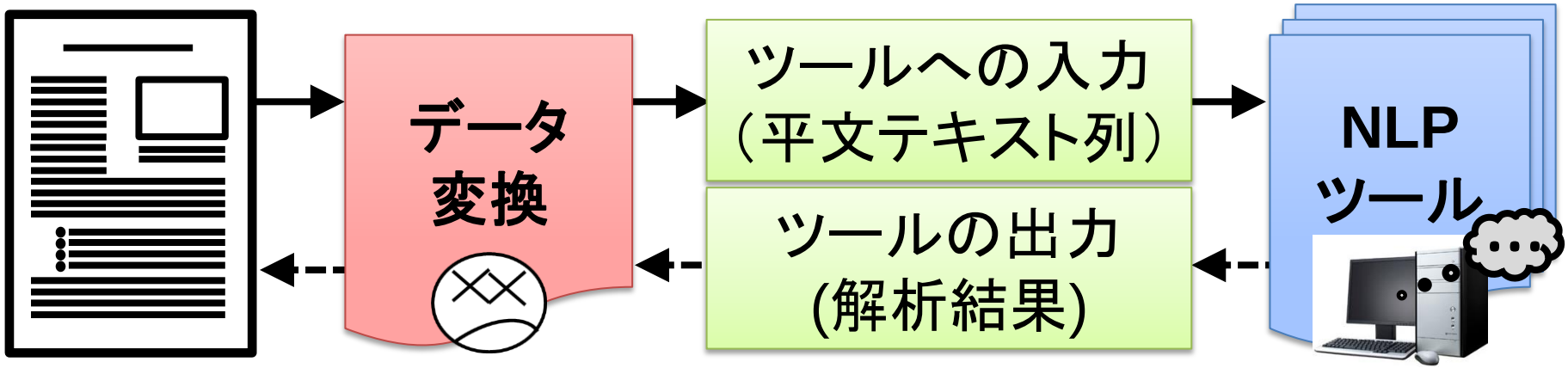
実文書を自然言語処理技術と適切に繋ぐ技術の重要性

原 忠義　　トピチ ゴラン
宮尾 祐介　　相澤 彰子
（国立情報学研究所）

本研究の目的

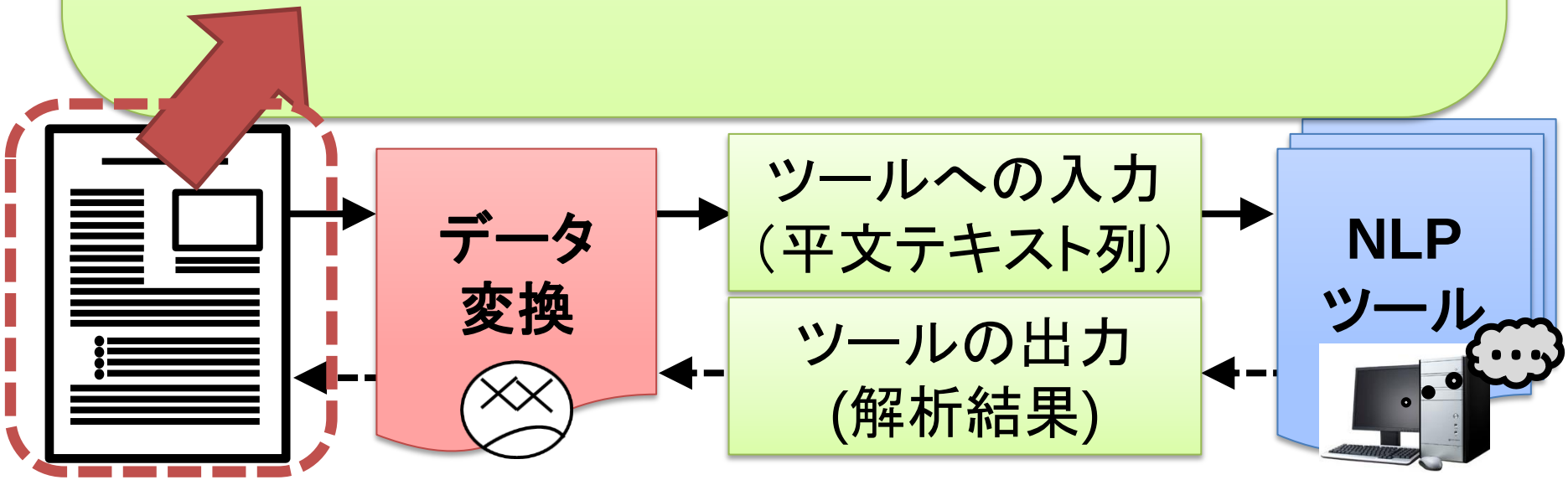
- 実文書をNLPツールで解析したいが…
  - 実文書テキスト: 様々な形で構造化
  - NLPツール: 多くが平文テキストを入力として想定
- データ変換が必要 → 利用者に委ねられる
  - その都度の変換作業は負担
  - 適切に行わないとツールの効率・性能が低下



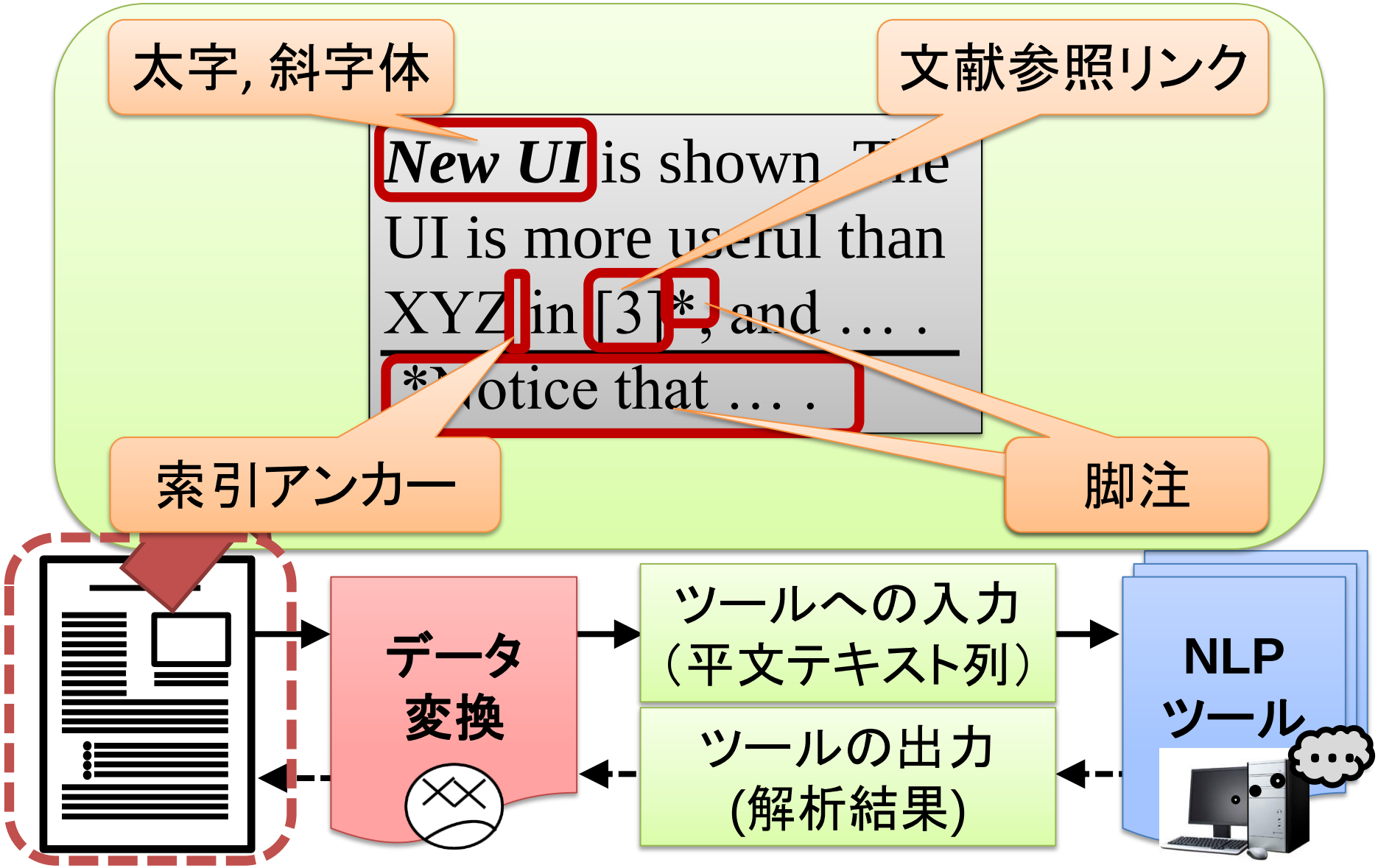
本研究の目的

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

*Notice that … .



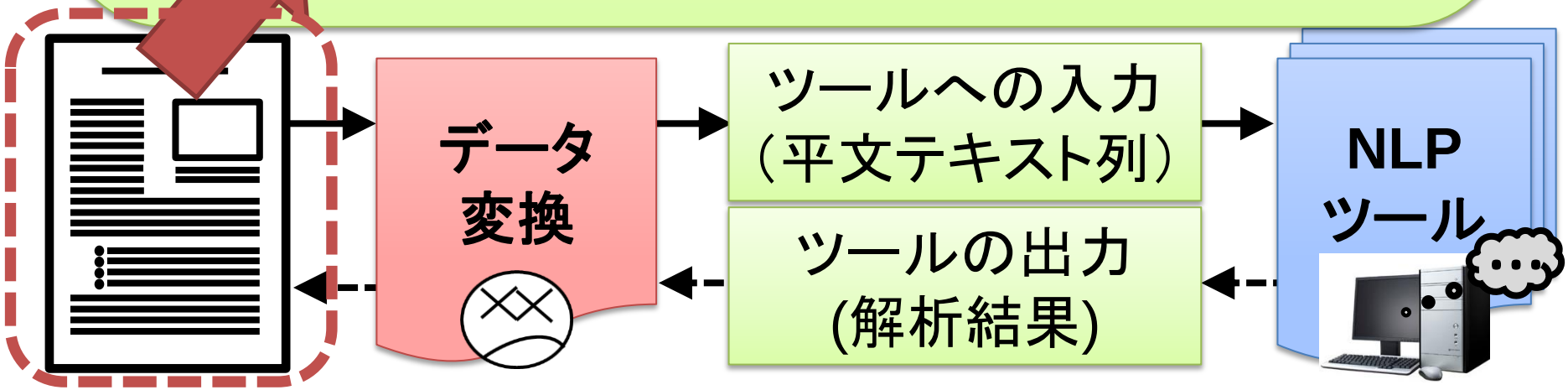
本研究の目的


太字, 斜字体
文献参照リンク
New UI is shown. The
UI is more useful than
XYZ in [3]*, and … .
*Notice that … .
索引アンカー
脚注
データ
変換
ツールへの入力
(平文テキスト列)
ツールの出力
(解析結果)
NLP
ツール

本研究の目的

タグ付テキスト

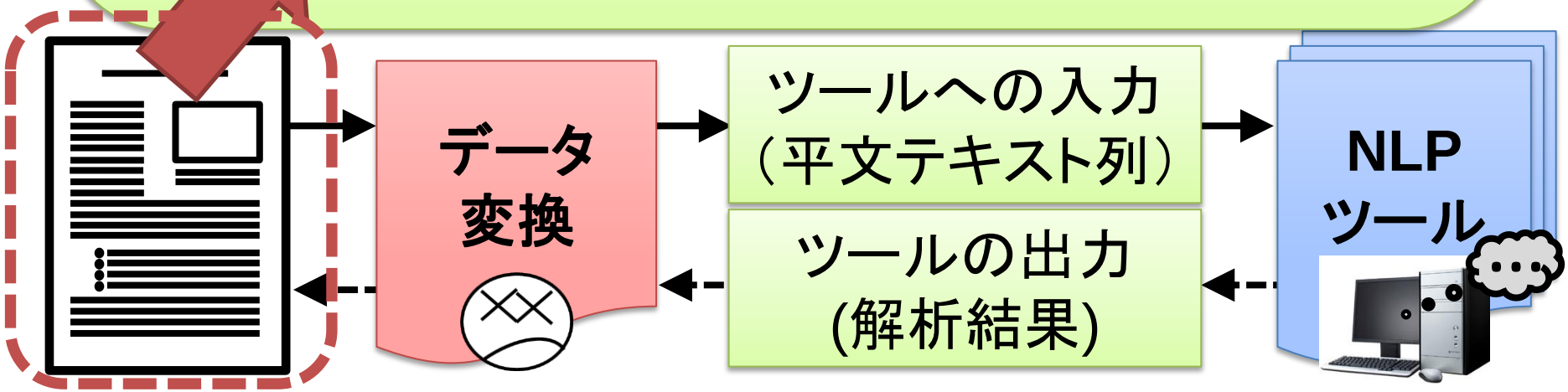
<text>New UI</text> is shown. The UI is more useful than XYZ<index>##</index> in <cite>[…]</cite><note>Notice that … .</note>, and …



本研究の目的

平文？テキスト

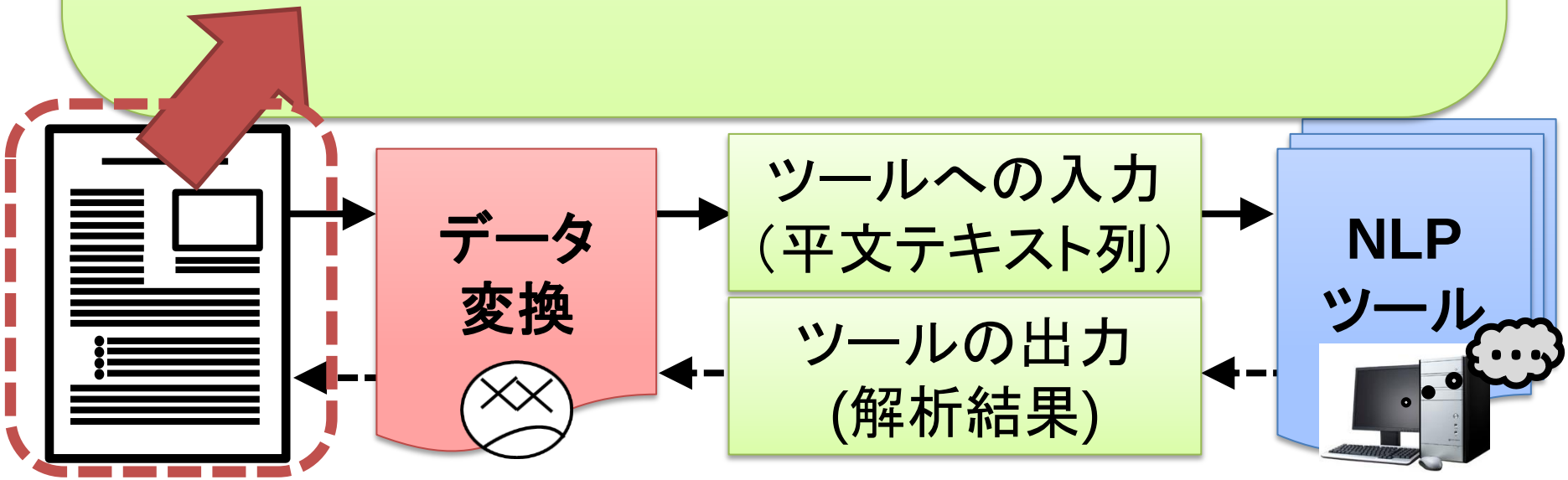
New UI is shown. The UI is more useful than XYZ ## in […] Notice that … . , and …



本研究の目的

平文？テキスト

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



本研究の目的

平文？テキスト

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

解析対象でない断片

文の埋め込み

自然言語以外の構造

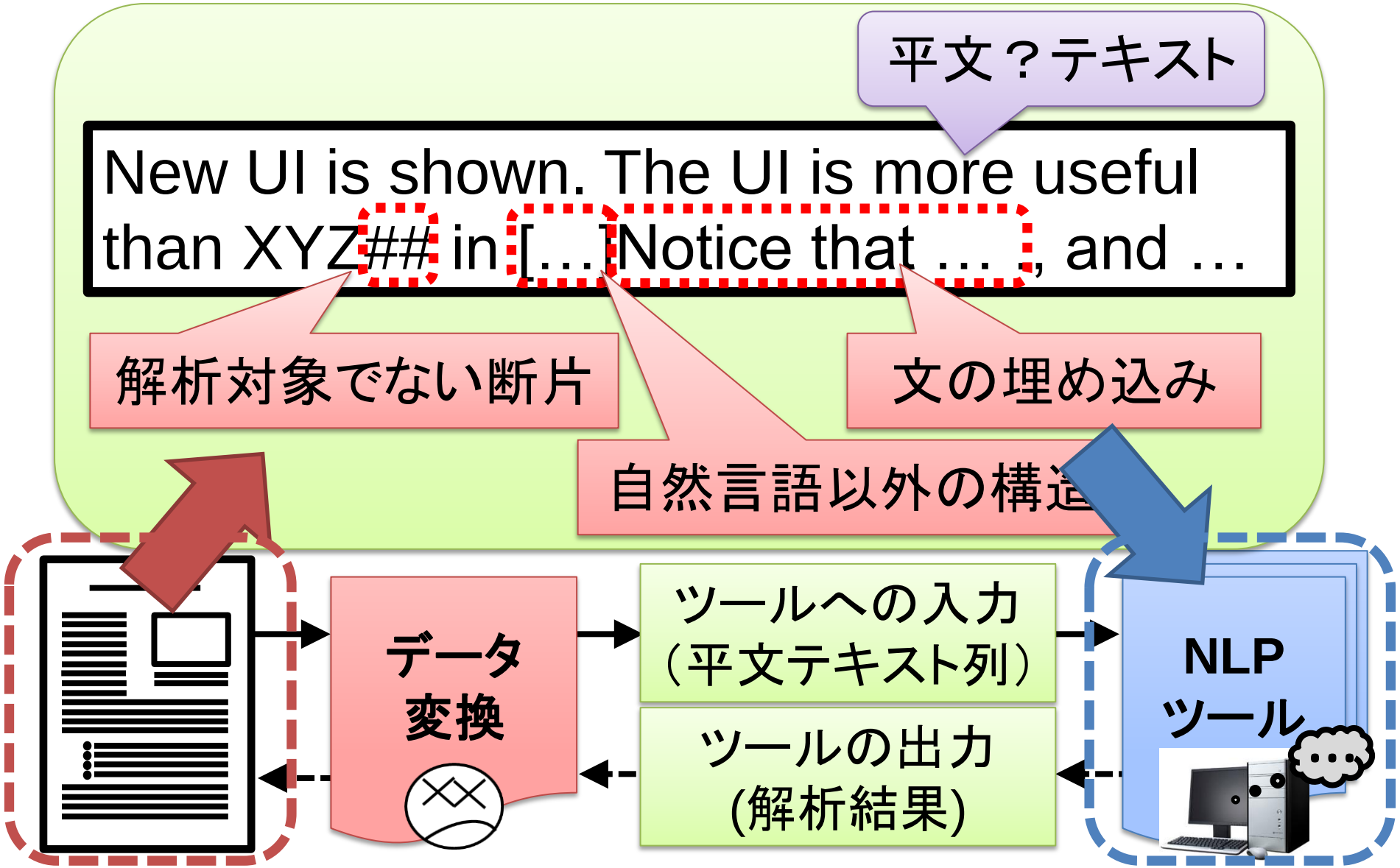
データ変換

ツールへの入力（平文テキスト列）

ツールの出力(解析結果)

