B)
<b>Decoration</b>	Set display style of region	Remove (A') from (B) → NLP → recover (A) to (B)
<b>Object</b>	Minimal object unit as text constituent	Replace (A) with (C) → NLP → recover (C) to (A)
<b>Meta-info</b>	Describe settings/ additional info.	Remove (A) from (B) → NLP → recover (A) to (B)

(A): tag&tagged region    (A'): tag    (B): original text  
(C): dummy word

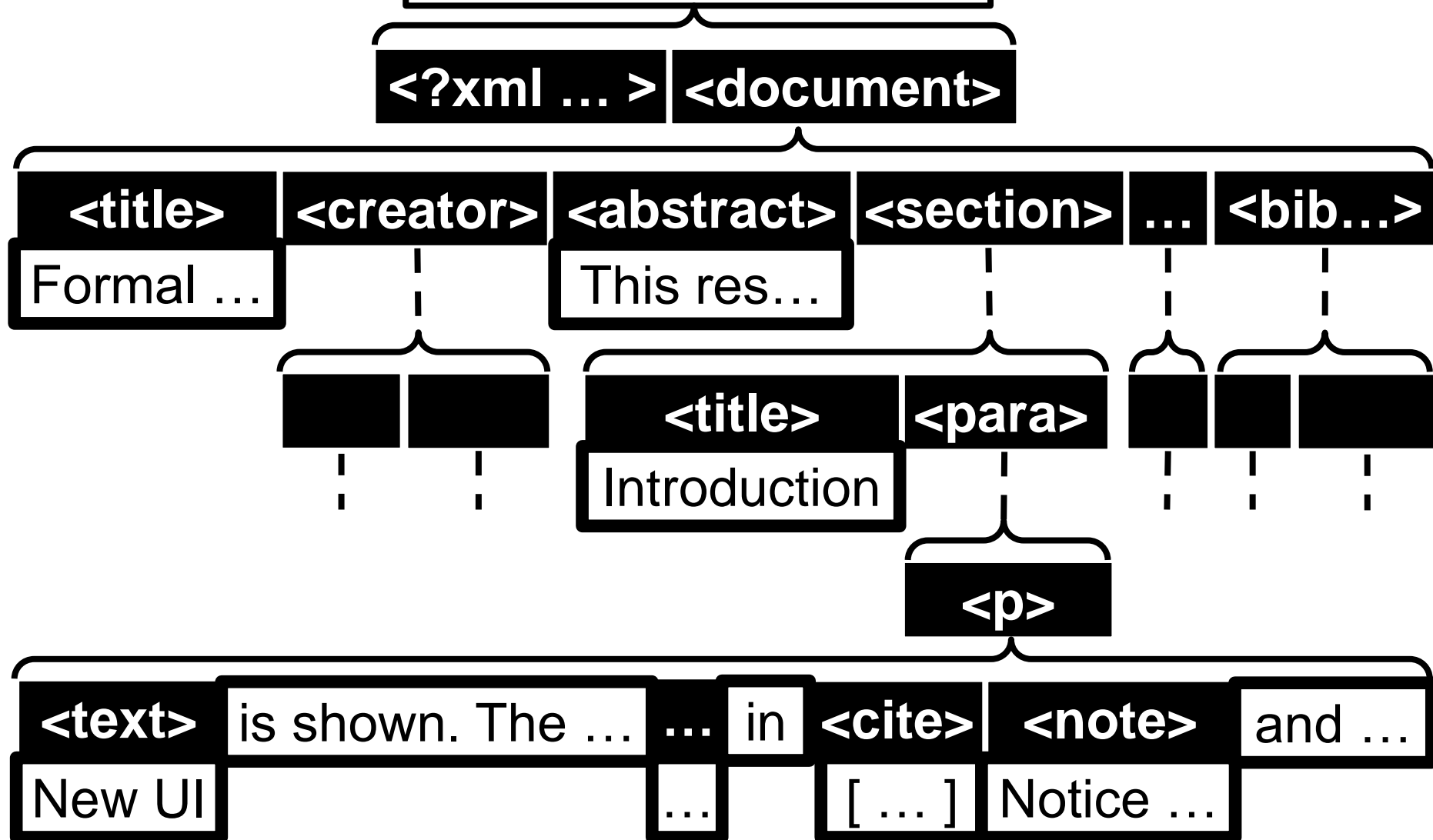
# Efficiency of Tag Classification

XML document

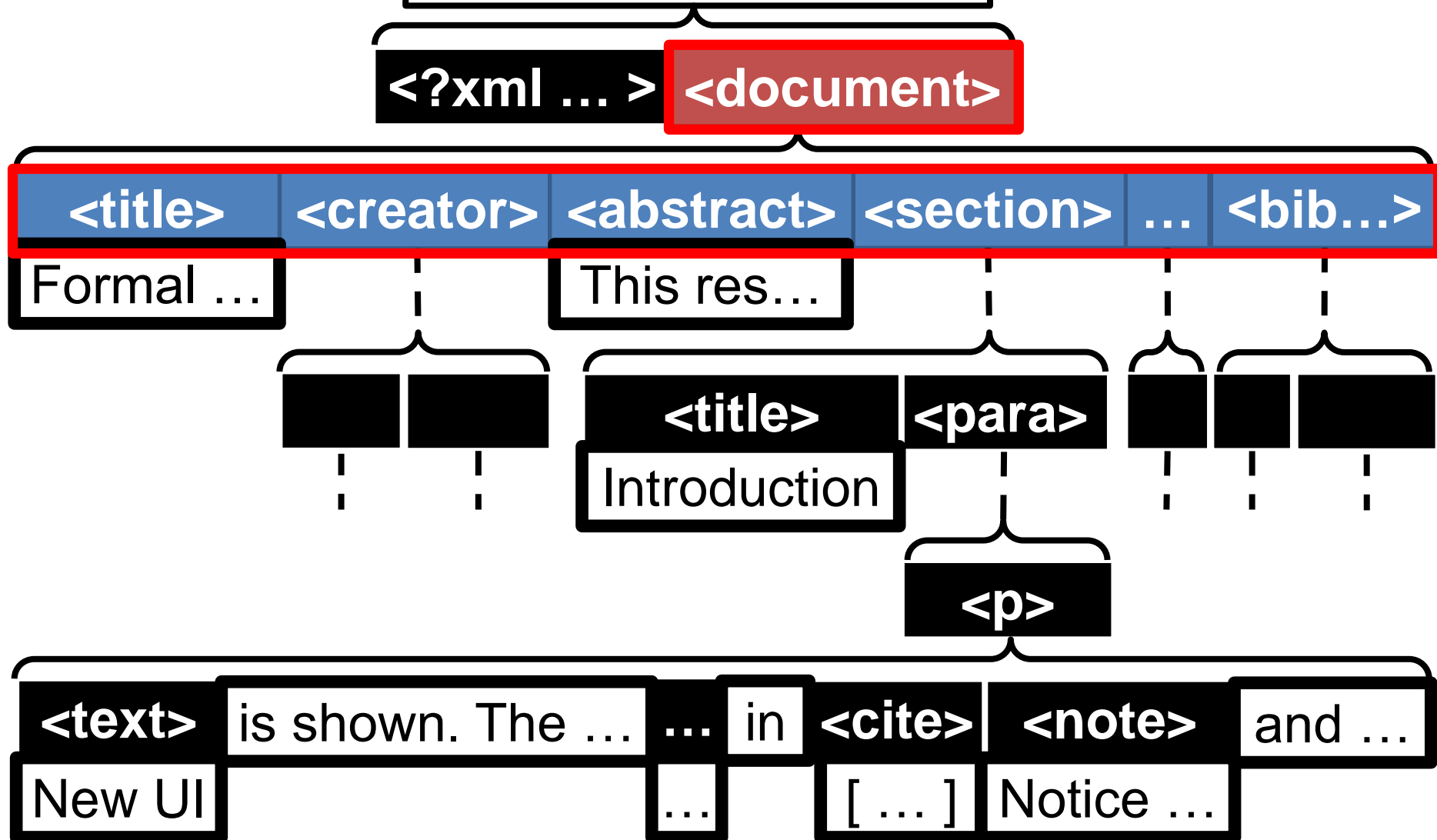
```
<document ...>
  <title>Formal approaches ... </title>
  <creator> ... </creator>
  <abstract><p>This research ... </p></abstract>
  <section><title>Introduction</title>
    <para><p><text>New UI</text> is shown. The
UI is more useful than XYZ<indexmark> ...
</indexmark> in <cite>[ ... ]</cite><note>Notice
that ... </note> and ... .</p></para>
  </section>
  <section> ... </section>
  <bibliography> ... </bibliography>
</document>
```

Classifying via naïve observation is inefficient

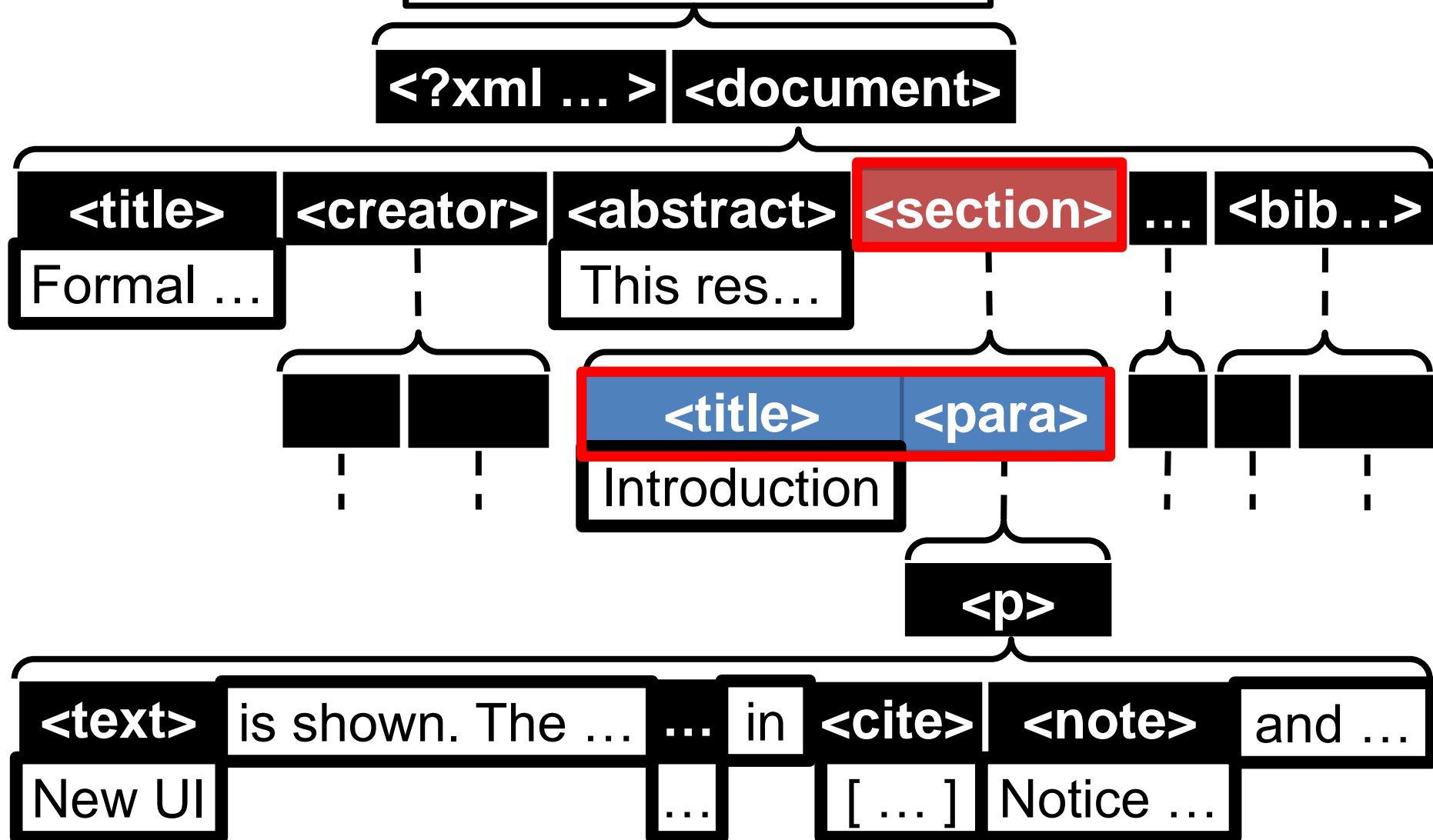
# Multi-layered Structure of XML Document:



# Multi-layered Structure of XML Document:



# Multi-layered Structure of XML Document:



# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document

**<?xml ... >** **<document>**

**<title>**

**<creator>**

**<abstract>**

**<section>**

**...**

**<bib...>**

Formal ...

This res...

**<title>**

**<para>**

Introduction

# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document

`<?xml ... >` `<document>`

`<title>`

Formal ...

`<creator>`

`<abstract>`

This res...

`<section>`

...

`<bib...>`

`<title>`

Introduction

`<para>`

# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

Indep.: `<document>`

XML document



**Classify**

Meta.: `<?xml ... >`

`<?xml ... >` `<document>`

`<title>`

`<creator>`

`<abstract>`

`<section>`

...

`<bib...>`

Formal ...

This res...

`<title>`

`<para>`

Introduction



# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost  
Indep.: `<document>`

XML document

`<?xml ... >` `<document>`

Meta.: `<?xml ... >`



*Unpack & report*

`<title>`

`<creator>`

`<abstract>`

`<section>`

...

`<bib...>`

Formal ...

This res...

`<title>`

`<para>`

Introduction

# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document



**Classify**

Indep.: <document> <section>

<title> <abstract>

Meta.: <?xml ... > <creator>

<bibliography>

<?xml ... > <document>

<title>

<creator>

<abstract>

<section>

...

<bib...>

Formal ...

This res...

<title>

<para>

Introduction

# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document

`<?xml ... >` `<document>`

Indep.: `<document>` `<section>`

`<title>` `<abstract>`

Meta.: `<?xml ... >` `<creator>`

`<bibliography>`

`<title>`

`<creator>`

`<abstract>`

`<section>`

...

`<bib...>`

Formal ...

This res...