NLP ツール

本研究の目的

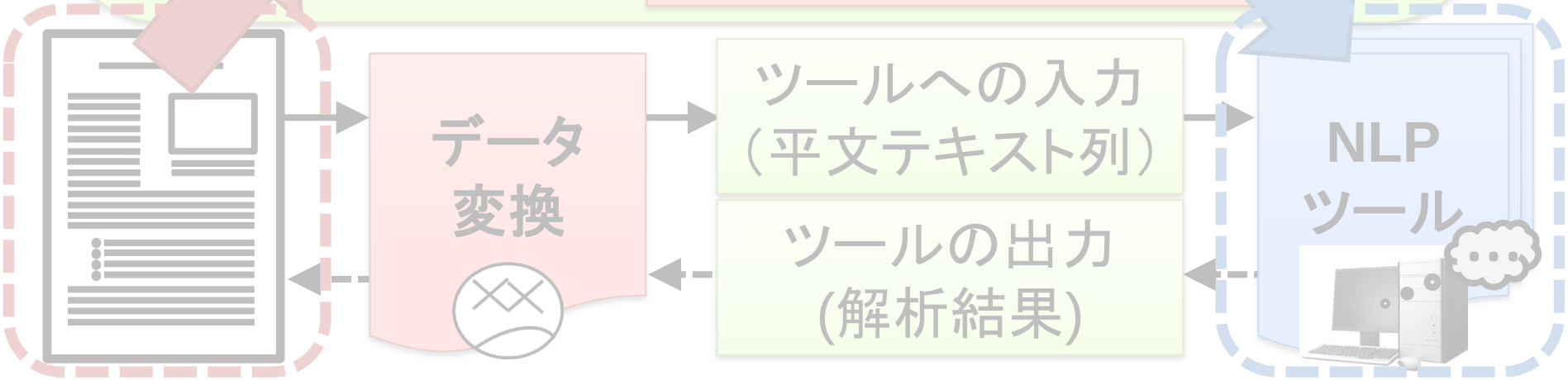

平文？テキスト
New UI is shown. The UI is more useful than XYZ## in […] Notice that … , and …
解析対象でない断片
文の埋め込み
自然言語以外の構造
データ変換
ツールへの入力（平文テキスト列）
ツールの出力(解析結果)
NLP ツール

本研究の目的

平文？テキスト

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

適切な変換の重要性を定量的に検証
→ 実文書解析の議論促進へ



本発表の概要

- 関連研究 – データ変換への注目度の低さ
- 研究方針 – XML文書上での重要性の検証
- 提案枠組 – タグ分類に基づく平文テキスト抽出
- 実験
  - 対象文書からの平文テキスト列抽出の実施
  - NLPツールの高被覆・効率的な適用実現の検証
- 議論
  - 実文書をNLP技術と適切に繋ぐ技術の重要性

対象文書⇔NLPツール入出力間のデータ変換に関する関連研究

- 統一的手法に取り組んだ研究は見当たらず
- ツール側からの変換ツール提供
  - 構文解析器*[1,2]: 品詞タグ付文 → タガーを同梱
    → タガー自体も1文毎に区切られた平文を想定
- 様々な文書・ツールでの解析を扱う統合枠組
  - UIMA*[3]（その実装*[4,5,6]）および GATE*[7]
    → NLPツールをシステムに統合する必要あり

---

*[1] C&C (Clark et al., 2007) *[2] Enju (Ninomiya et al., 2007) *[3] Ferruci et al., 2006
*[4] RASP4UIMA (Andersen et al., 2008) *[5] U-compare (Kano et al., 2011)
*[6] Kachako (Kano, 2012) *[7] Cunningham et al., 2013

本研究の方針

【目的】適切なデータ変換技術の重要性を検証

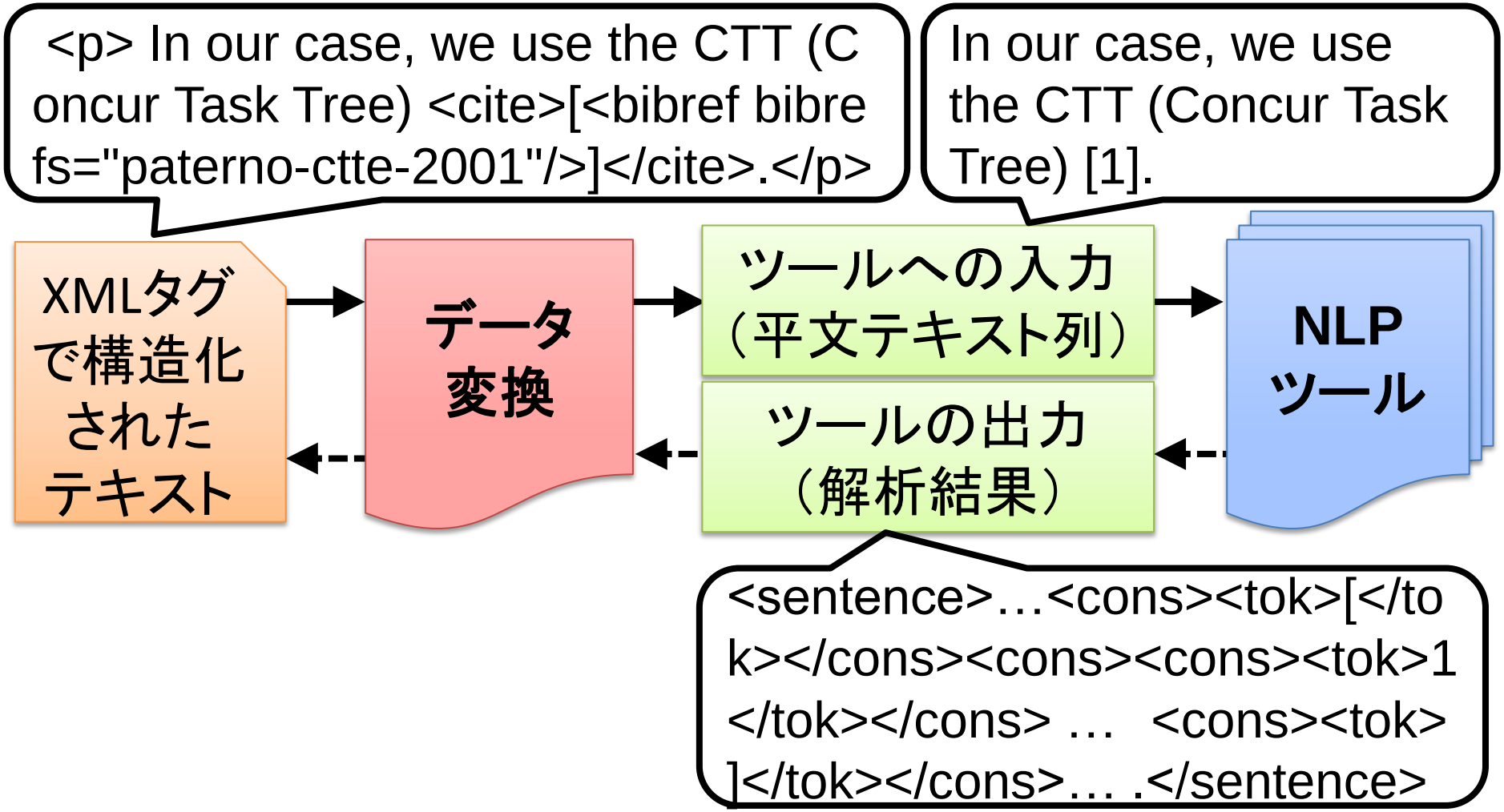
【方針１】対象をXML文書に絞る

- 平文を超える構造情報はタグで表されると仮定

【方針２】既存のNLPツールを、それ自体に手を加えず対象文書に適用可能にする枠組を提案

→有用性の検証から、データ変換の重要性を定量的に示す

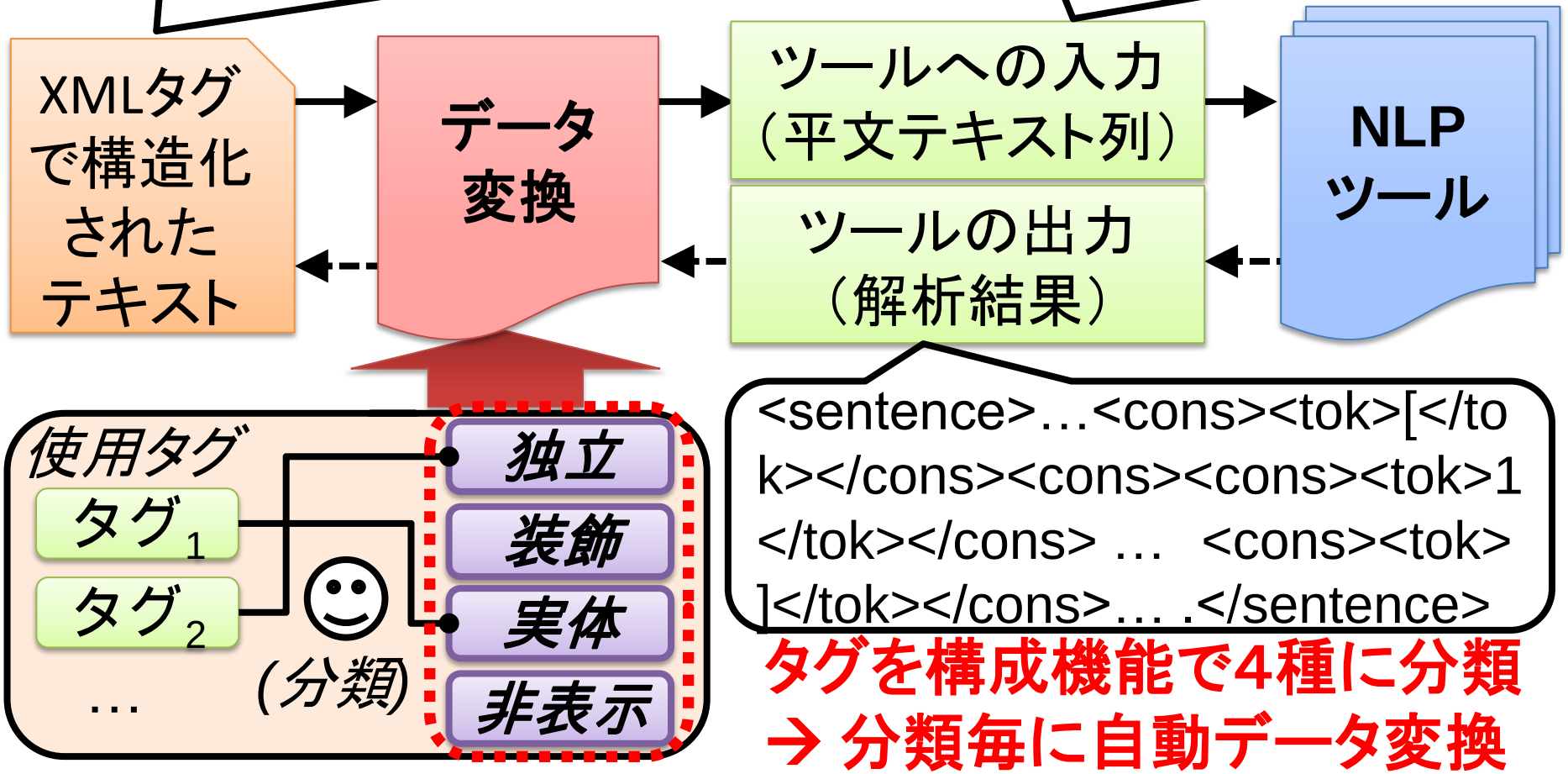
提案枠組概要


 In our case, we use the CTT (C oncur Task Tree) <cite>[<bibref bibre fs="paterno-ctte-2001"/>]</cite>.
In our case, we use the CTT (Concur Task Tree) [1].
XMLタグで構造化されたテキスト
データ変換
ツールへの入力（平文テキスト列）
ツールの出力（解析結果）
NLP ツール
<sentence>…<cons><tok>[</to k></cons><cons><cons><tok>1 </tok></cons> … <cons><tok>]</tok></cons>… .</sentence>

提案枠組概要

 In our case, we use the CTT (C oncur Task Tree) <cite>[<bibref bibre fs="paterno-ctte-2001"/>]</cite>.

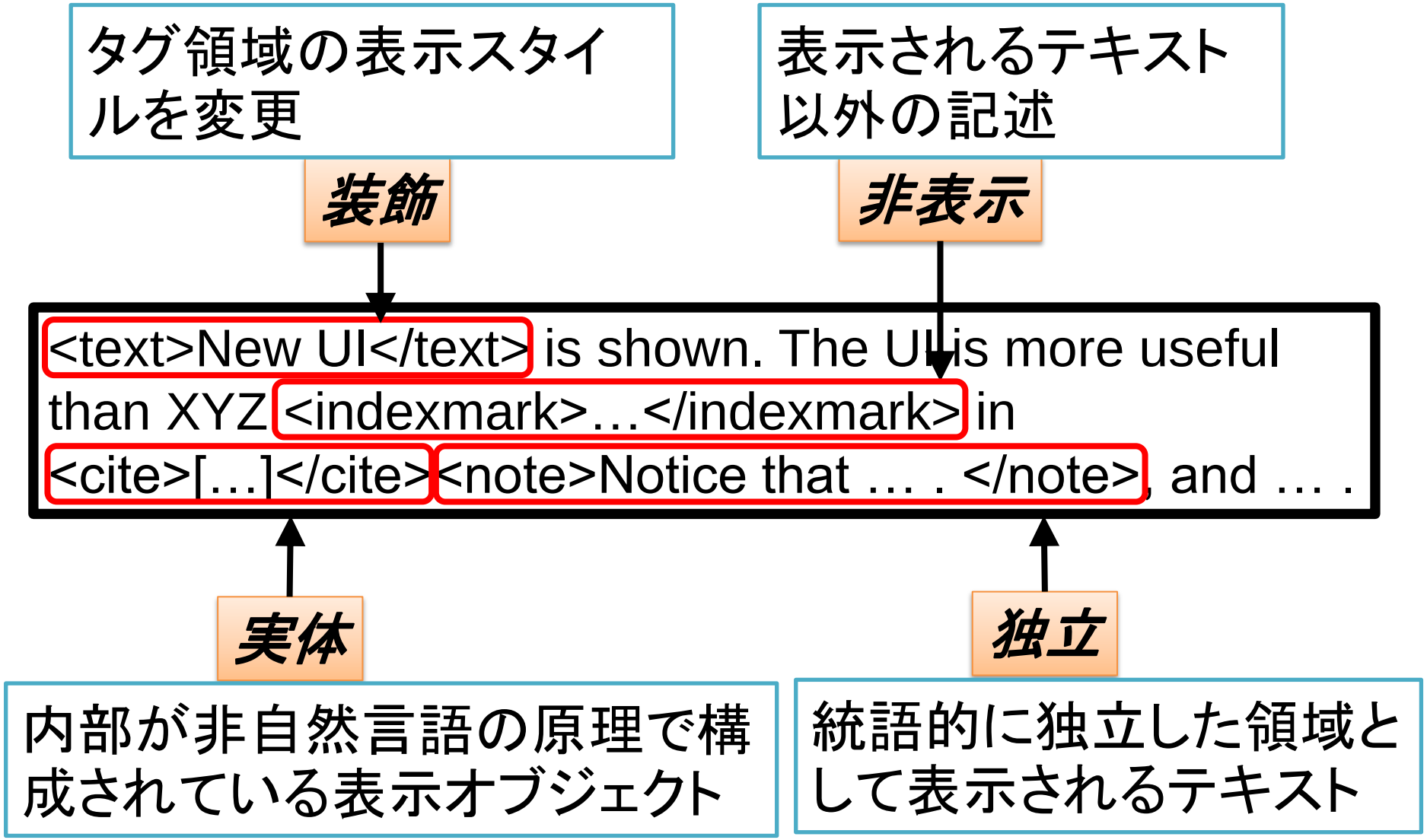
In our case, we use the CTT (Concur Task Tree) [1].