`<title>`

`<para>`

Introduction



**Unpack**

# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document

<?xml ... > <document>

Already classified  
→ unpacked automatically

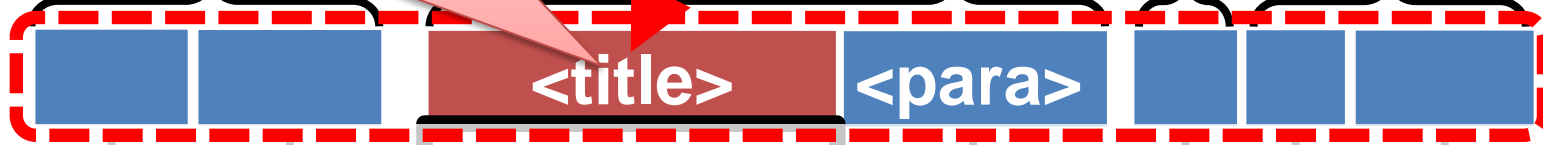
Indep.: <document> <section>

<title> <abstract>

Meta.: <?xml ... > <creator>

<bibliography>

<abstract> <section> ... <bib...>



Introduction

**Unpack**



# Procedure of Tag Classification

- Unpack regions enclosed by already-classified tags from topmost

XML document

<?xml ... > <document>

Already classified  
→ unpacked automatically

Indep.: <document> <section>

<title> <abstract>

Meta.: <?xml ... > <creator>

<bibliography>

<abstract> <section> ... <bib...>

<title>

<para>

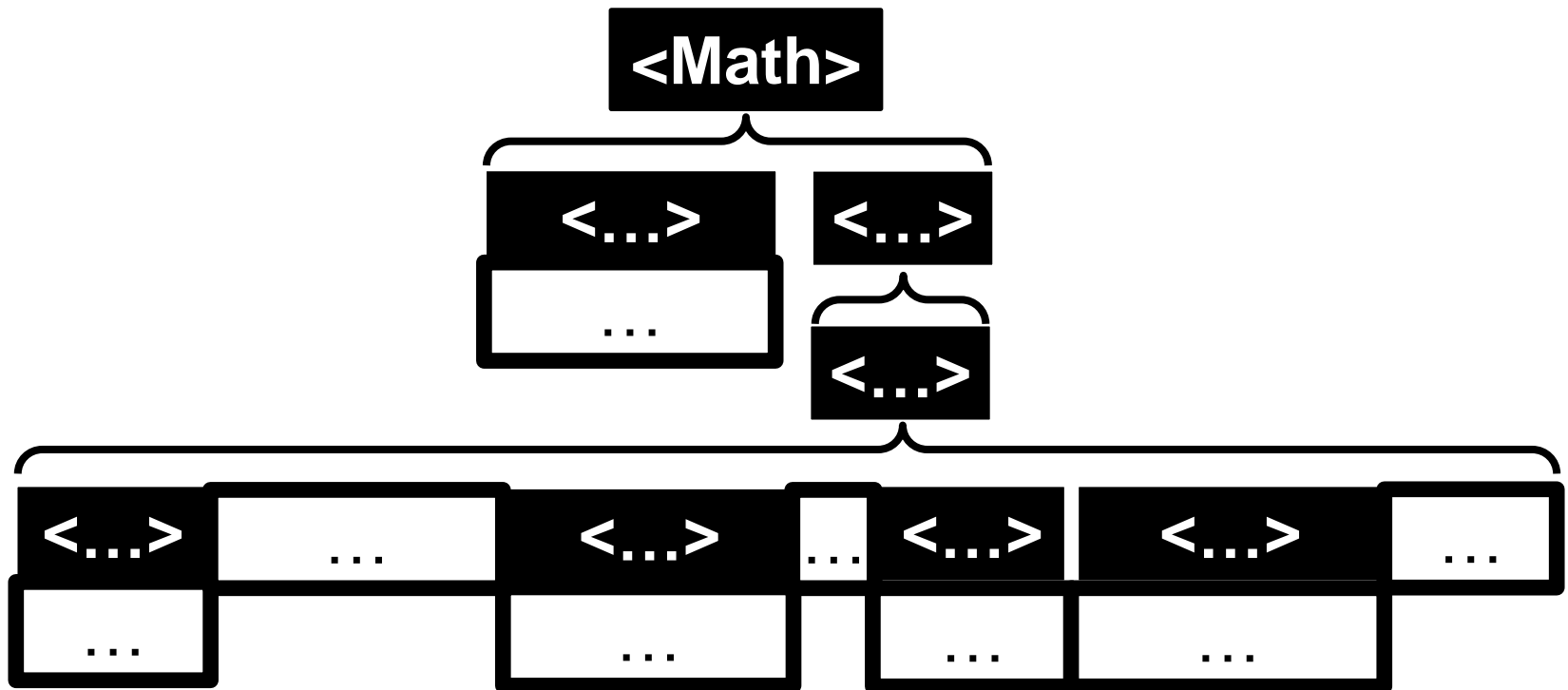
Introduction

**Unpack**



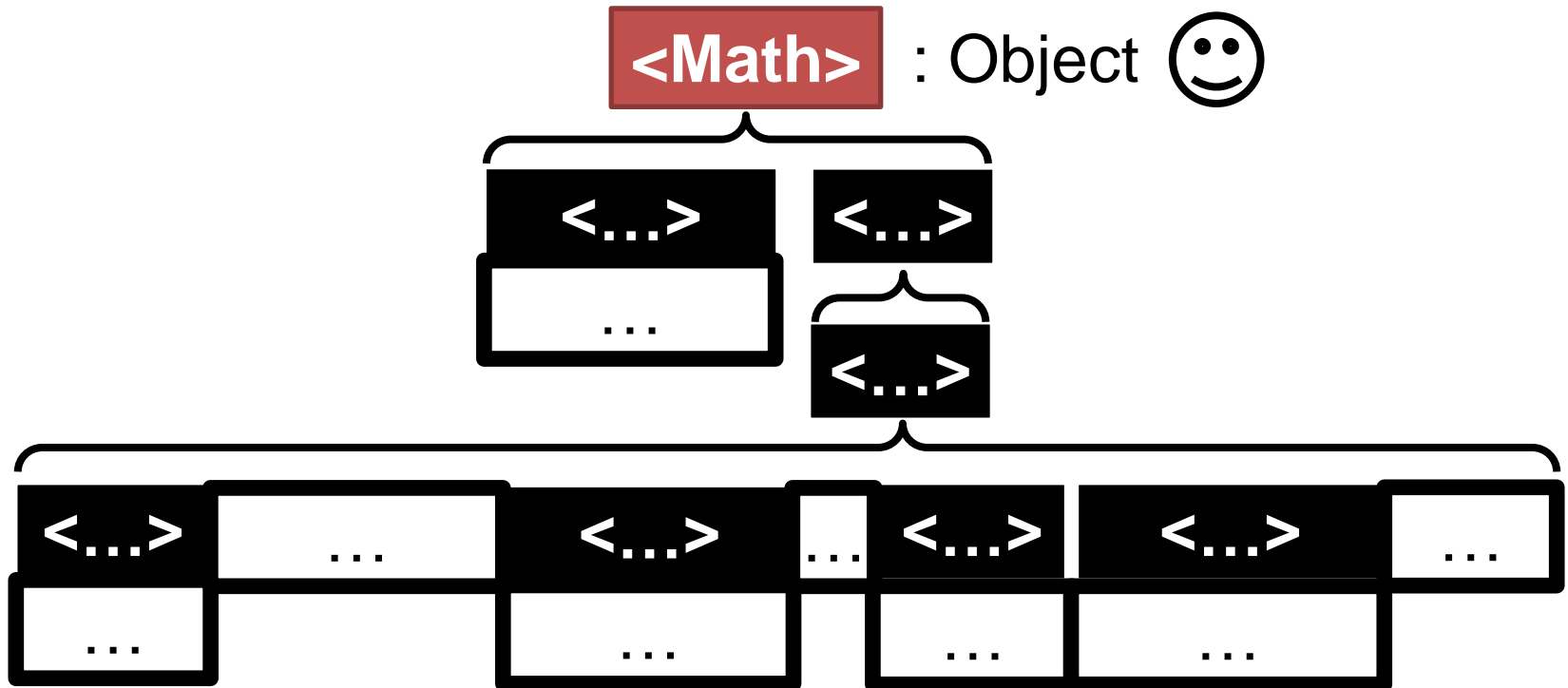
# Efficiency Brought by Procedure

- Regions enclosed by Meta-info/Object tags are not unpacked → **User labor is saved**



# Efficiency Brought by Procedure

- Regions enclosed by Meta-info/Object tags are not unpacked → **User labor is saved**