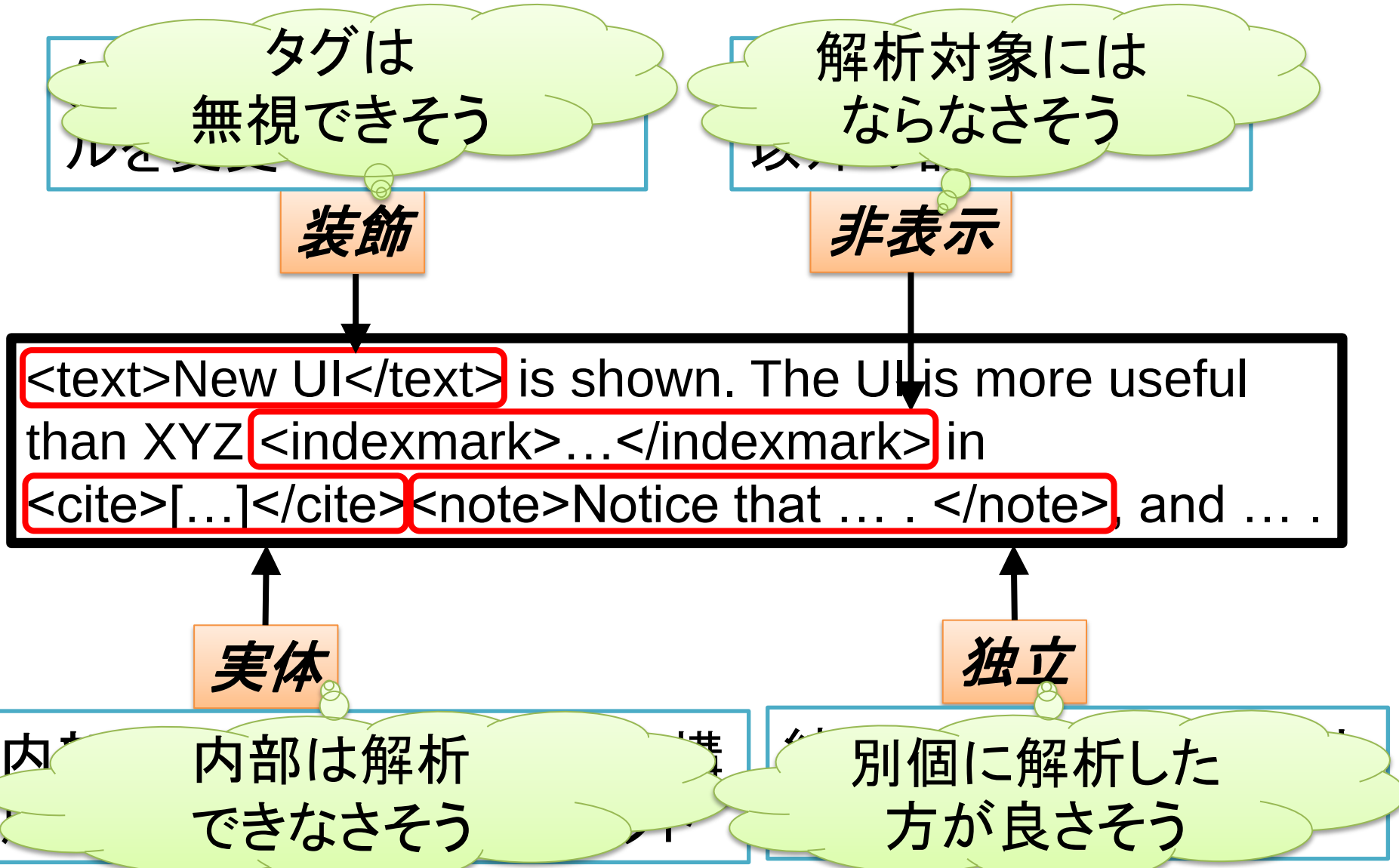
タグを構成機能で4種に分類

→ 分類毎に自動データ変換

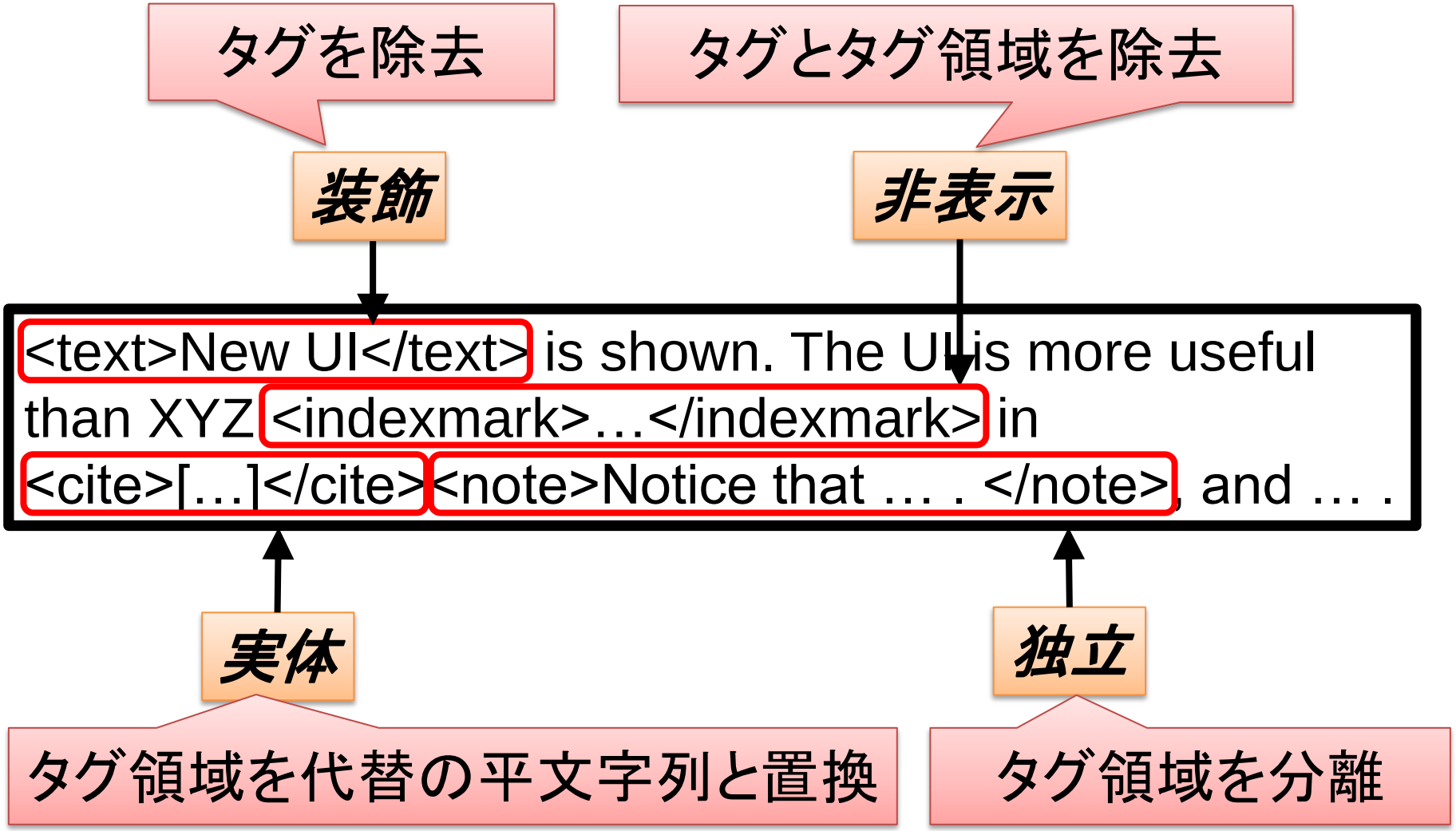
観察: 4種のテキスト構成機能



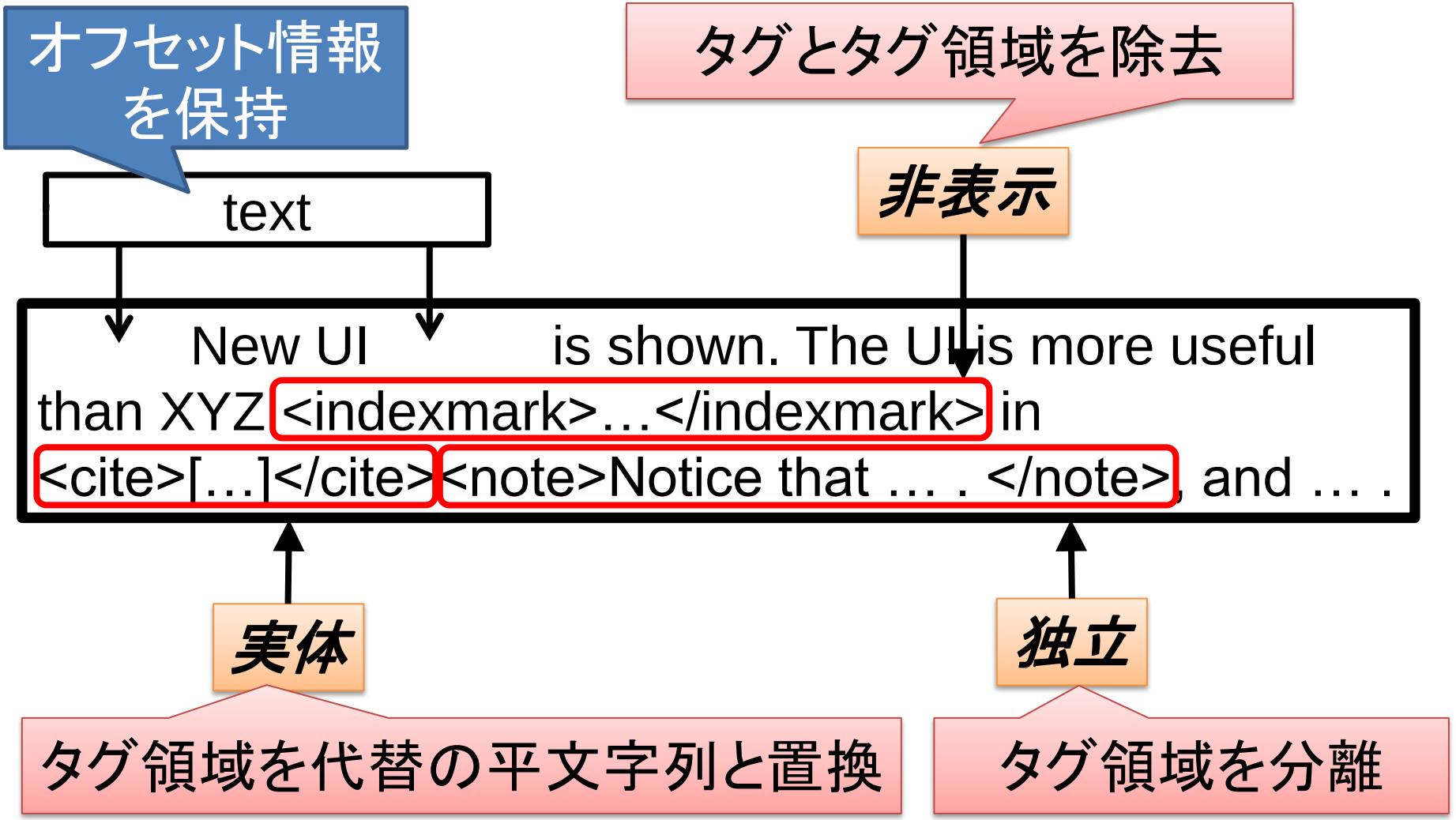
NLPツール適用の直感



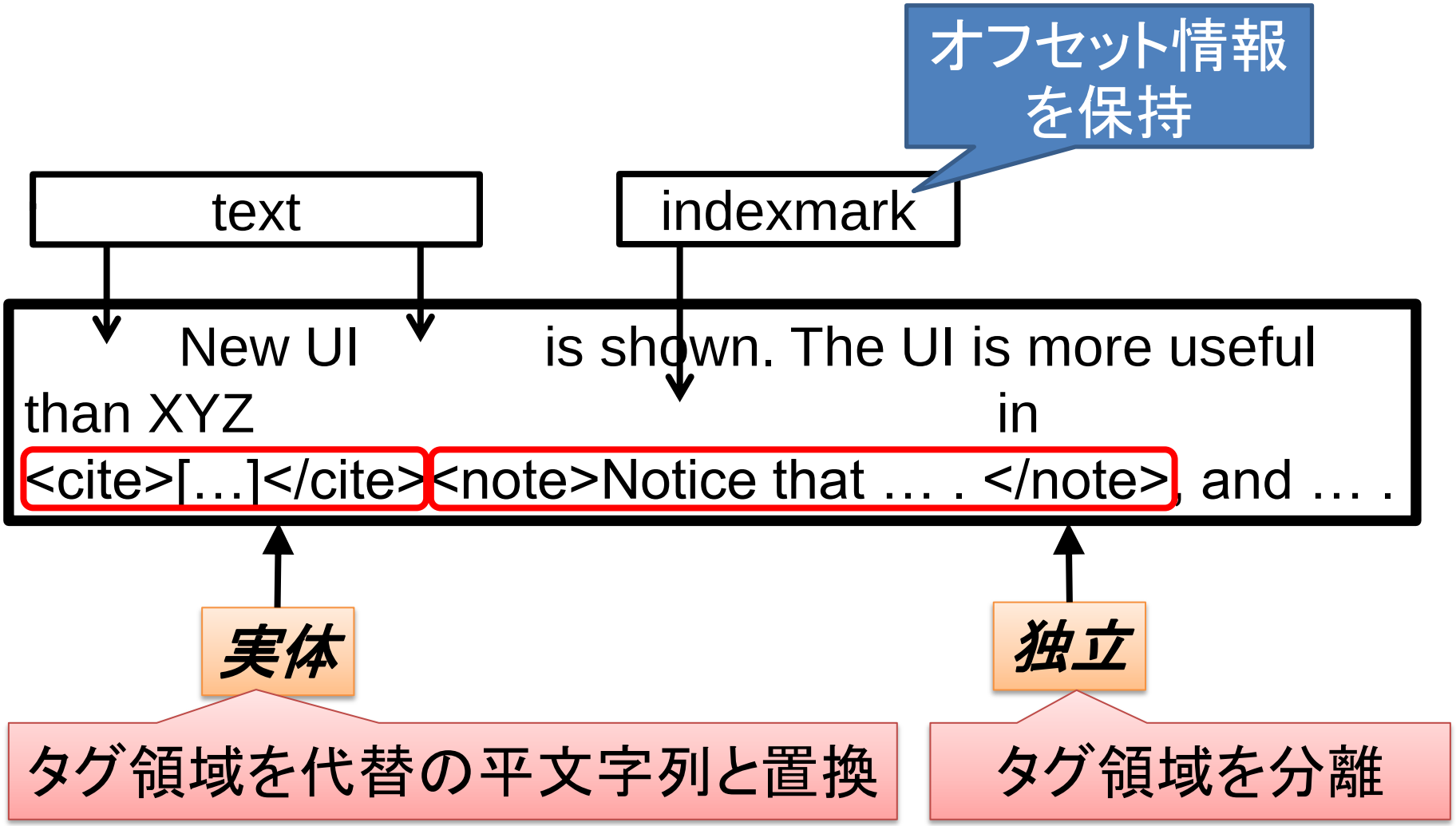
データ変換戦略 (1/3): NLPツール入力への変換



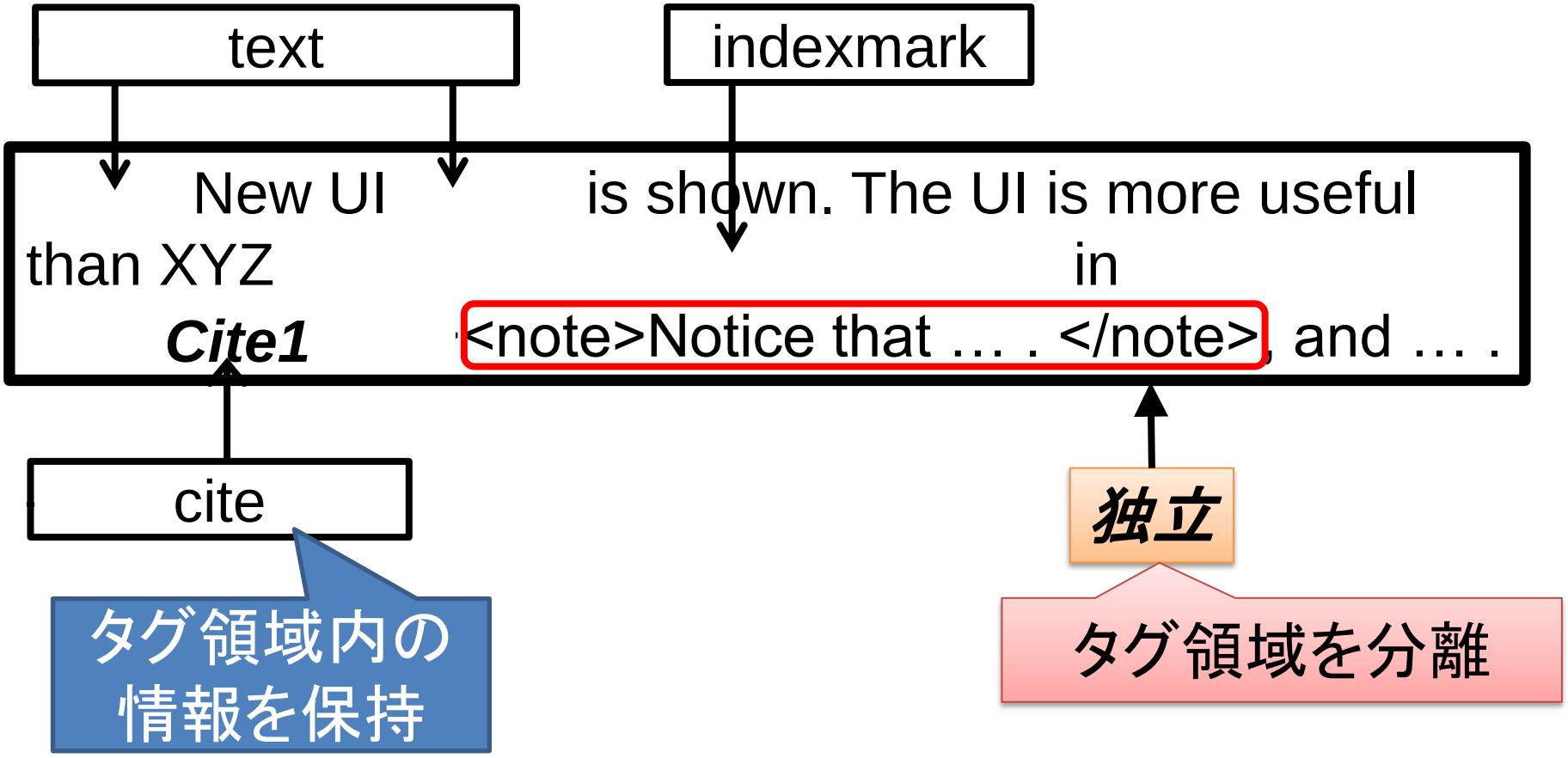
データ変換戦略 (1/3): NLPツール入力への変換



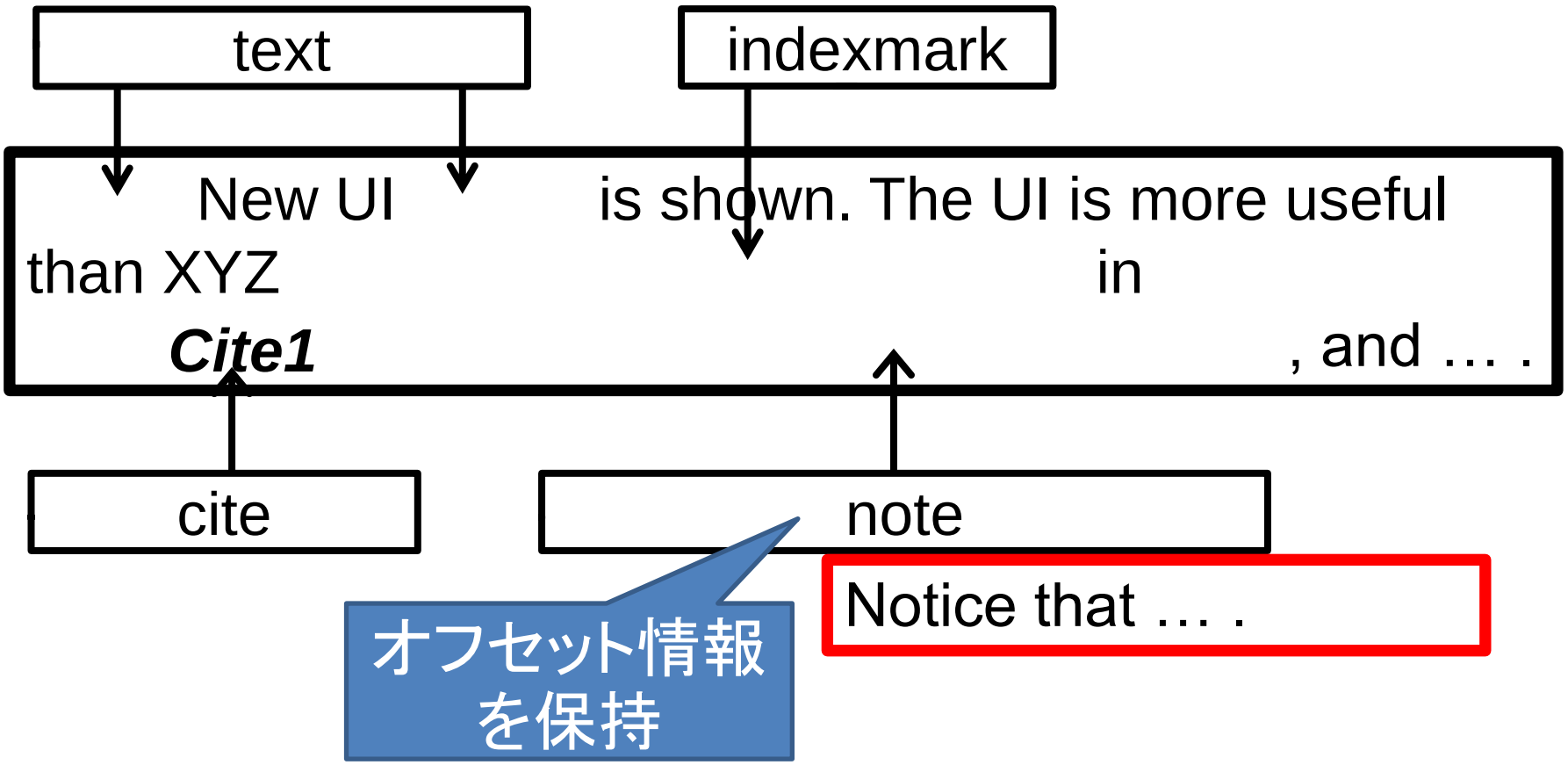