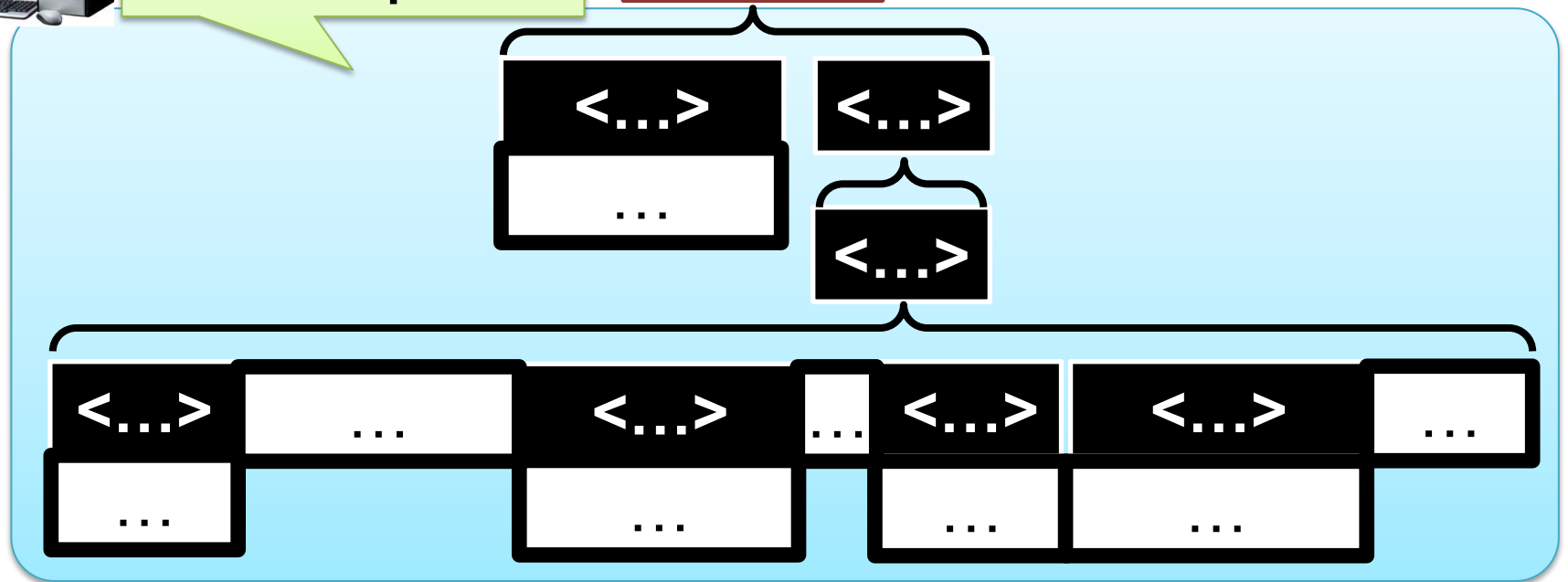
# Efficiency Brought by Procedure

- Regions enclosed by Meta-info/Object tags are not unpacked → **User labor is saved**



Not unpack

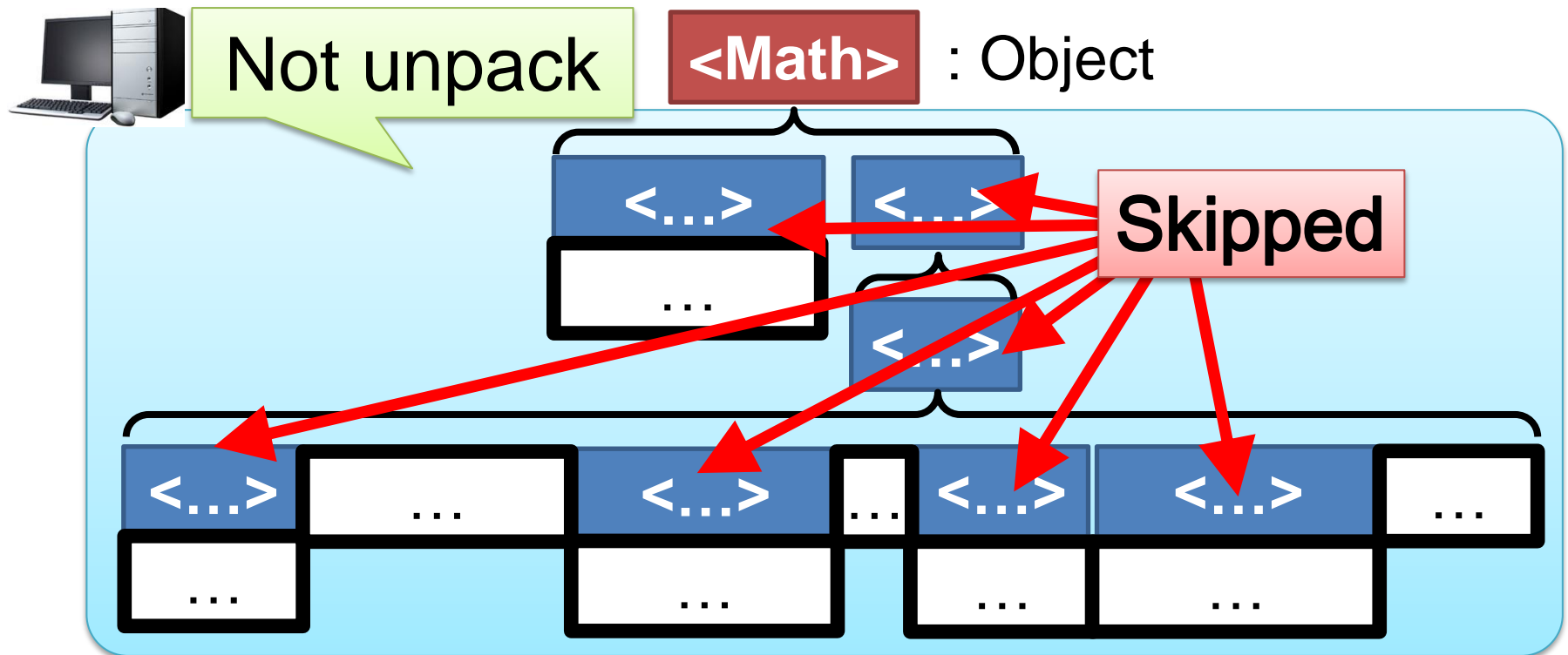
**<Math>** : Object





# Efficiency Brought by Procedure

- Regions enclosed by Meta-info/Object tags are not unpacked → **User labor is saved**



# GUI-based Tool for Classification & Conversion Procedure

| Unclassified tags   | Classify   | Independent tags  | Object tags   |
|---|--|---|---|
| Math<br>equationgroup<br>quote<br>text<br>enumerate<br>cite<br>break<br>ref<br>tabular<br><b>note</b><br>graphics<br>emph | Independent >><br>Decoration >><br>Object >><br>Meta-info >><br><< Release<br><b>Attribute</b><br><input type="text"/><br>(Only one) | document<br>title<br>abstract<br>section<br>subsection<br><br>p<br>para | <br><br><br><br><br><br>bibliography<br>creator<br>appendix |

Save the classification & Convert XML documents

## Context

[10009\_2006\_10.xml-format-tag-removed]:

... bloms of `<emph>specifying</emph>` contracts, `<emph>monitoring</emph>` their execution for performance. `<note class="footnote" mark="1"><emph>Performance</emph> in contract lingo refers to <emph>compliance</emph> with the <emph>promises</emph> (contractual commitments) stipulated in a contract; nonperformance is also termed <emph>breach of contract</emph>. </note> <emph>analyzing</emph> their ramifications for planning, pricing and other purposes prior to and du ...`

[10009\_2006\_10.xml-format-tag-removed]:

... `<emph><text>operational</text> semantics</emph>` is ideally suited to alleviating the above problems. `<note class="footnote" mark="2">`Our language is rendered in ordinary linear syntax, but we do not intend to limit the scope of the term "language" to specify linear sequences of characters only, but to include graphical objects and the like `</note>` Note that contracts are not put to a single use as programs are, whose sole u

# Experiments

- Extract plain text sequences from several types of documents
  - Examine efficiency of tag classification
- Apply NLP tools to obtained sequences
  - Compare performance with naïvely obtained sequences
  - Discuss impact of proper extraction of plain text

# Experimental Settings (1/3): Target Documents

| Article type                       | Domain           | Format | # used |
|------------------------------------|------------------|--------|--------|
| PubMed Central (PMC)* <sup>1</sup> | Scientific paper | XML    | 1,000  |
| arXiv.org* <sup>2</sup>            |                  | XHTML  | 300    |
| ACL 2014* <sup>3</sup>             |                  | XHTML  | 67     |
| Wikipedia* <sup>4</sup> entries    | Web page         | HTML†  | 300    |