データ変換戦略 (1/3): NLPツール入力への変換



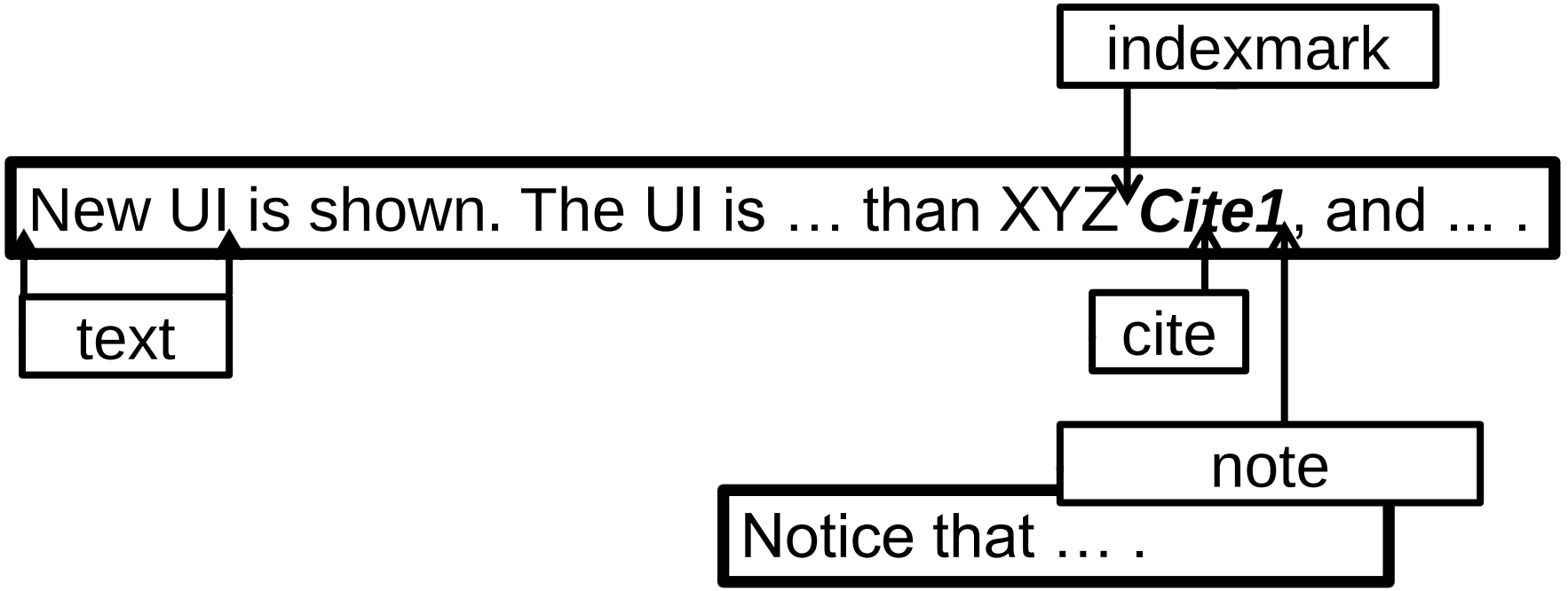
データ変換戦略 (1/3): NLPツール入力への変換



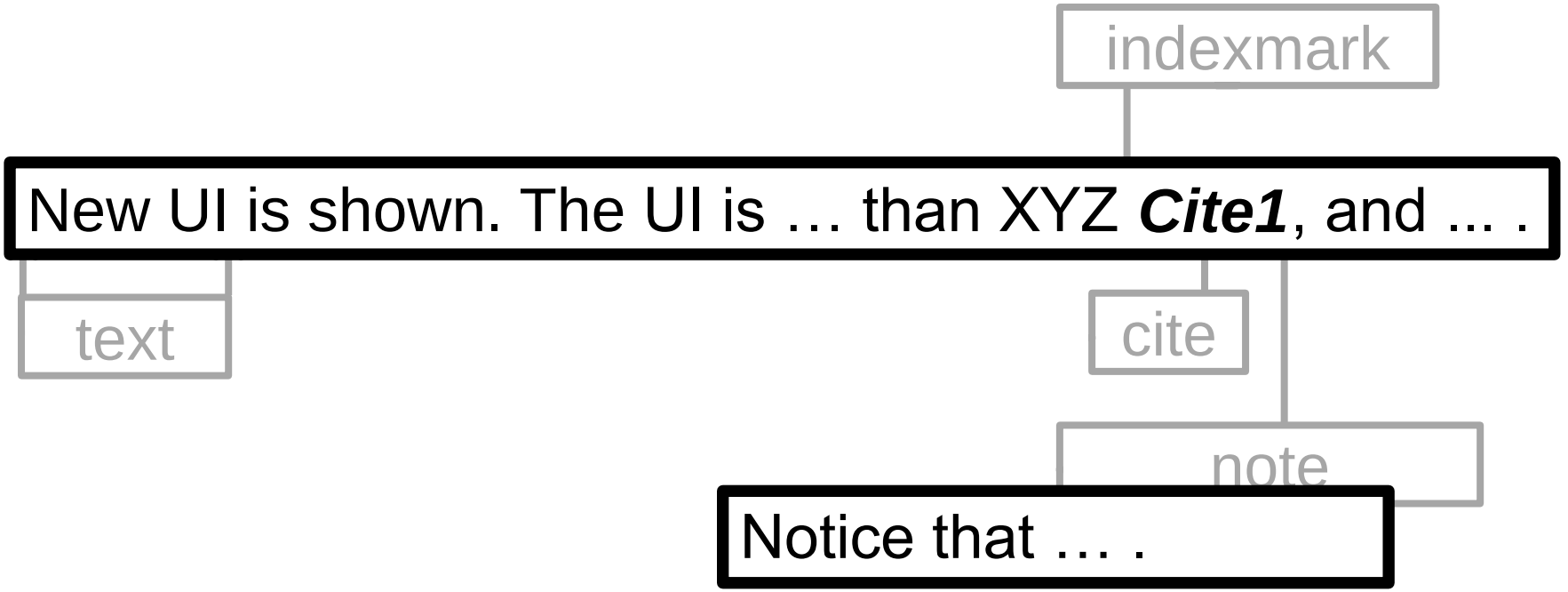
データ変換戦略 (1/3): NLPツール入力への変換



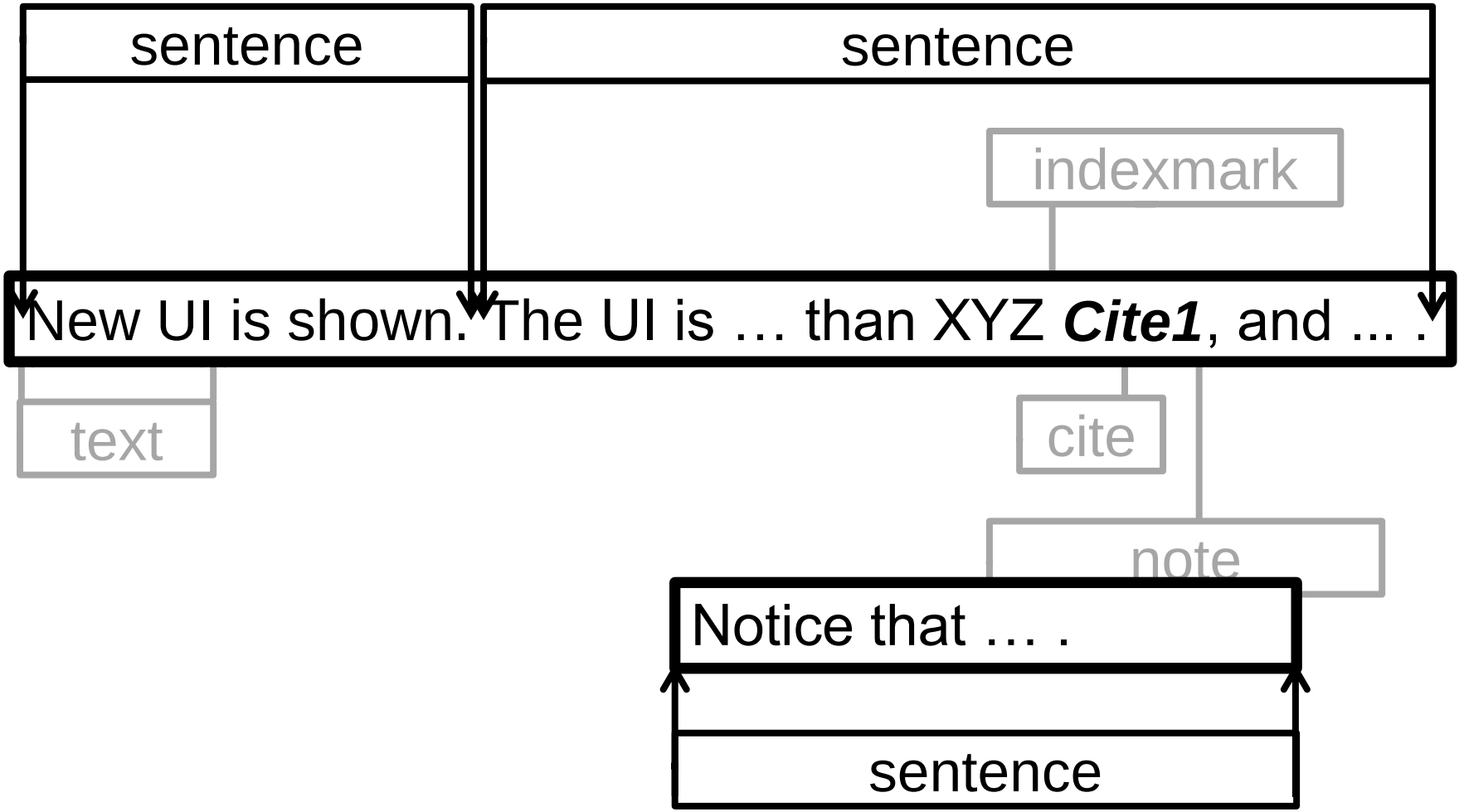
データ変換戦略 (1/3):
NLPツール入力への変換



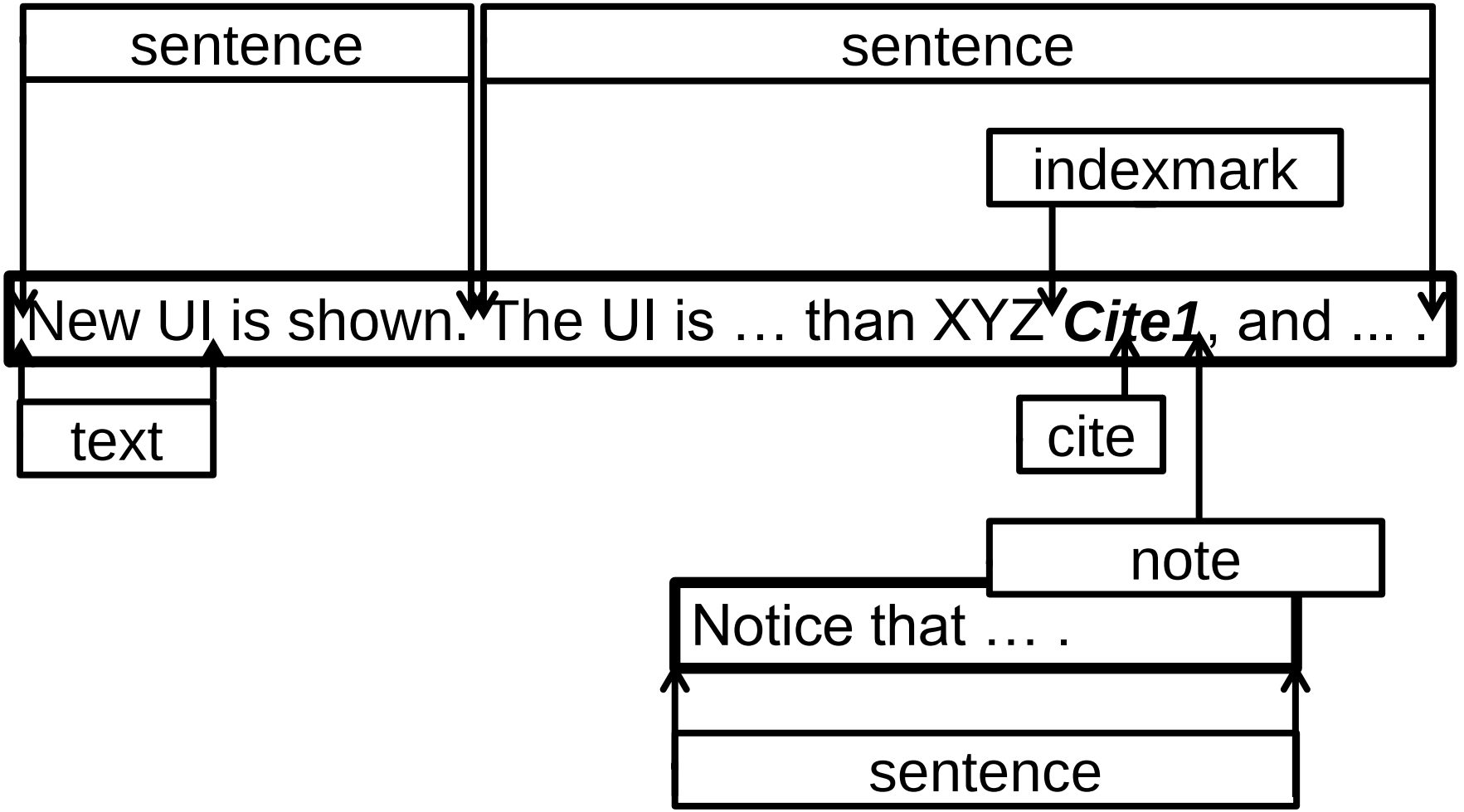
データ変換戦略 (1/3):
NLPツール入力への変換



データ変換戦略 (2/3): NLPツールの適用 (文区切り器)