(† XML-like: generated via intermediate XML files)

---

\*1 <http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/>

\*2 <http://arxiv.org/>

\*3 <http://anthology.aclweb.org/>

\*4 <http://www.wikipedia.org/>

# Experimental Settings (2/3): NLP Tools (Two Types of Parsers)

- **Enju parser** <sup>\*1</sup> (+ Genia Sentence Splitter<sup>\*2</sup>)
  - Deep syntactic/semantic analysis
  - Memory overflow of search space = failure
- **Stanford parser** <sup>\*3</sup>
  - Phrase structure & dependency analysis
  - Failure in long sentences terminated whole process → limit sentence length to 50 words
  - (>50 word sentences were skipped → failure)

---

\*1 Enju (Ninomiya et al., 2007)

\*2 GeniaSS (<http://www.nactem.ac.uk/y-matsu/geniass/>)

\*3 Stanford parser (de Marneffe et al., 2006)

# Experimental Settings (3/3): Comparison and Evaluation

- Three types of tag treatment compared:
  - **Remove**: simply remove all tags
  - **O/M**: Object/Meta-info → our framework  
Decoration / Independent → remove
  - **I/D/O/M**: process all tags using our framework
- Performance measured by:
  - # sentences detected by parser
  - Total parsing time
  - # (% of) sentences which could not be parsed

# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#)             | # tags (# types)               | # classified tags (# types) |               |             |              |                             | # seq.      |
|------------------------------|--------------------------------|-----------------------------|---------------|-------------|--------------|-----------------------------|-------------|
|                              |                                | I                           | D             | O           | M            | Total                       |             |
| <b>PMC</b><br><b>(1,000)</b> | <b>1,357 k</b><br><b>(421)</b> | 32k<br>(12)                 | 62k<br>(9)    | 48k<br>(9)  | 34k<br>(56)  | <b>177k</b><br><b>(85)</b>  | <b>26 k</b> |
| <b>arXiv</b><br><b>(300)</b> | <b>1,969 k</b><br><b>(210)</b> | 6k<br>(15*)                 | 47k<br>(12*)  | 60k<br>(8*) | 8k<br>(17*)  | <b>121k</b><br><b>(52*)</b> | <b>4 k</b>  |
| <b>ACL</b><br><b>(67)</b>    | <b>131 k</b><br><b>(66)</b>    | 3k<br>(24*)                 | 14 k<br>(29*) | 5k<br>(15*) | 2k<br>(19*)  | <b>24k</b><br><b>(87*)</b>  | <b>2 k</b>  |
| <b>Wiki.</b><br><b>(300)</b> | <b>224 k</b><br><b>(60)</b>    | 3k<br>(12*)                 | 11 k<br>(8*)  | 1k<br>(28*) | 11k<br>(67*) | <b>28k</b><br><b>(115*)</b> | <b>2 k</b>  |

(**I**: Independent, **D**: Decoration, **O**: Object, **M**: Meta-info)

# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#)             | # tags (# types)               | # classified tags (# types) |                             |                           |                            |                             | # seq.      |
|------------------------------|--------------------------------|-----------------------------|-----------------------------|---------------------------|----------------------------|-----------------------------|-------------|
|                              |                                | I                           | D                           | O                         | M                          | Total                       |             |
| <b>PMC</b><br><b>(1,000)</b> | <b>1,357 k</b><br><b>(421)</b> | <b>52k</b><br><b>(12)</b>   | <b>52k</b><br><b>(9)</b>    | <b>40k</b><br><b>(9)</b>  | <b>51k</b><br><b>(50)</b>  | <b>177k</b><br><b>(85)</b>  | <b>26 k</b> |
| <b>arXiv</b><br><b>(300)</b> | <b>1,969 k</b><br><b>(210)</b> | <b>6k</b><br><b>(15*)</b>   | <b>47k</b><br><b>(12*)</b>  | <b>60k</b><br><b>(8*)</b> | <b>8k</b><br><b>(17*)</b>  | <b>121k</b><br><b>(52*)</b> | <b>4 k</b>  |
| <b>ACL</b><br><b>(67)</b>    | <b>131 k</b><br><b>(66)</b>    | <b>3k</b><br><b>(24*)</b>   | <b>14 k</b><br><b>(29*)</b> | <b>5k</b><br><b>(15*)</b> | <b>2k</b><br><b>(19*)</b>  | <b>24k</b><br><b>(87*)</b>  | <b>2 k</b>  |
| <b>Wiki.</b><br><b>(300)</b> | <b>224 k</b><br><b>(60)</b>    | <b>3k</b><br><b>(12*)</b>   | <b>11 k</b><br><b>(8*)</b>  | <b>1k</b><br><b>(28*)</b> | <b>11k</b><br><b>(67*)</b> | <b>28k</b><br><b>(115*)</b> | <b>2 k</b>  |

Classify only 20% of total tag types

(**I**: Independent, **D**: Decoration, **O**: Object, **M**: Meta-info)



# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#) | # tags (# types) | # classified tags (# types) |              |             |             |               | # seq. |
|------------------|------------------|-----------------------------|--------------|-------------|-------------|---------------|--------|
|                  |                  | I                           | D            | O           | M           | Total         |        |
| PMC<br>(1,000)   | 1,357 k<br>(421) | 52k<br>(12)                 | 52k<br>(9)   | 40k<br>(9)  | 51k<br>(50) | 177k<br>(85)  | 26 k   |
| arXiv<br>(300)   | 1,969 k<br>(210) | 6k<br>(15*)                 | 47k<br>(12*) | 60k<br>(8*) | 8k<br>(17*) | 121k<br>(52*) | 4 k    |

Classify only 20% of total tag types

\* XHTML, HTML:

Tag names represent more abstract information  
 → Use combination of tag name & its selected attributes as a single tag

(I: Independent, D: Decoration, O: Object, M: Meta-info)

# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#) | # tags (# types) | # classified tags (# types) |           |          |           |            | # seq. |
|------------------|------------------|-----------------------------|-----------|----------|-----------|------------|--------|
|                  |                  | I                           | D         | O        | M         | Total      |        |
| PMC (1,000)      | 1,357 k (421)    | 52k (12)                    | 52k (9)   | 40k (9)  | 51k (50)  | 177k (85)  | 26 k   |
| arXiv (300)      | 1,969 k (210)    | 3k (15*)                    | 17k (12*) | 33k (8*) | 3k (17*)  | 121k (52*) | 4 k    |
| ACL (67)         | 131 k            | 3k                          | 14k       | 5k       | 2k        | 24k        | 2 k    |
| Wiki. (300)      | 224 k (60)       | 3k (12*)                    | 11k (8*)  | 1k (28*) | 11k (67*) | 28k (115*) | 2 k    |

Classify only 20% of total tag types

Focus on less than 20% of total occurrences

(I: Independent, D: Decoration, O: Object, M: Meta-info)

# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#) | # tags (# types) | # classified tags (# types) |           |          |          |            | # seq. |
|------------------|------------------|-----------------------------|-----------|----------|----------|------------|--------|
|                  |                  | I                           | D         | O        | M        | Total      |        |
| PMC (1,000)      | 1,357 k (421)    | 52k (12)                    | 52k (9)   | 40k (9)  | 51k (50) | 177k (85)  | 26 k   |
| arXiv (300)      | 1,969 k (210)    | 3k (15*)                    | 17k (12*) | 33k (8*) | 3k (17*) | 121k (52*) | 4 k    |
| ACL (67)         | 131 k            | 3k                          | 14k       | 5k       | 2k       | 24k        | 2 k    |
| Wiki. (300)      | 224 k            | 3k                          | 11k       | 1k       | 11k      | 28k        | 2 k    |

Classify only 20% of total tag types

Focus on less than 20% of total occurrences

Tags within regions enclosed by Object/ Meta-info tags were not considered

(I: Independent, D: Document, O: Object, M: Meta-info)