データ変換戦略 (3/3): 元のタグの復帰



データ変換戦略 (3/3): 元のタグの復帰

<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 … .**</sentence>**

タグ分類と変換戦略まとめ

分類	分類基準	変換戦略
独立	周囲のテキストから統語的に独立した領域	(A)を(B)から分離→(A), (B)にツール適用後, (A)を(B)に復帰
装飾	領域内の表示スタイルのみ変更	(A')のみをテキスト(B)から除去→ツール適用後に(A)を(B)に復帰
実体	テキスト構成要素として扱われる最小のオブジェクト単位を表す	(A)を(C)で置換→ツール適用後に(C)を(A)に復元
非表示	表示方法の設定や追加情報を記述する	(A)を(B)から除去→ツール(A)を(B)に復帰

(A): タグとタグ領域 (A'): タグ (B) 元テキスト (C) 代替文字列

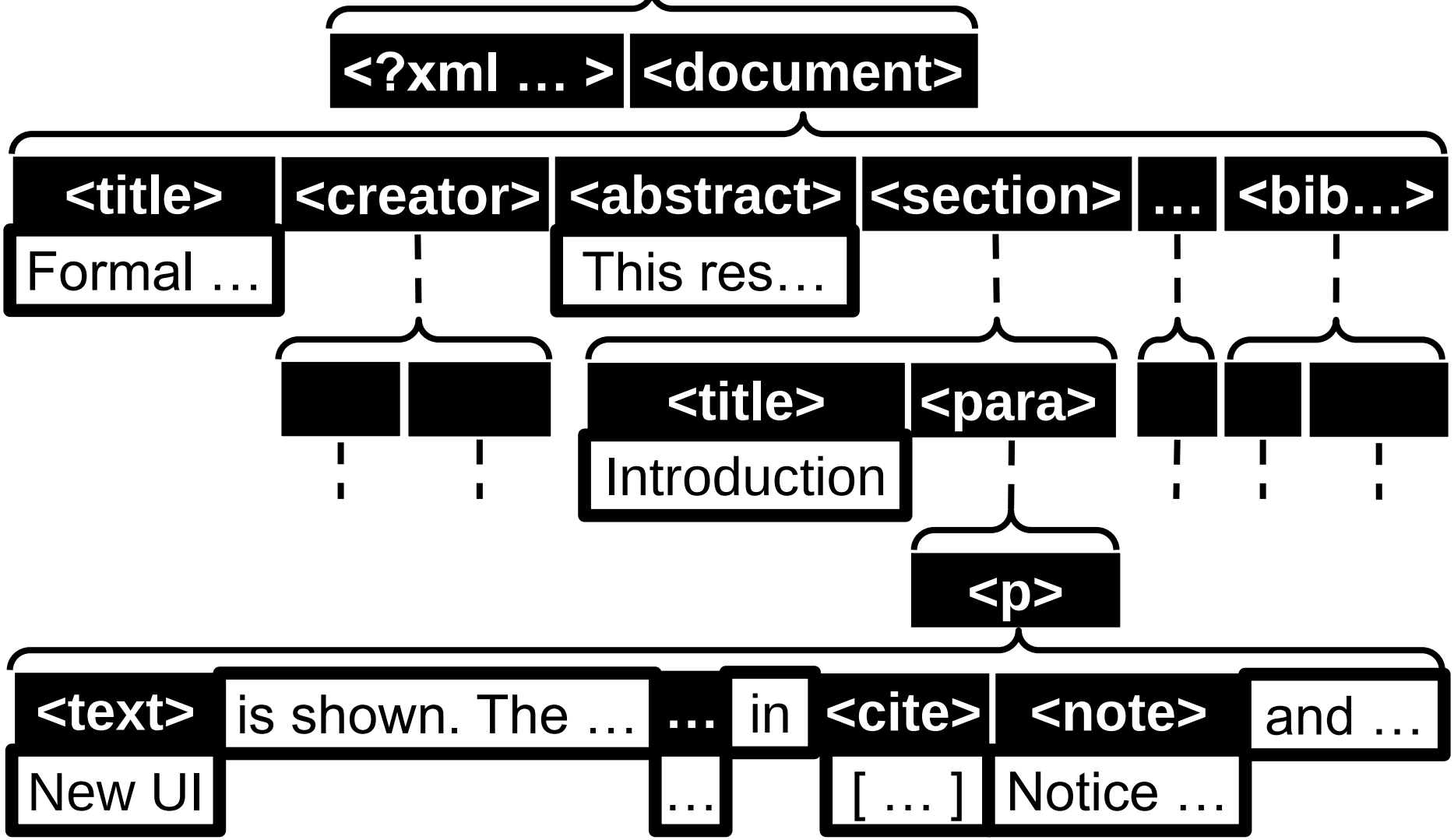
タグ分類の効率性

XML文書

```
<?xml …>
<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>
```

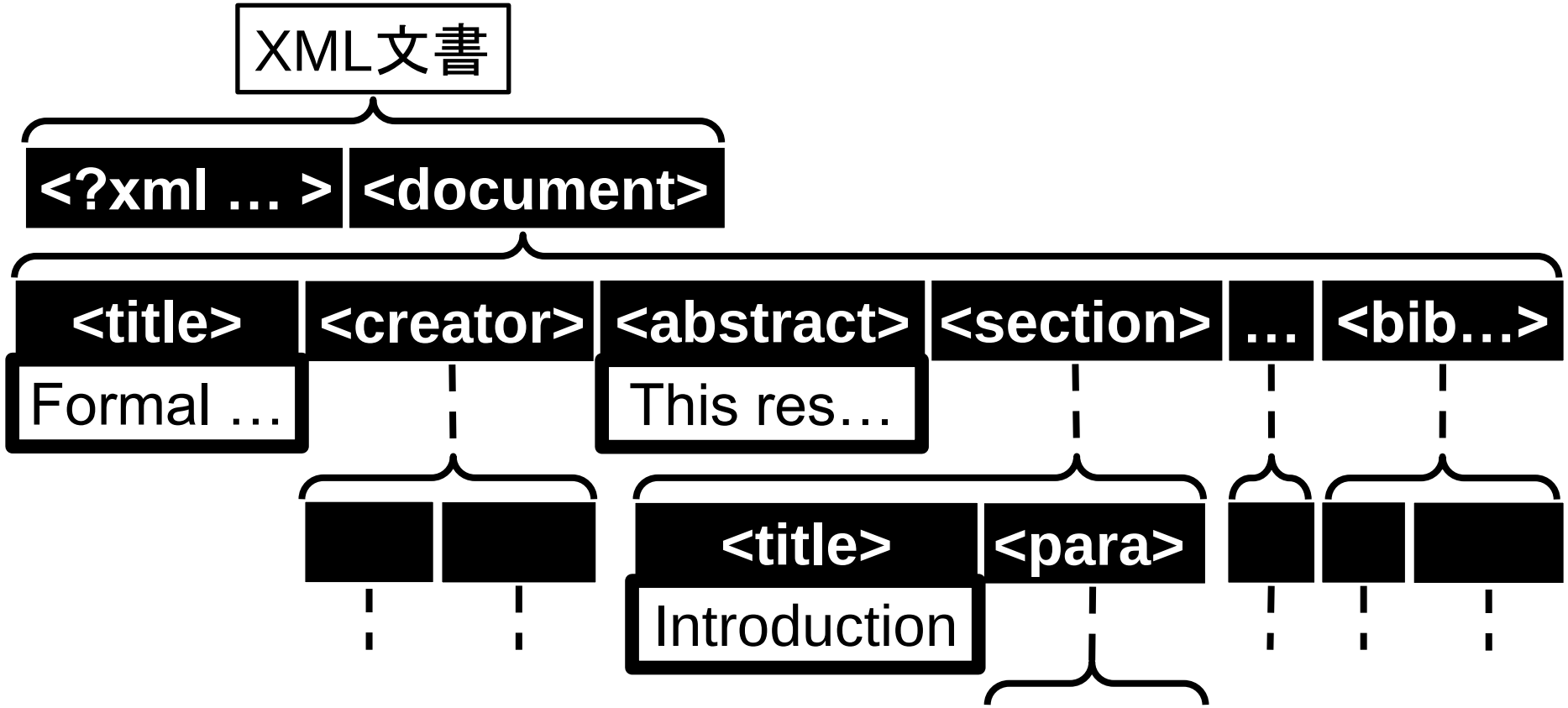
XML文書を直接観察して
タグ分類するのは非効率的

観察: XML文書の多層構造



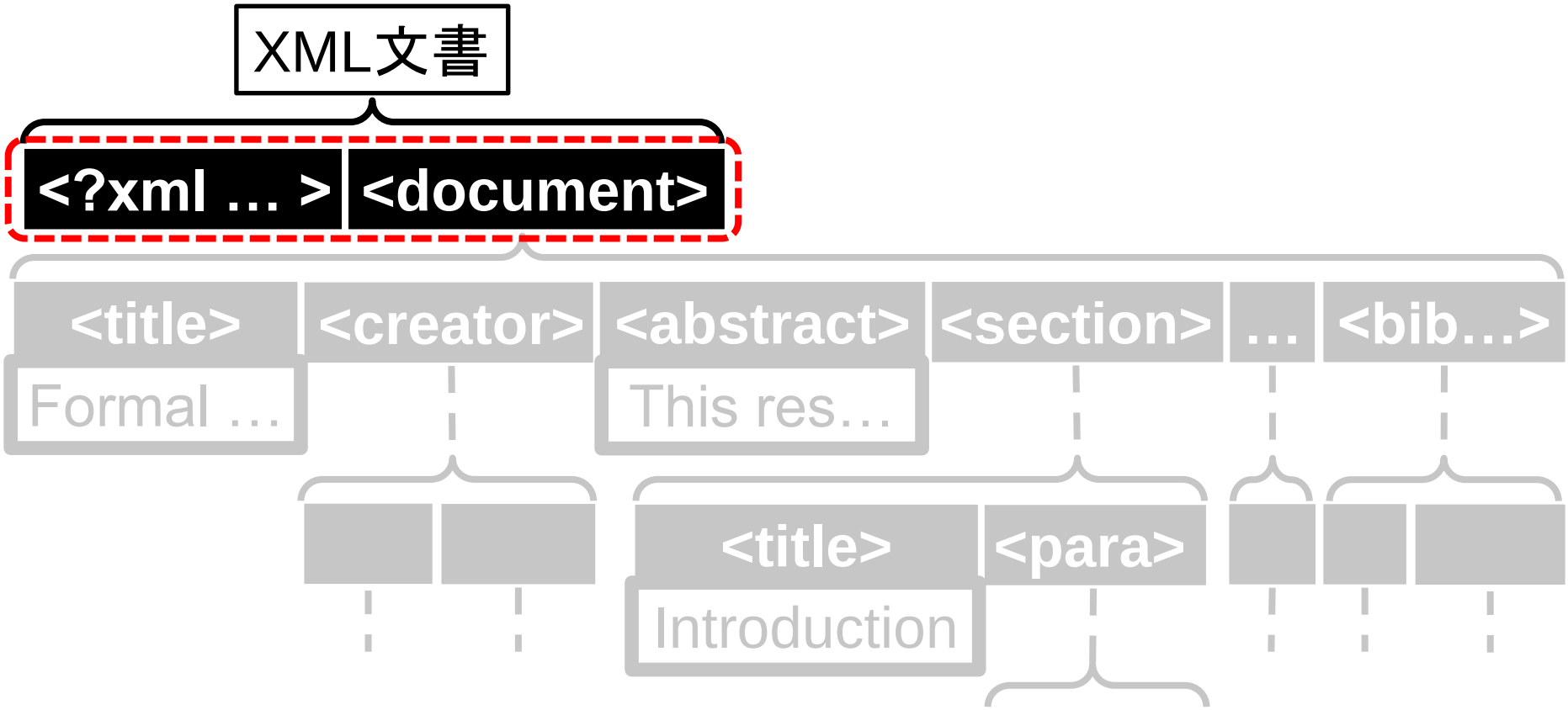
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



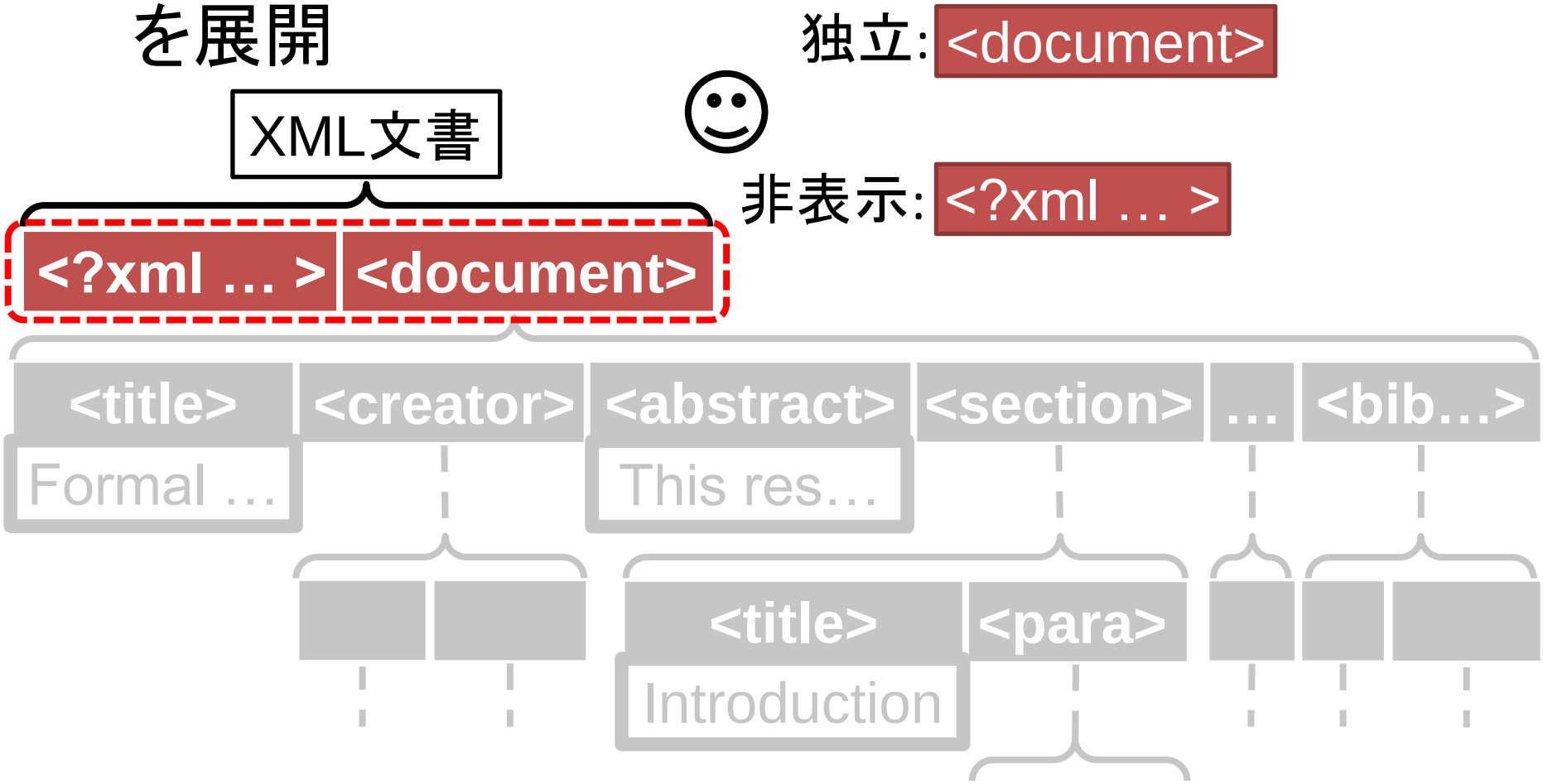
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



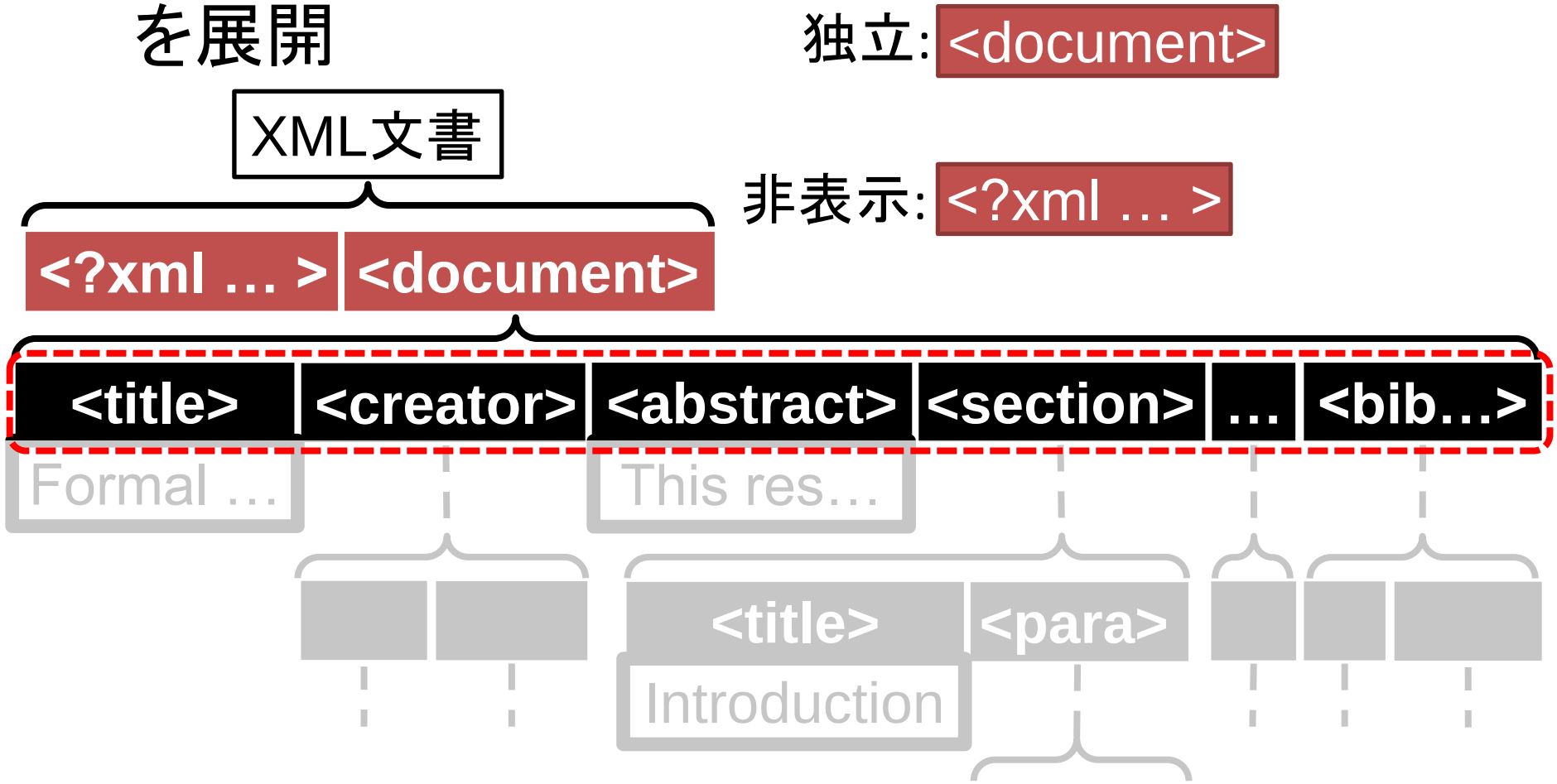
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