# Classified Tags and Obtained Sequences for Each Type of Article

| Article type (#) | # tags (# types) | # classified tags (# types) |              |             |             |               | # seq. |
|------------------|------------------|-----------------------------|--------------|-------------|-------------|---------------|--------|
|                  |                  | I                           | D            | O           | M           | Total         |        |
| PMC<br>(1,000)   | 1,357 k<br>(421) | 32k<br>(12)                 | 62k<br>(9)   | 48k<br>(9)  | 34k<br>(56) | 177k<br>(85)  | 26 k   |
| arXiv<br>(300)   | 1,969 k<br>(210) | 6k<br>(15*)                 | 47k<br>(12*) | 60k<br>(8*) | 8k<br>(17*) | 121k<br>(52*) | 4 k    |
| ACL              | 131 k            | 3k                          | 14 k         | 5k          | 2k          | 24k           | 2 k    |

Observe sequences for randomly-selected articles

→ They consisted of valid sentences:

- Could be directly input into NLP tools
- Thoroughly covered content of original articles

2 k

info)

# Impact of Tag Treatment on Performance of Enju Parser

| Art. (#)               | Treatment      | # sentences | Time (s) | # failure (%) |
|------------------------|----------------|-------------|----------|---------------|
| <b>PMC<br/>(1,000)</b> | <b>Remove</b>  | 159,327     | 209,783  | 4,721 (2.96)  |
|                        | <b>O/M</b>     | 112,285     | 135,752  | 810 (0.72)    |
|                        | <b>I/D/O/M</b> | 126,215     | 132,250  | 699 (0.55)    |
| <b>arXiv<br/>(300)</b> | <b>Remove</b>  | 74,762      | 108,831  | 2,047 (2.74)  |
|                        | <b>O/M</b>     | 41,265      | 89,200   | 411 (1.00)    |
|                        | <b>I/D/O/M</b> | 43,208      | 87,952   | 348 (0.81)    |
| <b>ACL<br/>(67)</b>    | <b>Remove</b>  | 19,571      | 15,142   | 115 (0.59)    |
|                        | <b>O/M</b>     | 9,819       | 9,481    | 63 (0.64)     |
|                        | <b>I/D/O/M</b> | 11,136      | 8,482    | 39 (0.35)     |
| <b>Wiki.<br/>(300)</b> | <b>Remove</b>  | 10,561      | 14,704   | 1,161 (10.99) |
|                        | <b>O/M</b>     | 5,026       | 6,743    | 67 (1.33)     |
|                        | <b>I/D/O/M</b> | 6,893       | 6,058    | 61 (0.88)     |

# Impact of Tag Treatment on Performance of Stanford Parser

| Art. (#)               | Treatment      | # sentences | Time (s) | # failure (%)  |
|------------------------|----------------|-------------|----------|----------------|
| <b>PMC<br/>(1,000)</b> | <b>Remove</b>  | 170,999     | 58,865   | 18,621 (10.89) |
|                        | <b>O/M</b>     | 126,176     | 50,741   | 11,881 (9.42)  |
|                        | <b>I/D/O/M</b> | 139,805     | 63,295   | 11,338 (8.11)  |
| <b>arXiv<br/>(300)</b> | <b>Remove</b>  | 75,672      | 27,970   | 10,590 (13.99) |
|                        | <b>O/M</b>     | 48,666      | 24,630   | 5,457 (11.21)  |
|                        | <b>I/D/O/M</b> | 50,504      | 26,360   | 5,345 (10.58)  |
| <b>ACL<br/>(67)</b>    | <b>Remove</b>  | 17,166      | 5,047    | 1,095 (6.38)   |
|                        | <b>O/M</b>     | 11,182      | 4,157    | 616 (5.51)     |
|                        | <b>I/D/O/M</b> | 12,402      | 4,871    | 587 (4.73)     |
| <b>Wiki.<br/>(300)</b> | <b>Remove</b>  | 14,883      | 3,114    | 1,651 (11.09)  |
|                        | <b>O/M</b>     | 6,173       | 2,248    | 282 (4.57)     |
|                        | <b>I/D/O/M</b> | 8,049       | 2,451    | 258 (3.21)     |

# Impact of Tag Treatment on Performance of Enju Parser

| Art. (#)               | Treatment     | # sentences | Time (s) | # failure (%) |
|------------------------|---------------|-------------|----------|---------------|
| <b>PMC<br/>(1,000)</b> | <b>Remove</b> | 159,327     | 209,783  | 4,721 (2.96)  |
|                        | <b>O/M</b>    | 112,285     | 135,752  | 810 (0.72)    |
|                        | I/D/O/M       | 126,215     | 132,250  | 699 (0.55)    |
| <b>arXiv<br/>(300)</b> | <b>Remove</b> | 74,762      | 108,831  | 2,047 (2.74)  |
|                        | <b>O/M</b>    | 41,265      | 89,200   | 411 (1.00)    |
|                        | I/D/O/M       | 43,208      | 87,952   | 348 (0.81)    |
| <b>ACL<br/>(67)</b>    | <b>Remove</b> | 19,571      | 15,142   | 115 (0.59)    |
|                        | <b>O/M</b>    | 9,819       | 9,481    | 63 (0.64)     |
|                        | I/D/O/M       | 11,136      | 8,482    | 39 (0.35)     |
| <b>Wiki.<br/>(300)</b> | <b>Remove</b> | 10,561      | 14,704   | 1,161 (10.99) |
|                        | <b>O/M</b>    | 5,026       | 6,743    | 67 (1.33)     |
|                        | I/D/O/M       | 6,893       | 6,058    | 61 (0.88)     |

# Impact of Tag Treatment on Performance of

Parsing failure: ~10%↓  
 → Much higher coverage

| Art. (#)       | Treatment | # sentences |         |                   |  |
|----------------|-----------|-------------|---------|-------------------|--|
| PMC<br>(1,000) | Remove    | 159,327     | 209,783 | 4,721 (2.96)      |  |
|                | O/M       | 112,285     | 135,752 | <b>810 (0.72)</b> |  |
|                | I/D/O/M   | 126,215     | 132,250 | 699 (0.55)        |  |
| arXiv<br>(300) | Remove    | 74,762      | 108,831 | 2,047 (2.74)      |  |
|                | O/M       | 41,265      | 89,200  | <b>411 (1.00)</b> |  |
|                | I/D/O/M   | 43,208      | 87,952  | 348 (0.81)        |  |
| ACL<br>(67)    | Remove    | 19,571      | 15,142  | 115 (0.59)        |  |
|                | O/M       | 9,819       | 9,481   | <b>63 (0.64)</b>  |  |
|                | I/D/O/M   | 11,136      | 8,482   | 39 (0.35)         |  |
| Wiki.<br>(300) | Remove    | 10,561      | 14,704  | 1,161 (10.99)     |  |
|                | O/M       | 5,026       | 6,743   | <b>67 (1.33)</b>  |  |
|                | I/D/O/M   | 6,893       | 6,058   | 61 (0.88)         |  |



# Impact of Tag Treatment on

Parsing time: 18 ~ 54%↓  
→ Drastic speed-up

Parsing failure: ~10%↓  
→ Much higher coverage

| Ar             |         |         |         |       |         |
|----------------|---------|---------|---------|-------|---------|
| PMC<br>(1,000) | Remove  | 155,827 | 209,783 | 4,721 | (2.96)  |
|                | O/M     | 112,285 | 135,752 | 810   | (0.72)  |
|                | I/D/O/M | 126,215 | 132,250 | 699   | (0.55)  |
| arXiv<br>(300) | Remove  | 74,762  | 108,831 | 2,047 | (2.74)  |
|                | O/M     | 41,265  | 89,200  | 411   | (1.00)  |
|                | I/D/O/M | 43,208  | 87,952  | 348   | (0.81)  |
| ACL<br>(67)    | Remove  | 19,571  | 15,142  | 115   | (0.59)  |
|                | O/M     | 9,819   | 9,481   | 63    | (0.64)  |
|                | I/D/O/M | 11,136  | 8,482   | 39    | (0.35)  |
| Wiki.<br>(300) | Remove  | 10,561  | 14,704  | 1,161 | (10.99) |
|                | O/M     | 5,026   | 6,743   | 67    | (1.33)  |
|                | I/D/O/M | 6,893   | 6,058   | 61    | (0.88)  |