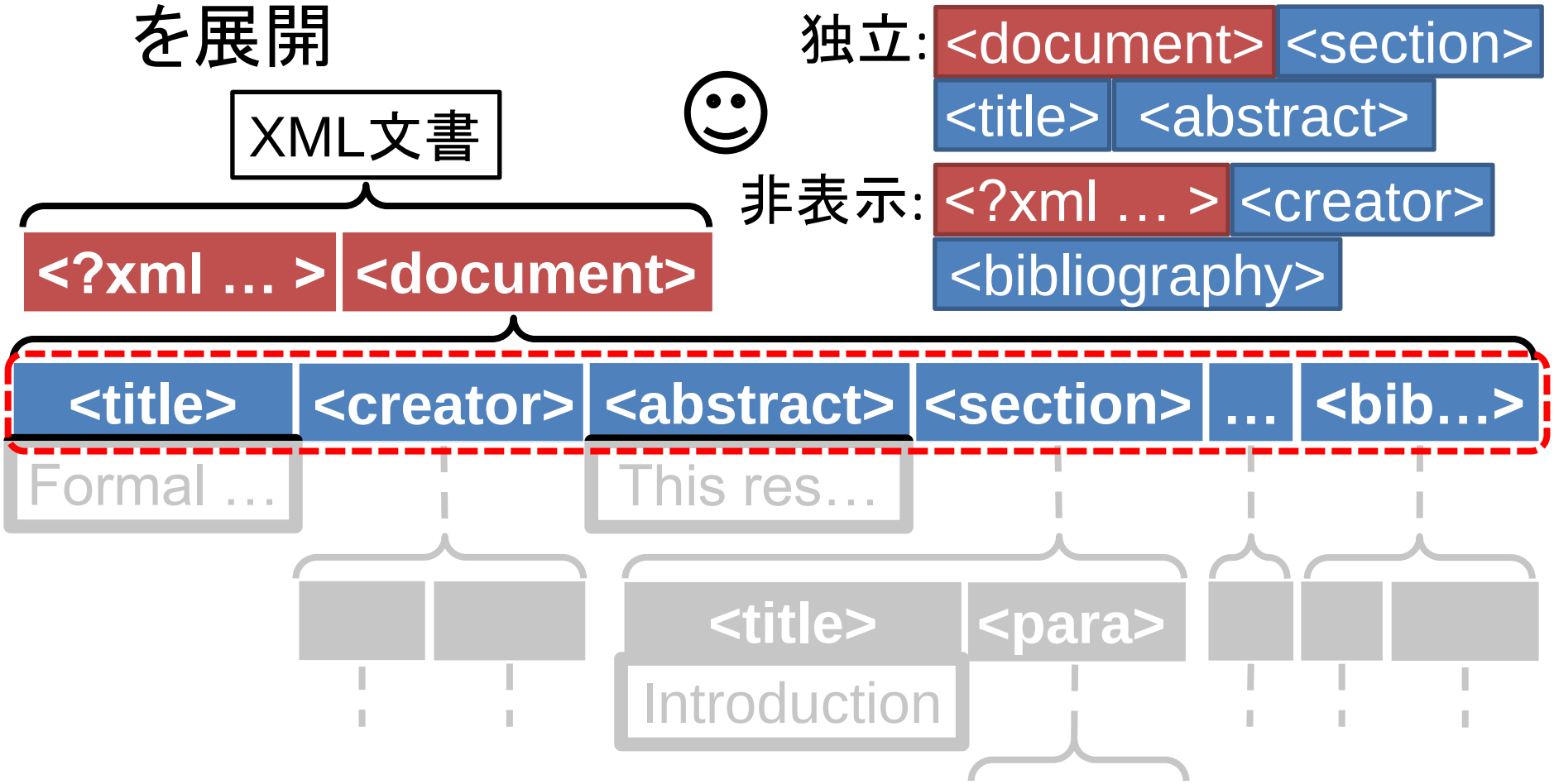
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



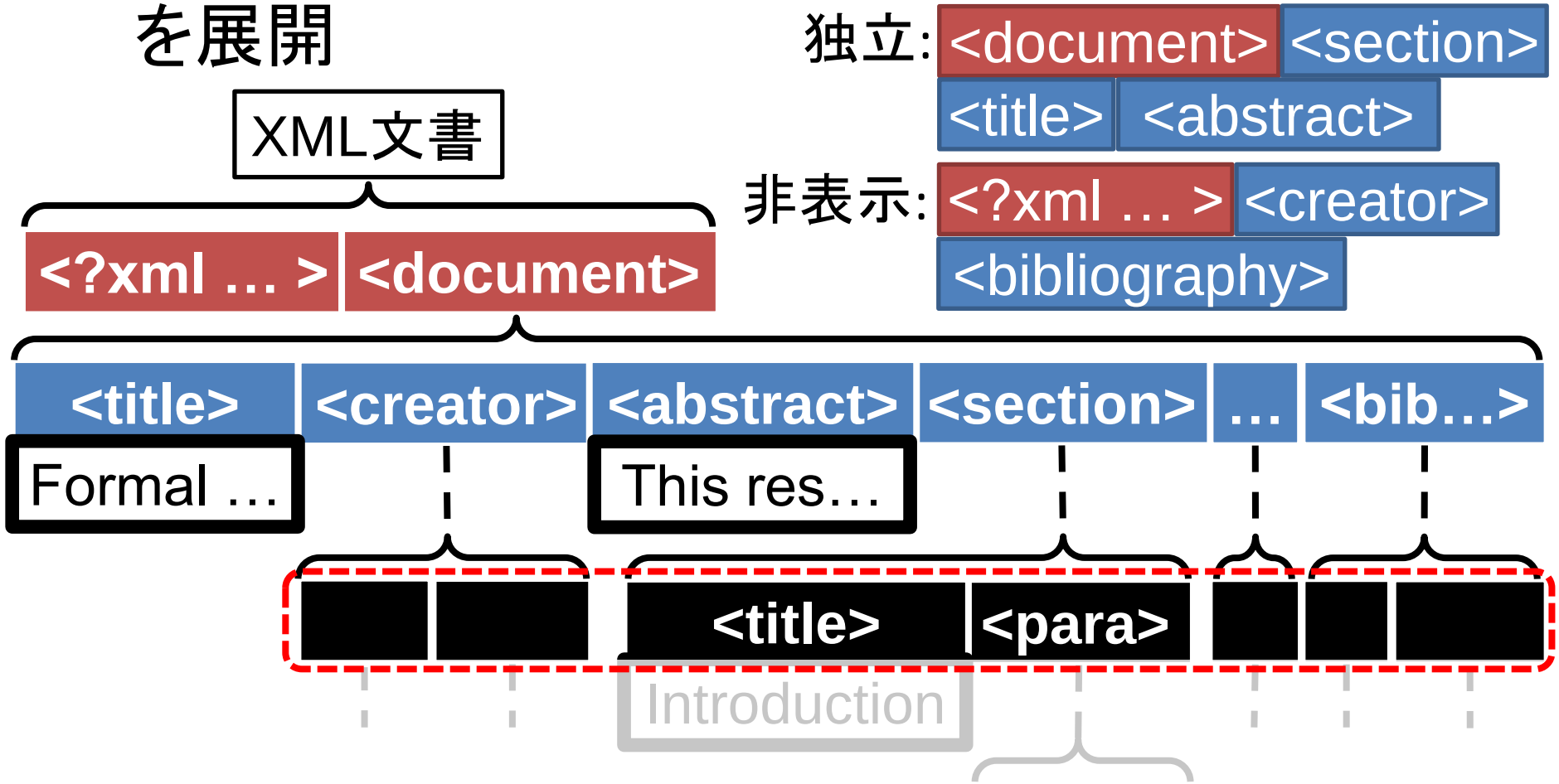
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



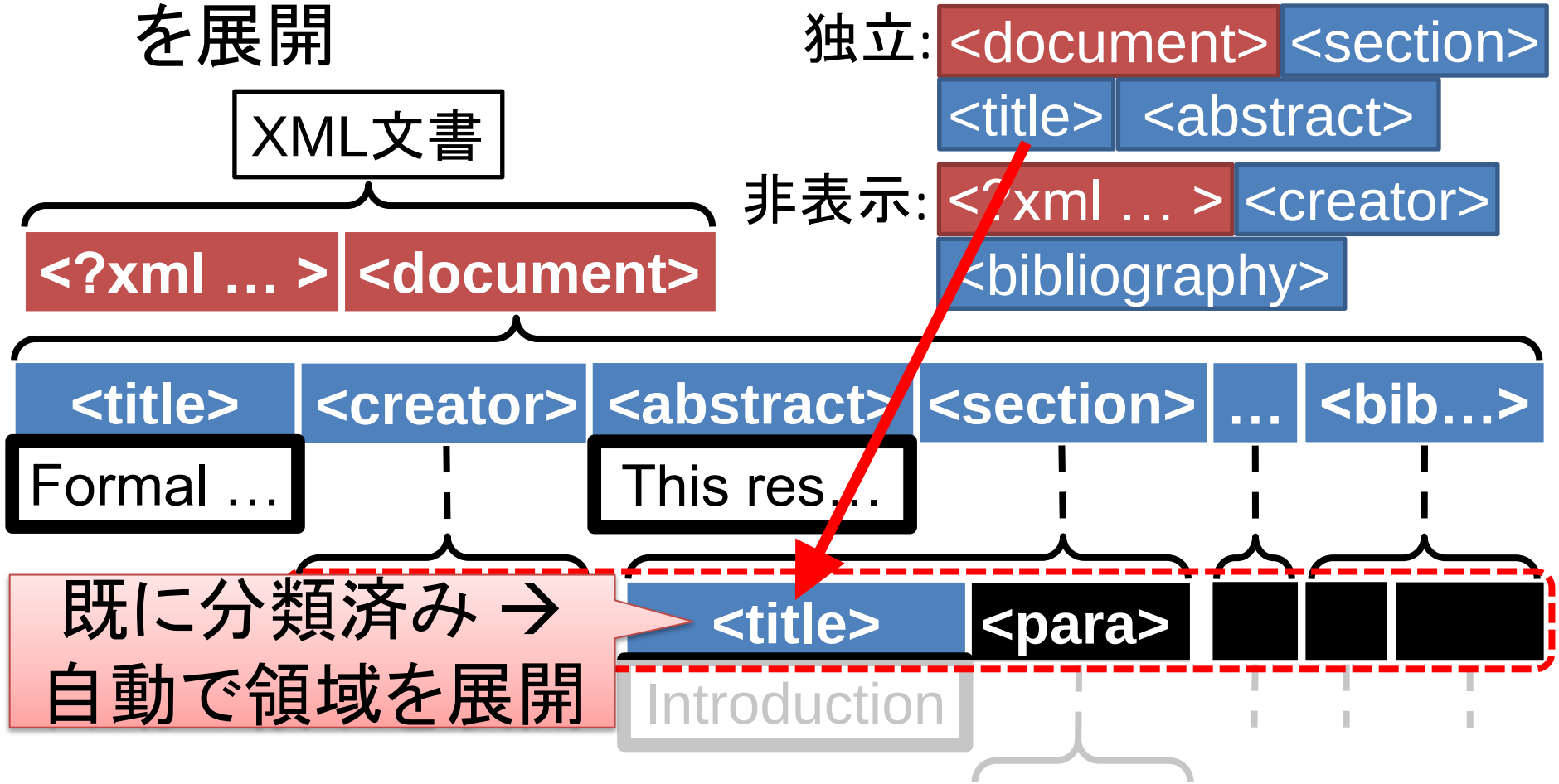
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



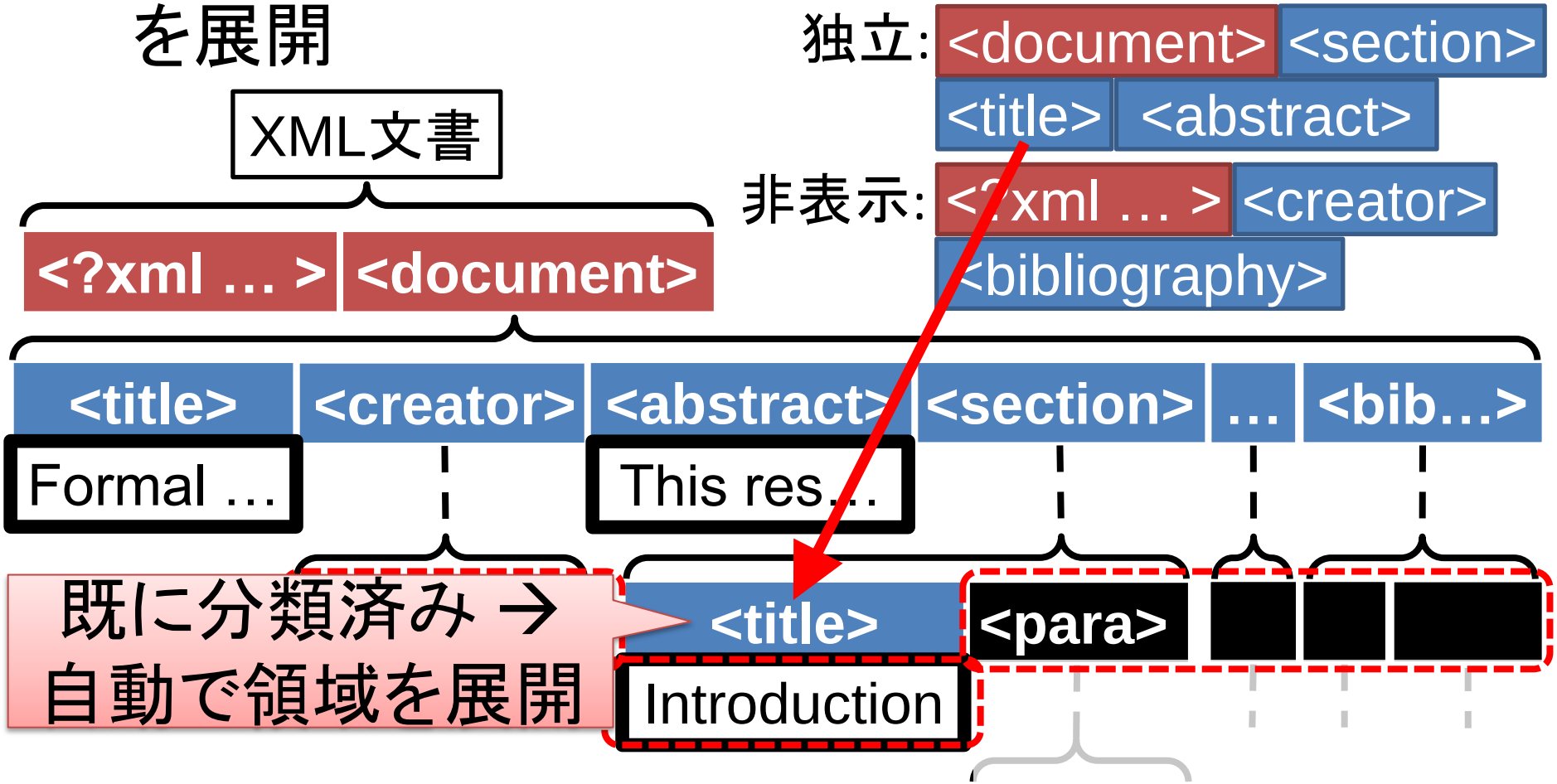
タグ分類の手続き

- 最上層のタグから, 既に分類済みのタグ領域を展開



タグ分類の手続き

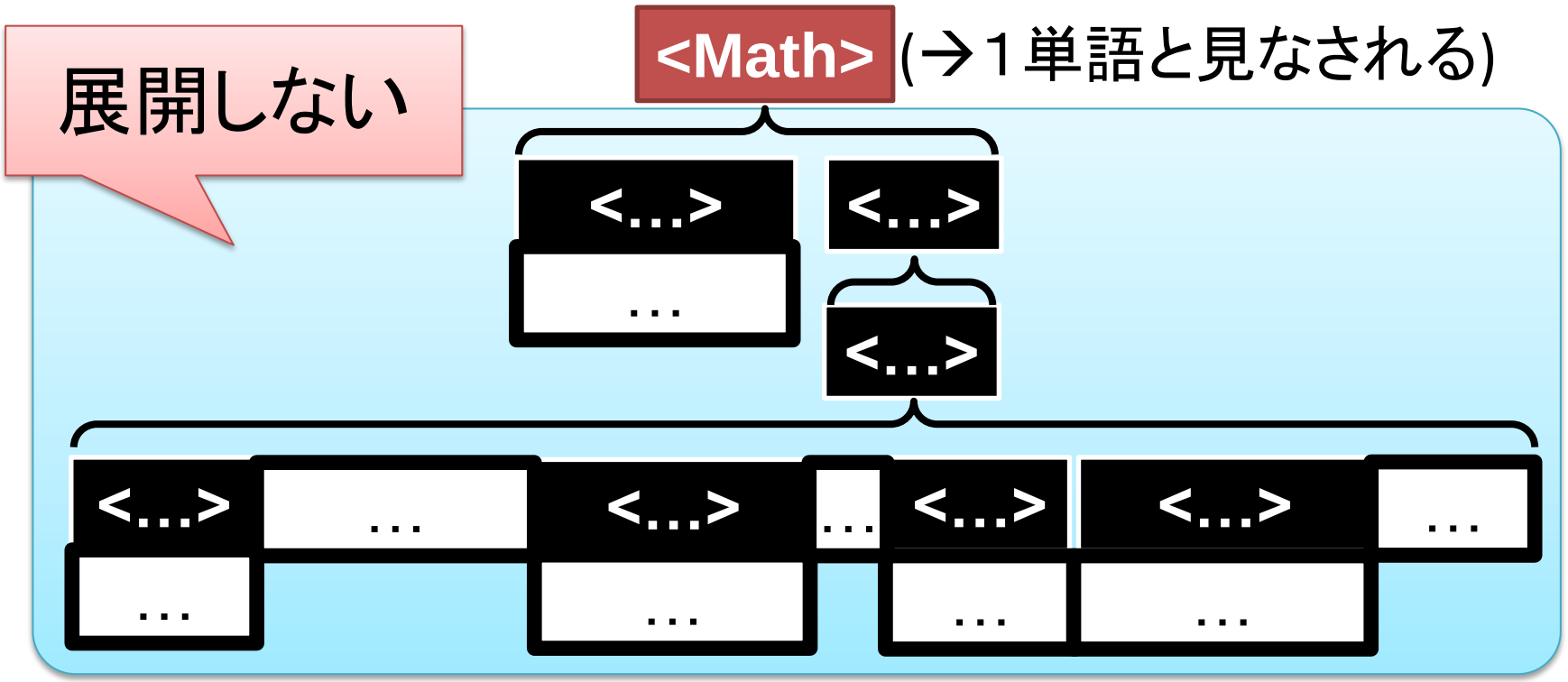
- 最上層のタグから, 既に分類済みのタグ領域を展開



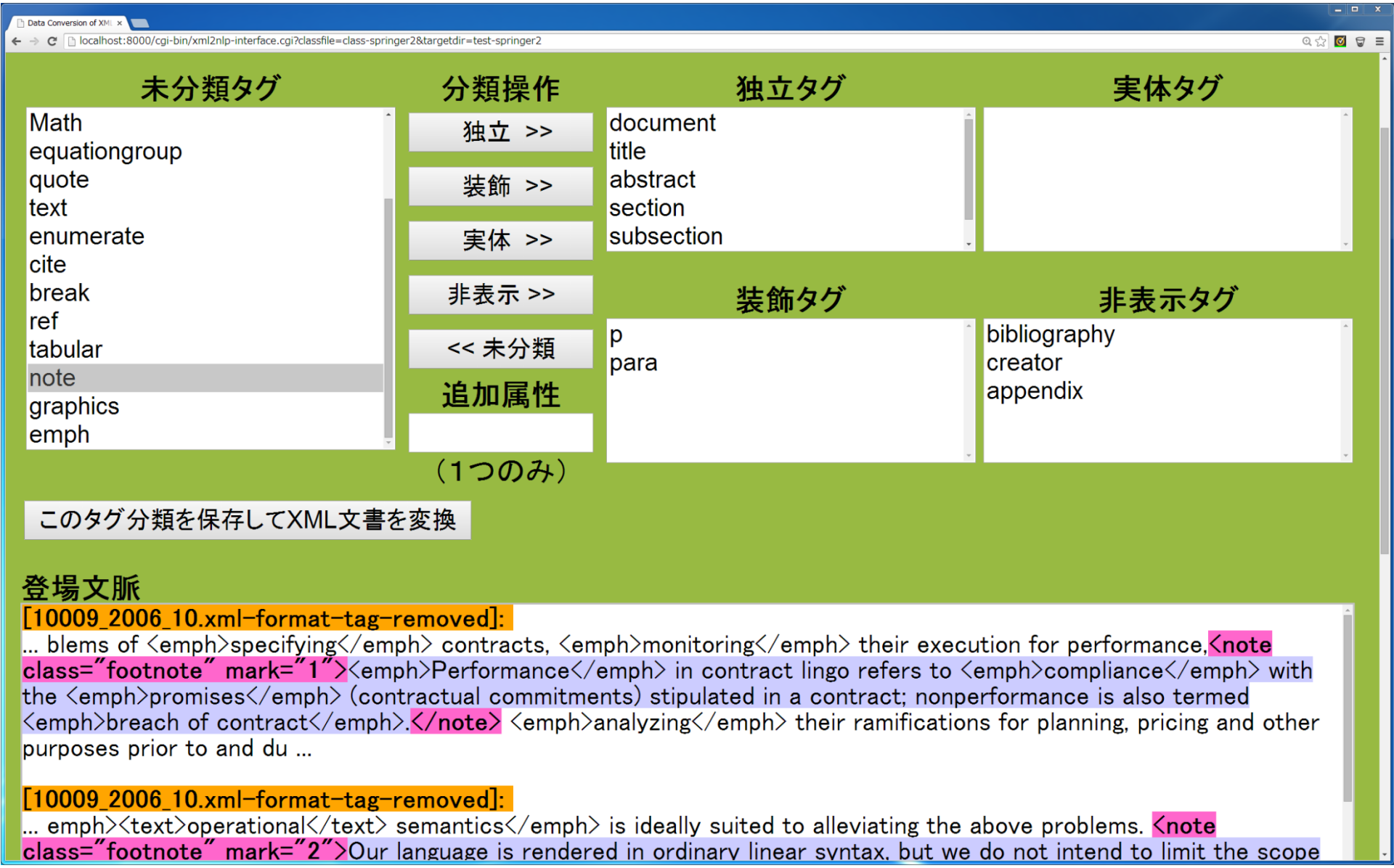
本手続きでもたらされる効率性

- 非表示／実体タグの領域は展開しない
  → 最小限の分類作業に

実体: <Math>



変換ツールの実装


Data Conversion of XML
localhost:8000/cgi-bin/xml2nlp-interface.cgi?classfile=class-springer2&targetdir=test-springer2
未分類タグ
Math
equationgroup
quote
text
enumerate
cite
break
ref
tabular
note
graphics
emph
分類操作
独立 >>
装飾 >>
実体 >>
非表示 >>
<< 未分類
追加属性
（1つのみ）
独立タグ
document
title
abstract
section
subsection
実体タグ
装飾タグ
p
para
非表示タグ
bibliography
creator
appendix
このタグ分類を保存してXML文書を変換
登場文脈
[10009_2006_10.xml-format-tag-removed]:
... blems 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

未分類タグ

Math
cite
note
emph
graphics

変換を実行

分類操作

独立 >>

装飾 >>

実体 >>

非表示 >>

独立タグ

title
section

実体タグ

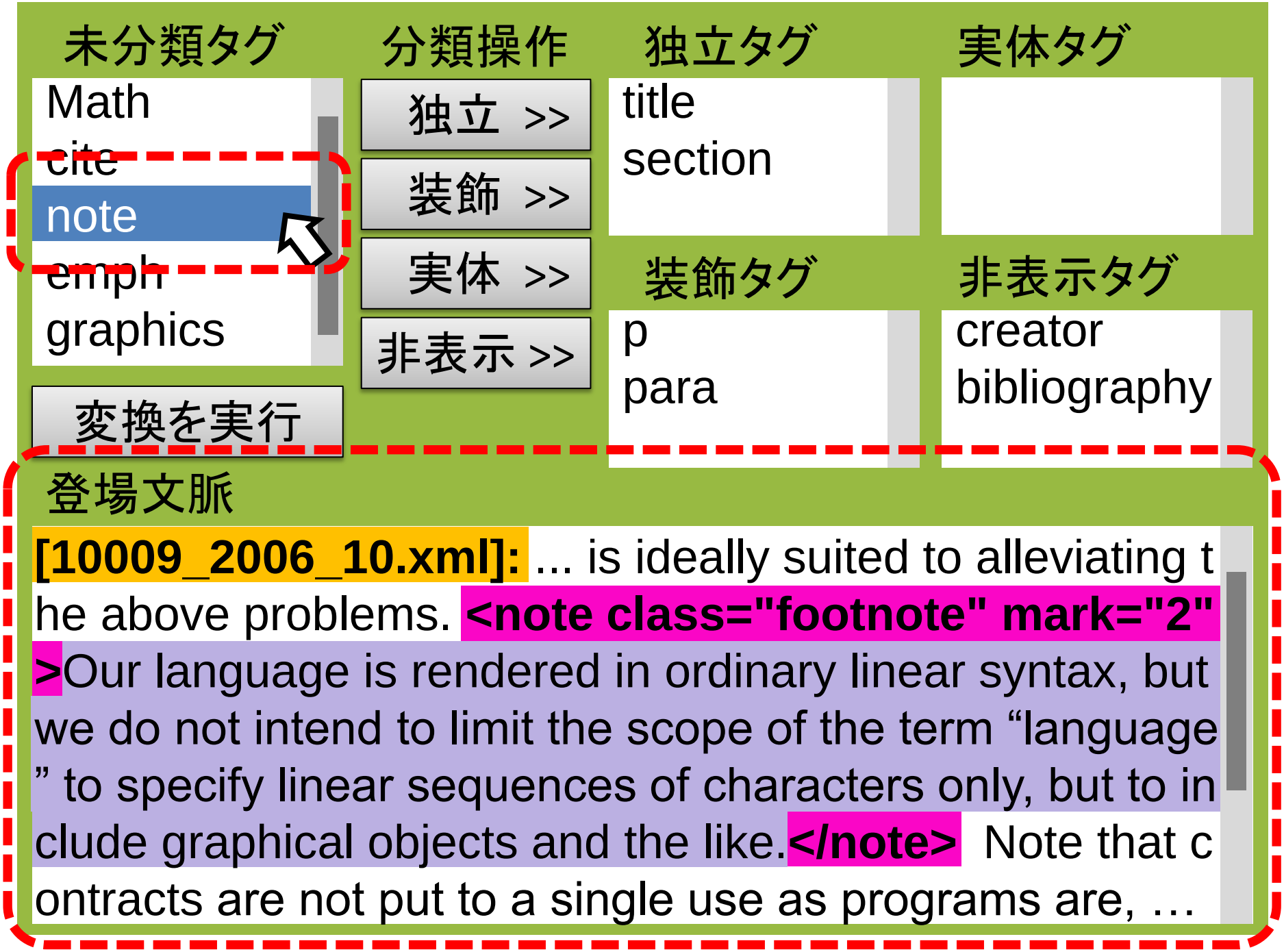
装飾タグ

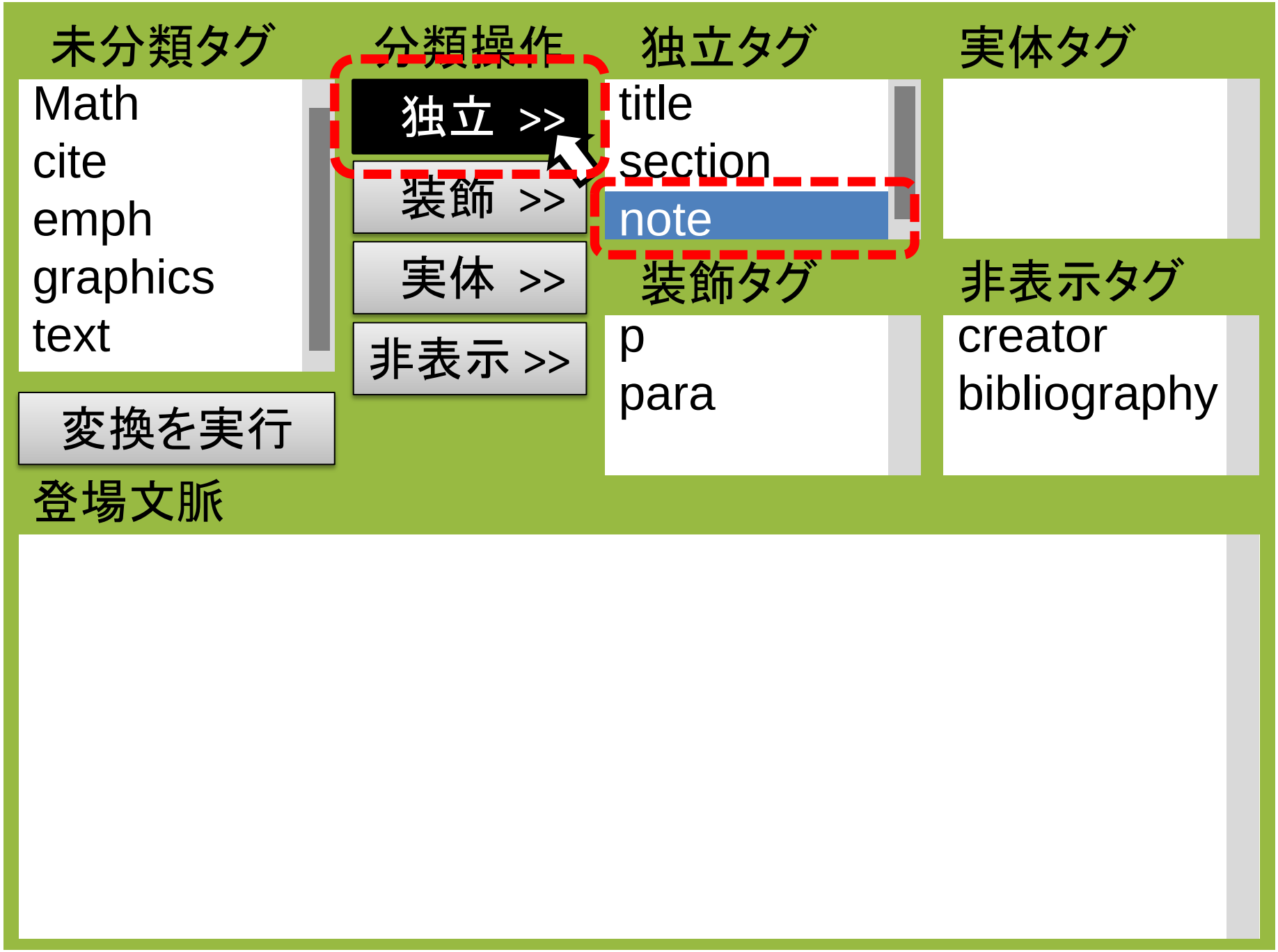
p
para

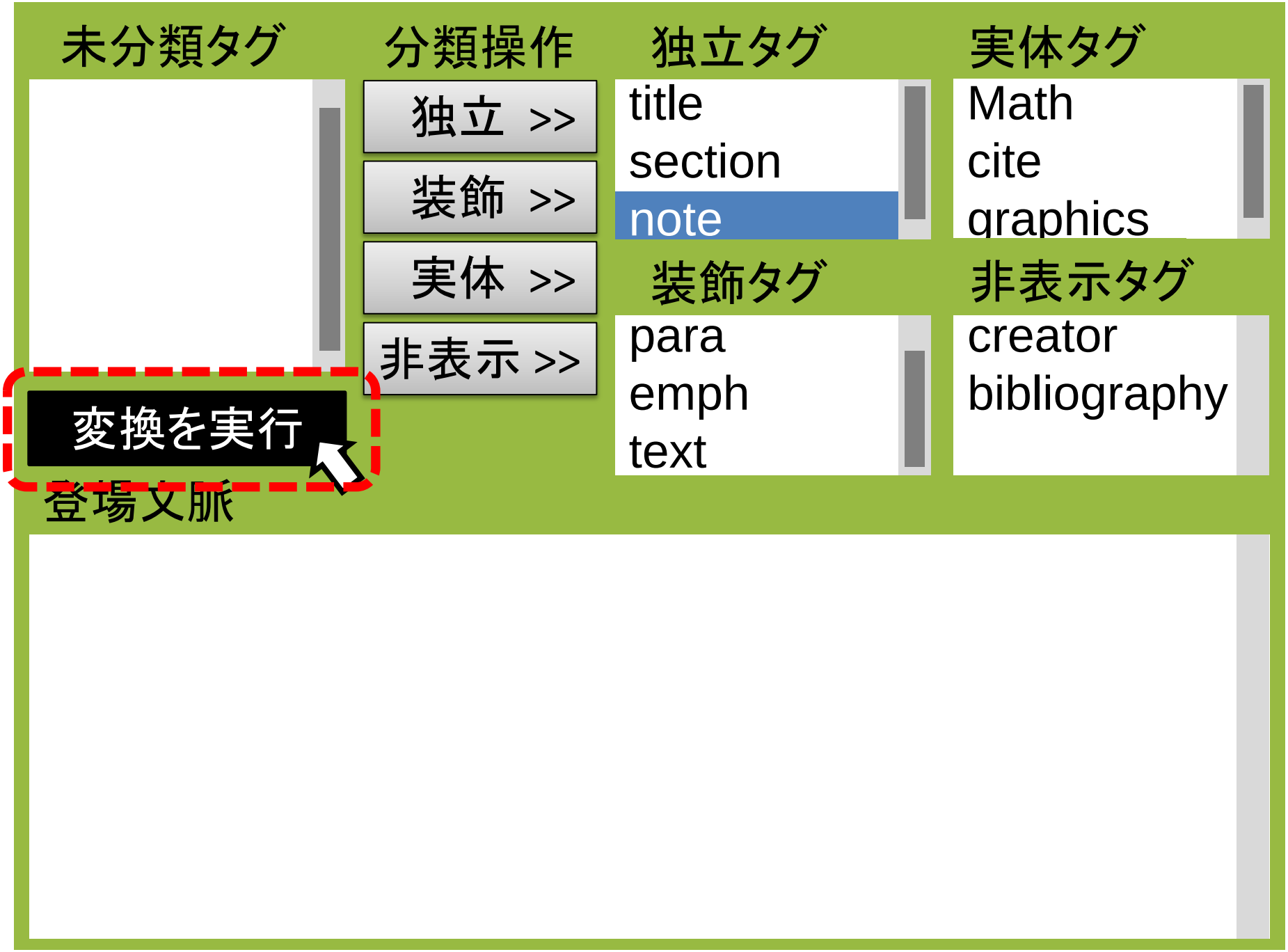
非表示タグ

creator
bibliography

登場文脈


未分類タグ
分類操作
独立タグ
実体タグ
Math
cite
note
emph
graphics
独立 >>
装飾 >>
実体 >>
非表示 >>
title
section
装飾タグ
非表示タグ
p
para
creator
bibliography
変換を実行
登場文脈
[10009_2006_10.xml]: ... 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, …