# Impact of Tag Treatment on Performance of Stanford Parser

| Art. (#)       | Treatment | # sentences | Time (s) | # failure (%)  |
|----------------|-----------|-------------|----------|----------------|
| PMC<br>(1,000) | Remove    | 170,999     | 58,865   | 18,621 (10.89) |
|                | O/M       | 126,176     | 50,741   | 11,881 (9.42)  |
|                | I/D/O/M   | 139,805     | 63,295   | 11,338 (8.11)  |
| arXiv<br>(300) | Remove    | 75,672      | 27,970   | 10,590 (13.99) |
|                | O/M       | 48,666      | 24,630   | 5,457 (11.21)  |
|                | I/D/O/M   | 50,504      | 26,360   | 5,345 (10.58)  |
| ACL<br>(67)    | Remove    | 17,166      | 5,047    | 1,095 (6.38)   |
|                | O/M       | 11,182      | 4,157    | 616 (5.51)     |
|                | I/D/O/M   | 12,402      | 4,871    | 587 (4.73)     |
| Wiki.<br>(300) | Remove    | 14,883      | 3,114    | 1,651 (11.09)  |
|                | O/M       | 6,173       | 2,248    | 282 (4.57)     |
|                | I/D/O/M   | 8,049       | 2,451    | 258 (3.21)     |

# Impact of Tag Treatment on

Parsing time: 12 ~ 28%↓  
→ Drastic speed-up

Parsing failure: 1 ~ 7%↓  
→ Much higher coverage

| Ar             |         |         |        |        |         |  |
|----------------|---------|---------|--------|--------|---------|--|
| PMC<br>(1,000) | Remove  | 170,999 | 58,865 | 18,621 | (10.89) |  |
|                | O/M     | 126,176 | 50,741 | 11,881 | (9.42)  |  |
|                | I/D/O/M | 139,805 | 63,295 | 11,338 | (8.11)  |  |
| arXiv<br>(300) | Remove  | 75,672  | 27,970 | 10,590 | (13.99) |  |
|                | O/M     | 48,666  | 24,630 | 5,457  | (11.21) |  |
|                | I/D/O/M | 50,504  | 26,360 | 5,345  | (10.58) |  |
| ACL<br>(67)    | Remove  | 17,166  | 5,047  | 1,095  | (6.38)  |  |
|                | O/M     | 11,182  | 4,157  | 616    | (5.51)  |  |
|                | I/D/O/M | 12,402  | 4,871  | 587    | (4.73)  |  |
| Wiki.<br>(300) | Remove  | 14,883  | 3,114  | 1,651  | (11.09) |  |
|                | O/M     | 6,173   | 2,248  | 282    | (4.57)  |  |
|                | I/D/O/M | 8,049   | 2,451  | 258    | (3.21)  |  |

Non-NL(Object)/-target(Meta-info) parts were excluded

Only target word sequences were input

## Performance of Stanford Parser

→ Drastic speed-up

→ Much higher coverage

|                |         |         |        |                |
|----------------|---------|---------|--------|----------------|
| PMC<br>(1,000) | Remove  | 170,999 | 58,865 | 18,621 (10.89) |
|                | O/M     | 126,176 | 50,741 | 11,881 (9.42)  |
|                | I/D/O/M | 139,805 | 63,295 | 11,338 (8.11)  |
| arXiv<br>(300) | Remove  | 75,672  | 27,970 | 10,590 (13.99) |
|                | O/M     | 48,666  | 24,630 | 5,457 (11.21)  |
|                | I/D/O/M | 50,504  | 26,360 | 5,345 (10.58)  |
| ACL<br>(67)    | Remove  | 17,166  | 5,047  | 1,095 (6.38)   |
|                | O/M     | 11,182  | 4,157  | 616 (5.51)     |
|                | I/D/O/M | 12,402  | 4,871  | 587 (4.73)     |
| Wiki.<br>(300) | Remove  | 14,883  | 3,114  | 1,651 (11.09)  |
|                | O/M     | 6,173   | 2,248  | 282 (4.57)     |
|                | I/D/O/M | 8,049   | 2,451  | 258 (3.21)     |

# Impact of Tag Treatment on Performance of Enju Parser

| Art. (#)               | Treatment      | # sentences | Time (s) | # failure (%) |
|------------------------|----------------|-------------|----------|---------------|
| <b>PMC<br/>(1,000)</b> | Remove         | 159,327     | 209,783  | 4,721 (2.96)  |
|                        | <b>O/M</b>     | 112,285     | 135,752  | 810 (0.72)    |
|                        | <b>I/D/O/M</b> | 126,215     | 132,250  | 699 (0.55)    |
| <b>arXiv<br/>(300)</b> | Remove         | 74,762      | 108,831  | 2,047 (2.74)  |
|                        | <b>O/M</b>     | 41,265      | 89,200   | 411 (1.00)    |
|                        | <b>I/D/O/M</b> | 43,208      | 87,952   | 348 (0.81)    |
| <b>ACL<br/>(67)</b>    | Remove         | 19,571      | 15,142   | 115 (0.59)    |
|                        | <b>O/M</b>     | 9,819       | 9,481    | 63 (0.64)     |
|                        | <b>I/D/O/M</b> | 11,136      | 8,482    | 39 (0.35)     |
| <b>Wiki.<br/>(300)</b> | Remove         | 10,561      | 14,704   | 1,161 (10.99) |
|                        | <b>O/M</b>     | 5,026       | 6,743    | 67 (1.33)     |
|                        | <b>I/D/O/M</b> | 6,893       | 6,058    | 61 (0.88)     |

# Impact of Tag Treatment on Performance of Enju Parser

| App (#)     | Treatment | # Tokens | Time (s) | # Errors (%)  |
|-------------|-----------|----------|----------|---------------|
| PWC (1,000) | O/M       | 112,285  | 135.752  | 810 (0.72)    |
|             | I/D/O/M   | 126,215  | 132.250  | 699 (0.55)    |
|             | Remove    | 74,762   | 108.831  | 2,047 (2.74)  |
| arXiv (300) | O/M       | 41,265   | 89.200   | 411 (1.00)    |
|             | I/D/O/M   | 43,208   | 87.952   | 348 (0.81)    |
|             | Remove    | 19,571   | 15.142   | 115 (0.59)    |
| ACL (67)    | O/M       | 9,819    | 9.481    | 63 (0.64)     |
|             | I/D/O/M   | 11,136   | 8.482    | 39 (0.35)     |
|             | Remove    | 10,561   | 14.704   | 1,161 (10.99) |
| Wiki. (300) | O/M       | 5,026    | 6.743    | 67 (1.33)     |
|             | I/D/O/M   | 6,895    | 6.058    | 61 (0.88)     |
|             | Remove    | 10,561   | 14.704   | 1,161 (10.99) |

Parsing time: 1~11%↓

Parsing failure: 0.2~0.5%↓

Embedded sentences (Independent) were separated  
 → **Correct & shorter sentences** were increased



Parser could perform better

Parsing time: 1~11%↓

Parsing failure: 0.2~0.5%↓

| Dataset                    | Method  | Time (s) | Time (s) | Failures | Failures (%) |
|----------------------------|---------|----------|----------|----------|--------------|
| P110 (1,000)               | O/M     | 112,285  | 135,752  | 810      | (0.72)       |
|                            | I/D/O/M | 126,215  | 132,250  | 699      | (0.55)       |
| arXiv (300)                | Remove  | 74,762   | 108,831  | 2,047    | (2.74)       |
|                            | O/M     | 41,265   | 89,200   | 411      | (1.00)       |
|                            | I/D/O/M | 43,208   | 87,952   | 348      | (0.81)       |
| ACL (67)                   | Remove  | 19,571   | 15,142   | 115      | (0.59)       |
|                            | O/M     | 9,819    | 9,481    | 63       | (0.64)       |
|                            | I/D/O/M | 11,130   | 8,482    | 39       | (0.35)       |
| Detected sentences: 5~37%↑ | Remove  | 10,561   | 14,704   | 1,161    | (10.99)      |
|                            | O/M     | 5,026    | 6,743    | 67       | (1.33)       |
|                            | I/D/O/M | 6,895    | 6,058    | 61       | (0.88)       |

# Impact of Tag Treatment on Performance of Stanford Parser

| Dataset     | Tag Treatment | Total Tokens | Parse Time (s) | Parse Failure (%) |
|-------------|---------------|--------------|----------------|-------------------|
| All (1,000) | Remove        | 126,176      | 50,741         | 11,881 (9.42)     |
|             | O/M           | 139,805      | 63,295         | 11,338 (8.11)     |
|             | I/D/O/M       | 139,805      | 63,295         | 11,338 (8.11)     |
| arXiv (300) | Remove        | 75,672       | 27,970         | 10,590 (13.99)    |
|             | O/M           | 48,666       | 24,630         | 5,457 (11.21)     |
|             | I/D/O/M       | 50,504       | 26,360         | 5,345 (10.58)     |
| ACL (67)    | Remove        | 17,166       | 5,047          | 1,095 (6.38)      |
|             | O/M           | 11,182       | 4,157          | 616 (5.51)        |
|             | I/D/O/M       | 12,402       | 4,871          | 587 (4.73)        |
| Wiki. (300) | Remove        | 14,883       | 3,114          | 1,651 (11.09)     |
|             | O/M           | 6,173        | 2,248          | 282 (4.57)        |
|             | I/D/O/M       | 8,049        | 2,451          | 258 (3.21)        |

Parsing time: 7~25%↑

Parsing failure: 0.6~1.5%↓



Embedded sentences (Independent) were separated  
 → Correct & shorter sentences were increased

| Dataset             | Method  | Count   | Percentage (%) | Count  | Percentage (%) |
|---------------------|---------|---------|----------------|--------|----------------|
| Pile (1,000)        | O/M     | 126,176 | 50.741         | 11,881 | (9.42)         |
|                     | I/D/O/M | 139,805 | 63.295         | 11,338 | (8.11)         |
| arXiv (300)         | Remove  | 75,672  | 27,970         | 10,590 | (13.99)        |
|                     | O/M     | 48,666  | 24,630         | 5,457  | (11.21)        |
|                     | I/D/O/M | 50,504  | 26.360         | 5,345  | (10.58)        |
| ACL (67)            | Remove  | 17,166  | 5,047          | 1,095  | (6.38)         |
|                     | O/M     | 11,182  | 4,157          | 616    | (5.51)         |
|                     | I/D/O/M | 12,402  | 4.871          | 587    | (4.73)         |
| Detected sentences: | Remove  | 14,883  | 3,114          | 1,651  | (11.09)        |
|                     | O/M     | 6,173   | 2,248          | 282    | (4.57)         |
|                     | I/D/O/M | 8,049   | 2.451          | 258    | (3.21)         |

Parsing time: 7~25%↑

Parsing failure: 0.6~1.5%↓

Detected sentences:  
5~37%↑

Embedded sentences (Independent) were separated  
 → **Correct & shorter sentences** were increased

→ Skipped (>50 word) sentences were decreased

Parsing time: 7~25%↑

Parsing failure: 0.6~1.5%↓

|                               |         |         |        |                |
|-------------------------------|---------|---------|--------|----------------|
| ArXiv<br>(1,000)              | O/M     | 126,176 | 50,741 | 11,881 (9.42)  |
|                               | I/D/O/M | 139,805 | 63,295 | 11,338 (8.11)  |
| arXiv<br>(300)                | Remove  | 75,672  | 27,970 | 10,590 (13.99) |
|                               | O/M     | 48,666  | 24,630 | 5,457 (11.21)  |
|                               | I/D/O/M | 50,504  | 26,360 | 5,345 (10.58)  |
| ACL<br>(67)                   | Remove  | 17,166  | 5,047  | 1,095 (6.38)   |
|                               | O/M     | 11,182  | 4,157  | 616 (5.51)     |
|                               | I/D/O/M | 12,402  | 4,871  | 587 (4.73)     |
| Detected sentences:<br>5~37%↑ | Remove  | 14,883  | 3,114  | 1,651 (11.09)  |
|                               | O/M     | 6,173   | 2,248  | 282 (4.57)     |
|                               | I/D/O/M | 8,049   | 2,451  | 258 (3.21)     |

# Discussion: What Thorough & Efficient Document Processing Brings About

- Shallow analysis with simple approach (word count etc.) → removing tags will be enough
  - # words is not affected by embedded sentences
  - Some non-NL seq. canceled by many NL seq.
- Detailed/precise analysis (discourse analysis/translation/grammar extraction/etc.)
  - Even subtle utterance cannot be overlooked
  - Seq. other than body text should be excluded
  - = Presumed condition in most NLP challenges

# Significance of Bridging Real-world Documents and NLP Technologies

- True goal of NLP challenges = analyze any real-world(, richly formatted) documents
- Proper framework enables conventional NLP tools to process real-world text without significant loss of performance

“Adequately bridging target real-world documents and NLP technologies”  
= Crucial task for utilizing full benefit brought by NLP in ubiquitous application

# Summary

- We proposed framework for data conversion between XML-tagged text and I/O of NLP
  - According to classification of tag functions
- We succeeded in obtaining plain text seq. from target doc. by classifying 20% of tags
  - Much more thorough & efficient parsing of target doc. than with naively tag-removed text
  - Emphasize significance of bridging real-world documents and NLP technologies

# Future Work

- Release tool for converting XML doc. into plain text seq. utilizing our framework
  - Share further discussion on applying NLP tools to various real-world documents
- Treat tag & tagged regions more flexibly
  - Treatment of textual parts in Object regions
- Apply NLP to various formats of documents
  - OCR data / presentation slides / etc.

# Downloadable Package Almost Ready (will be Released in September)

PlaneText - Google Chrome  
kmcs.nii.ac.jp/apps/planetext/dataset/arxiv50

## PlaneText

**Unknown Tags**

| Tag       | Attribute | Word | Value | Instance              |
|-----------|-----------|------|-------|-----------------------|
| xmlns:div | id        | emph | emph  | 0904.0684.xhtml (128) |
| xmlns:a   | class     |      |       | 0904.0684.xhtml (162) |
| xmlns:ol  |           |      |       | 0904.0684.xhtml (163) |
| xmlns:ul  |           |      |       | 0904.0684.xhtml (163) |
| xmlns:em  |           |      |       | 0904.0684.xhtml (163) |
| m:math    |           |      |       | 0904.0684.xhtml (163) |

**Classified Tags**

| Independent   | Decoration                                  | Object | Metainfo   | Options   |
|---|---|--------|--|---|
| xmlns:html<br>xmlns:body<br>xmlns:div[class: main]<br>xmlns:div[class: conte]<br>xmlns:div[class: docu]<br>xmlns:div[class: abstr | xmlns:p[class: p]<br>xmlns:div[class: para] |        | xmlns:head<br>xmlns:div[class: foote<br>xmlns:div[class: RDFa<br>xmlns:div[class: creat<br>xmlns:div[class: creat<br>xmlns:div[class: keyw | <input checked="" type="checkbox"/> Autosubmit<br>Submit<br>Docs 5 Set<br><a href="#">Config</a> , <a href="#">Data</a> |

knots, any two band presentations for the same knot are stably equivalent.

*Dedicated to the memory of Xiao-Song Lin*

0/50

<http://kmcs.nii.ac.jp/planetext/>

Thank you !