未分類タグ
Math
cite
emph
graphics
text
分類操作
独立 >>
装飾 >>
実体 >>
非表示 >>
変換を実行
独立タグ
title
section
note
装飾タグ
p
para
実体タグ
非表示タグ
creator
bibliography
登場文脈


未分類タグ
分類操作
独立タグ
実体タグ
独立 >>
装飾 >>
実体 >>
非表示 >>
title
section
note
Math
cite
graphics
装飾タグ
非表示タグ
para
emph
text
creator
bibliography
変換を実行
登場文脈

実験

- 複数の文書群に対して提案枠組を適用
  - 平文テキストを獲得
  - タグ分類の効率性を検証
- 上記平文テキストに NLP ツールを適用
  - 平文テキスト列を適切に得られることがいかに重要か、単純なアプローチと比較検証

実験設定 (1/3): 対象文書群（5種）

文書群	ジャンル	言語	スタイル	使用文書数
PubMed Central (PMC)*1	科学論文	英	XML	1,000
arXiv.org*2	科学論文	英	XHTML	300
Wikipedia*3 エントリ	ウェブ	英	HTML†	300
言語処理学会論文誌「自然言語処理」*4	科学論文	日	XHTML††	384(全)
	科学論文	英	XHTML††	68(全)

(† XML ファイルから生成されたもの)
(†† LaTeXML*5 を用い XML から変換したものが公開中)

---

*1 http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/ *2 http://arxiv.org/
*3 http://www.wikipedia.org/ *4 http://nlp20.nii.ac.jp/resources/
*5 http://dlmf.nist.gov/LaTeXML/

実験設定(2/3): NLPツール

- 英語文書：構文解析器（2種）
  - Enju パーザ*1 (+ GeniaSS*2): 深い統語／意味解析
    - （検索空間のメモリオーバーフロー＝構文解析失敗）
  - Stanford パーザ*3: 句構造／依存関係解析
    - ≤50単語の文を解析（メモリ不足による強制終了対策）
    - （50単語以下 = 成功 / 50単語超 = 自動的に解析失敗）
- 日本語文書：形態素解析器（2種）
  - JUMAN*4／MeCab*5

---

*1 Enju (Ninomiya et al., 2007) *2 http://www.nactem.ac.uk/y-matsu/geniass/
*3 de Marneffe et al., 2006 *4 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
*5 http://mecab.sourceforge.jp

実験設定 (3/3): 比較手法と評価尺度

- 比較手法
  - **単純除去**:　タグを単純除去
  - **実/非のみ**:　実体・非表示タグ → 提案枠組で処理
    装飾・独立タグ　→ 単純除去
  - **提案枠組**:　提案枠組で全タグを処理
- 評価尺度
  - 検出される文の数
  - 解析時間
  - 解析失敗文の数(%)

【訂正】表3
「独/非（独立＋非表示）」は
「実/非（実体＋非表示）」の
誤りです (7/1 正誤表公開)

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	**1,357 k (421)**	32k (12)	62k (9)	48k (9)	34k (56)	**177k (85)**	**26 k**
arXiv (300)	**1,969 k (210)**	6k (15*)	47k (12*)	60k (8*)	8k (17*)	**121k (52*)**	**4 k**
Wiki. (300)	**224 k (60)**	3k (12*)	11 k (8*)	1k (28*)	11k (67*)	**28k (115*)**	**2 k**
JNL-E (68)	**142 k (57)**	8k (25*)	12 k (16*)	6k (9)	23k (19)	**29k (69*)**	**6 k**
JNL-J (384)	**699 k (58)**	50k (23*)	56 k (18*)	32k (10*)	14k (21*)	**153k (72*)**	**38 k**

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	**1,357 k (421)**	32k (12)	32k (9)	18k (9)	34k (56)	**177k (85)**	**26 k**
arXiv (300)	**1,969 k (210)**	6k (15*)	47k (12*)	60k (8*)	8k (17*)	**121k (52*)**	**4 k**
Wiki. (300)	**224 k (60)**	3k (12*)	11 k (8*)	1k (28*)	11k (67*)	**28k (115*)**	**2 k**
JNL-E (68)	**142 k (57)**	8k (25*)	12 k (16*)	6k (9)	23k (19)	**29k (69*)**	**6 k**
JNL-J (384)	**699 k (58)**	50k (23*)	56 k (18*)	32k (10*)	14k (21*)	**153k (72*)**	**38 k**

1/5 以下のみ分類

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	**1,357 k (421)**	32k (13)	32k (9)	18k (9)	34k (56)	**177k (85)**	**26 k**
arXiv (300)	**1,969 k (210)**	6k (15*)	47k (12*)	60k (8*)	8k (17*)	**121k (52*)**	**4 k**
Wiki. (300)	**224 k (60)**	3k (13*)	11 k (8*)	[illegible]	11k (67*)	**28k (115*)**	**2 k**
JI (6						**29k (69*)**	**6 k**
JI (384)	**(58)**	(23*)	(18*)	(10*)	(21*)	**153k (72*)**	**38 k**

1/5 以下のみ分類

* XHTML, HTML:
タグ名のみでは情報が抽象的
→ 属性･属性値とのペアで1つのタグ名

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	1,357 k (421)	32k (12)	32k (9)	13k (9)	34k (56)	177k (85)	26 k
arXiv (300)	1,969 k (210)	9k (15*)	17k (12*)	33k (8*)	9k (17*)	121k (52*)	4 k
Wiki. (300)	224 k (60)	3k	11 k	1k	11k	28k (15*)	2 k
JNL-E (68)	142 k (57)	3k (25*)	12 k (16*)	3k (9)	23k (19)	29k (69*)	6 k
JNL-J (384)	699 k (58)	50k (23*)	50 k (18*)	32k (10*)	14k (21*)	153k (72*)	38 k

1/5 以下のみ分類

出現数の 20% 以下のみに注目

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	1,357 k (421)	32k (12)	32k (9)	13k (9)	34k (56)	177k (85)	26 k
arXiv (300)	1,969 k (210)	9k (15*)	17k (12*)	33k (8*)	9k (17*)	121k (52*)	4 k
Wiki. (300)	224 k (60)	3k	11 k	1k	11k	28k (15*)	2 k
JNL-E (68)	142 k (57)	3k (25*)	12 k (16*)	3k (9)	23k (19)	29k (69*)	6 k
JNL-J (384)	(38)	(25*)	(18*)	(10*)	(21*)	3k (72*)	38 k

1/5 以下のみ分類

出現数の 20% 以下のみに注目

実体・非表示タグで囲まれた領域の内部がそれ以上展開されなかったため

各種文書群に対して分類されたタグと得られた平文テキスト列

文書 (数)	全タグ数 (異なり数)	分類されたタグの数 (異なり数)					獲得列数
		独立	装飾	実体	非表示	合計	
PMC (1,000)	**1,357 k (421)**	32k (12)	62k (9)	48k (9)	34k (56)	**177k (85)**	**26 k**
arXiv	**1,969 k**	6k	47k	60k	8k	**121k**	**4 k**
							2 k
							6 k
JNL-J (384)	**699 k (58)**	50k (23*)	56 k (18*)	32k (10*)	14k (21*)	**153k (72*)**	**38 k**

文書をランダムに選択し観察 →
- 平文テキストがNLPツールに直接入力可能
- 文書全体を被覆可能な「適切な」文から構成されていることを確認

タグの扱い方がEnjuパーザの性能に与える影響

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
PMC (1,000)	単純除去	159,327	209,783	4,721 (2.96)
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
JNL-E (68)	単純除去	23,196	24,881	271 (1.17)
	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

タグの扱い方がStandordパーザの性能に与える影響

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
PMC (1,000)	単純除去	170,999	58,865	18,621(10.89)
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

タグの扱い方が Enju パーザ の性能に与える影響

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
PMC	単純除去	159,327	209,783	4,721 (2.96)
			135,752	810 (0.72)
			132,250	699 (0.55)
			108,831	2,047 (2.74)
(300)	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki.	単純除去	10,561	14,704	1,161(10.99)
(300)	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
JNL-E	単純除去	23,196	24,881	271 (1.17)
(68)	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

提案枠組: 1%未満
解析失敗 1～10% 減
＝ 文書被覆率の大幅改善

タグの扱い方が Enju パーザ の性能に与える影響

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
PMC	単純除去	159,327	209,783	4,721 (2.96)
			135,752	810 (0.72)
			132,250	699 (0.55)
			108,831	2,047 (2.74)
(300)	実/非のみ	41,265	89,200	411 (1.00)
			87,952	348 (0.81)
			14,704	1,161(10.99)
			6,743	67 (1.33)
			6,058	61 (0.88)
JNL-E (68)	単純除去	23,196	24,881	271 (1.17)
	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

提案枠組: 1%未満
解析失敗 1～10% 減
＝ 文書被覆率の大幅改善

提案枠組:
解析時間 25～59% 減
＝ 大幅短縮を達成

タグの扱い方が Standord パーザ の性能に与える影響

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
PMC	単純除去	170,999	58,865	18,621(10.89)
			50,741	11,881 (9.42)
			63.295	11.338 (8.11)
			27,970	10,590(13.99)
(300)	実/非のみ	48,666	24,630	5,457(11.21)
			26.360	5.345(10.58)
			3,114	1,651(11.09)
			2,248	282 (4.57)
			2.451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10.154	892 (4.48)

提案枠組:
解析失敗 2～8% 減
＝ 文書被覆率の大幅改善

提案枠組 or 実/非のみ:
解析時間 12～28% 減
＝ 大幅短縮を達成

タグを単純に除去すると、構文解析器は以下により混乱

- 文の埋め込み挿入（独立タグ）
- 自然言語ではない要素で構成された表現（実体タグ）
- 文の表示に直接関係しないシーケンス（非表示タグ）

提案枠組:
解析失敗 2～8% 減
＝ 文書被覆率の大幅改善

提案枠組 or 実/非のみ:
解析時間 12～28% 減
＝ 大幅短縮を達成

PMC	単純除去	170,999		58,865	18,621(10.89)
				50,741	11,881 (9.42)
				63.295	11.338 (8.11)
				27,970	10,590(13.99)
(300)	実/非のみ	48,666		24,630	5,457(11.21)
				26.360	5.345(10.58)
				3,114	1,651(11.09)
				2,248	282 (4.57)
				2.451	258 (3.21)
JNL-E	単純除去	24,942		9,069	1,577 (6.32)
(68)	実/非のみ	17,572		7,865	1,058 (6.02)
	提案枠組	19,925		10.154	892 (4.48)

タグを単純に除去すると、構文解析器は以下により混乱

- 文の埋め込み挿入（独立タグ）
- **自然言語ではない要素で構成された表現（実体タグ）**
- **文の表示に直接関係しないシーケンス（非表示タグ）**

PMC (1,000)	単純除去	170,999	58,865	18,621(10.89)
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

タグを単純に除去すると、構文解

非自然言語箇所の排除

- 文の埋め込み挿入（独立タグ）
- **自然言語ではない要素で構成された表現（実体タグ）**
- **文の表示に直接関係しないシーケンス（非表示タグ）**

PMC (1,000)	単純除去	170,999	58,865	18,621(10.89)
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

タグの扱い方が Enju パーザ の性能に与える影響

解析時間 1～12%減

解析失敗文 0.2～0.9%減

文書		数		
PMC (1,000)	単純除去			
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
JNL-E (68)	単純除去	23,196	24,881	271 (1.17)
	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

タグの扱い方が Enju パーザ の性能に与える影響

解析時間 1～12%減

解析失敗文 0.2～0.9%減

検出文数
5 ～37% 増

文				
PMC (1,000)	単純除去			
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
	除去	23,196	24,881	271 (1.17)
	のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

独立タグ：文区切りで検出できない文境界を示し得る
→ 文の数が増える＝**文長の減少**

解析時間 1～12%減

解析失敗文 0.2～0.9%減

検出文数
5 ～37% 増

文				
PMC (1,000)	単純除去			
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
	除去	23,196	24,881	271 (1.17)
	のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

独立タグ：文区切りで検出できない文境界を示し得る
→ 文の数が増える＝**文長の減少**

検索空間不足を抑制

解析時間 1～12%減

解析失敗文 0.2～0.9%減

検出文数
5 ～37% 増

PMC (1,000)	単純除去			
	実/非のみ	112,285	135,752	810 (0.72)
	提案枠組	126,215	132,250	699 (0.55)
arXiv (300)	単純除去	74,762	108,831	2,047 (2.74)
	実/非のみ	41,265	89,200	411 (1.00)
	提案枠組	43,208	87,952	348 (0.81)
Wiki. (300)	単純除去	10,561	14,704	1,161(10.99)
	実/非のみ	5,026	6,743	67 (1.33)
	提案枠組	6,893	6,058	61 (0.88)
	単純除去	23,196	24,881	271 (1.17)
	実/非のみ	15,606	21,304	183 (1.17)
	提案枠組	17,929	18,683	50 (0.28)

タグの扱い方が Standord パーザ の性能に与える影響

解析時間 7～29%増

解析失敗文 0.6～1.5%減

文		数		
PMC (1,000)	単純除去	[illegible]	[illegible]	[illegible]
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
JNL-E (68)	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

独立タグ：文区切りで検出できない文境界を示し得る
→ 文の数が増える＝**文長の減少**

解析時間 7～29%増

解析失敗文 0.6～1.5%減

検出文数
4 ～30% 増

PMC (1,000)	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
	除去	24,942	9,069	1,577 (6.32)
	のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

独立タグ：文区切りで検出できない文境界を示し得る
→ 文の数が増える＝**文長の減少**

50単語以下の（解析対象となる）文の増加

解析時間 7～29%増

解析失敗文 0.6～1.5%減

検出文数
4 ～30% 増

PMC (1,000)	単純除去			
	実/非のみ	126,176	50,741	11,881 (9.42)
	提案枠組	139,805	63,295	11,338 (8.11)
arXiv (300)	単純除去	75,672	27,970	10,590(13.99)
	実/非のみ	48,666	24,630	5,457(11.21)
	提案枠組	50,504	26,360	5,345(10.58)
Wiki. (300)	単純除去	14,883	3,114	1,651(11.09)
	実/非のみ	6,173	2,248	282 (4.57)
	提案枠組	8,049	2,451	258 (3.21)
	単純除去	24,942	9,069	1,577 (6.32)
	実/非のみ	17,572	7,865	1,058 (6.02)
	提案枠組	19,925	10,154	892 (4.48)

タグの扱い方がJUMAN / MeCabの性能に与える影響

JUMAN

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
JNL-J (384)	単純除去	96,668	122	10 (0.01)
	実/非のみ	76,277	86	8 (0.01)
	提案枠組	114,250	59	2 (0.00)

MeCab

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
JNL-J (384)	単純除去	97,312	7	10 (0.01)
	実/非のみ	78,461	6	8 (0.01)
	提案枠組	116,424	6	2 (0.00)

タグの扱い方が JUMAN / MeCab の性能に与える影響

JUMAN

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
JNL-J (384)	単純除去	96,668	122	10 (0.01)
	実/非のみ	76,277	86	8 (0.01)
	提案枠組	114,290	59	2 (0.00)

提案枠組: 解析時間51%減
＝ 大幅短縮を達成

提案枠組: 解析失敗 0.0%
＝ 文書被覆率ほぼ 100%

文書(数)	手法	検出文数	解析時間(秒)	失敗文数(%)
JNL-J (384)	単純除去	97,312	7	10 (0.01)
	実/非のみ	78,461	6	8 (0.01)
	提案枠組	116,424	6	2 (0.00)

対象文書全体の高被覆かつ効率的な処理が意味するもの

- 浅い分析（単語数のカウント等）で良いタスク
  → タグの単純除去でも事足りる
  - 文の挿入は単語数に影響せず
  - 非自然言語要素は大量のテキストで打ち消せる
- 詳細かつ正確な分析要求（談話分析・翻訳・文法抽出など）にはより厳密な解析が必要
  - ごく狭い箇所で発せられた情報も慎重に吟味
  - その分析は、本文とは関係のない要素の混入を確実に排除したテキストに対して行う必要がある

＝ 近年 NLP が確立してきたものそのもの

対象文書全体の高被覆かつ効率的

自然言語現象の正確な解析に集中するために、
データセットは平文テキスト列の形に整備
→新たなタスクの発見・取り組み・成果報告

- 非自然言語要素は大量のテキストで打ち消せる

- 詳細かつ正確な分析要求（談話分析・翻訳・文法抽出など）にはより厳密な解析が必要
  - ごく狭い箇所で発せられた情報も慎重に吟味
  - その分析は、本文とは関係のない要素の混入を確実に排除したテキストに対して行う必要がある

＝ 近年 NLP が確立してきたものそのもの

実文書を自然言語処理技術と適切に繋ぐ技術の重要性

- 多くの有用な成果は、本来目指す最終目的「実世界の文書の解析」へ還元させてこそ
- 本研究の実験結果が示唆するもの：
  「適切な枠組で支援すれば、従来のNLPツールには、実文書テキストを『本来期待される性能で』解析する能力が既に備わっている」
- 「実文書とNLP技術を適切に繋ぐ」：
  NLPをあらゆる局面で利用する際、その本来の効力をフル活用するための重要なタスク

本研究のまとめ

- XMLタグ付テキストとNLPツール入出力との間の変換枠組を提案
  - タグを4種のテキスト構成機能に分類

  →分類に基づき平文テキスト列へ自動的に変換
- 対象文書中の 20%以下のタグを分類することで平文テキスト列を獲得成功
- タグの単純除去に比べ、NLPツールの大幅な解析性能改善（高被覆・効率化）を確認

→実文書をNLPと適切に繋ぐ技術の重要性示唆

今後の展望

- 提案枠組ツールの公開（近日）
  - 多種多様な実文書へのNLPツール適用の議論の提供・共有
- より柔軟・繊細な構造解析の検討
  - （例）実体タグ内と周囲のテキストの繋ぎ方
- より多様な文書形式への適用可能性の検討
  - OCRで得られた文書・スライドデータ等との接続

ありがとうございました