

科学技術論文の構造解析 に基づく知識獲得

進藤 裕之
奈良先端科学技術大学院大学
2017/11/21
@富士通研究所

進藤 裕之

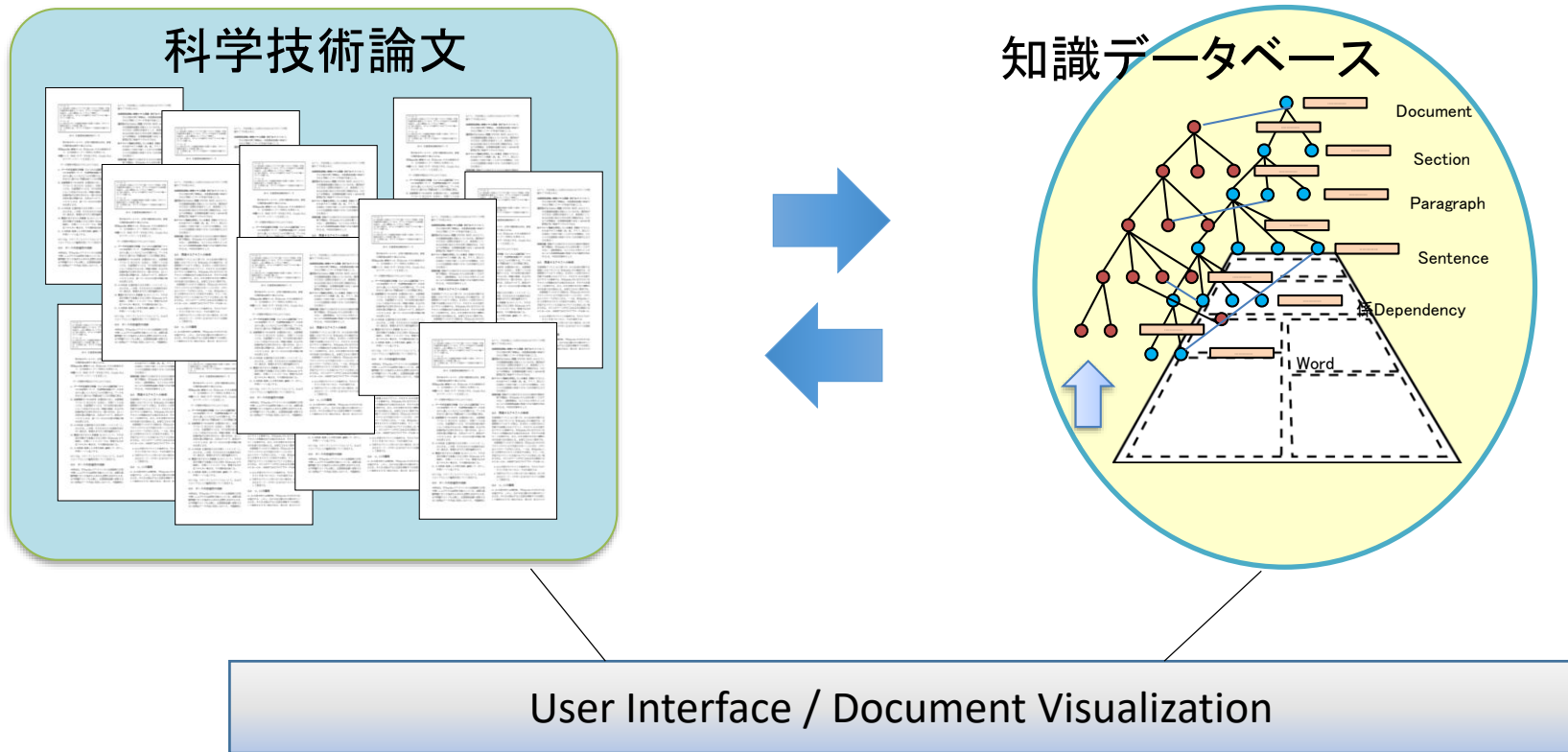
- 2009～2014 NTTコミュニケーション科学基礎研究所
- 2013 奈良先端大 博士後期課程修了
- 2014～ 奈良先端大 助教

研究テーマ:

- 構文解析, 意味解析
- 質問応答(QA)システム
- 論文解析

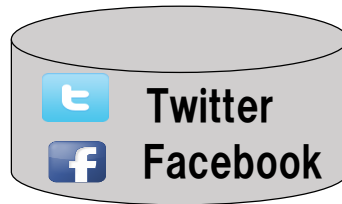
論文CREST

膨大な科学技術論文からの知識獲得・編集・検索



社会脳CREST

脳活動と言語情報から個人の社会的態度を予測する

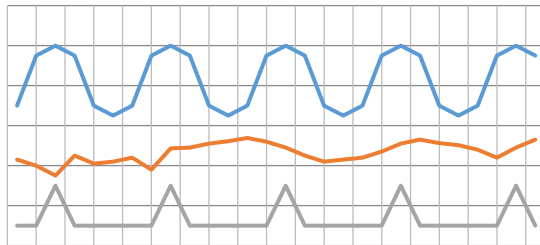


impulsivity
(衝動性)

Empathy
(共感性)

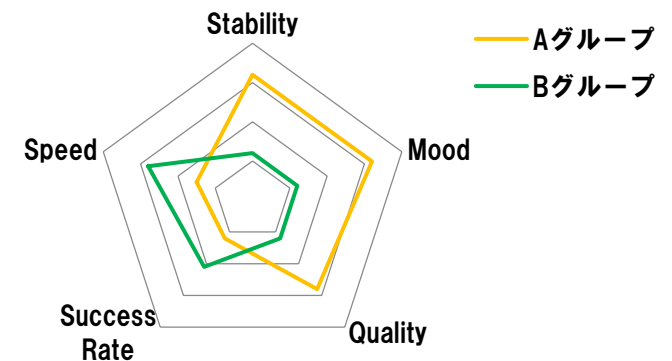
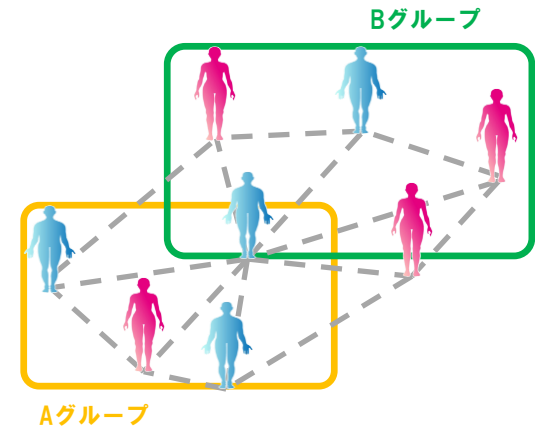
social type

fMRI



stress sensor

予測



Quiz Bowl

早押しクイズ(知識データベースから答えを探す)

問題例:

Later in its existence, this polity's leader was chosen by a group that included three bishops and six laymen, up from the seven who traditionally made the decision. Free imperial cities in this polity included Basel and Speyer. Dissolved in 1806, its key events included the Investiture Controversy and the Golden Bull of 1356. Led by Charles V, Frederick Barbarossa, and Otto I, for 10 points, name this polity, which ruled most of what is now Germany through the Middle Ages and rarely ruled its titular city.



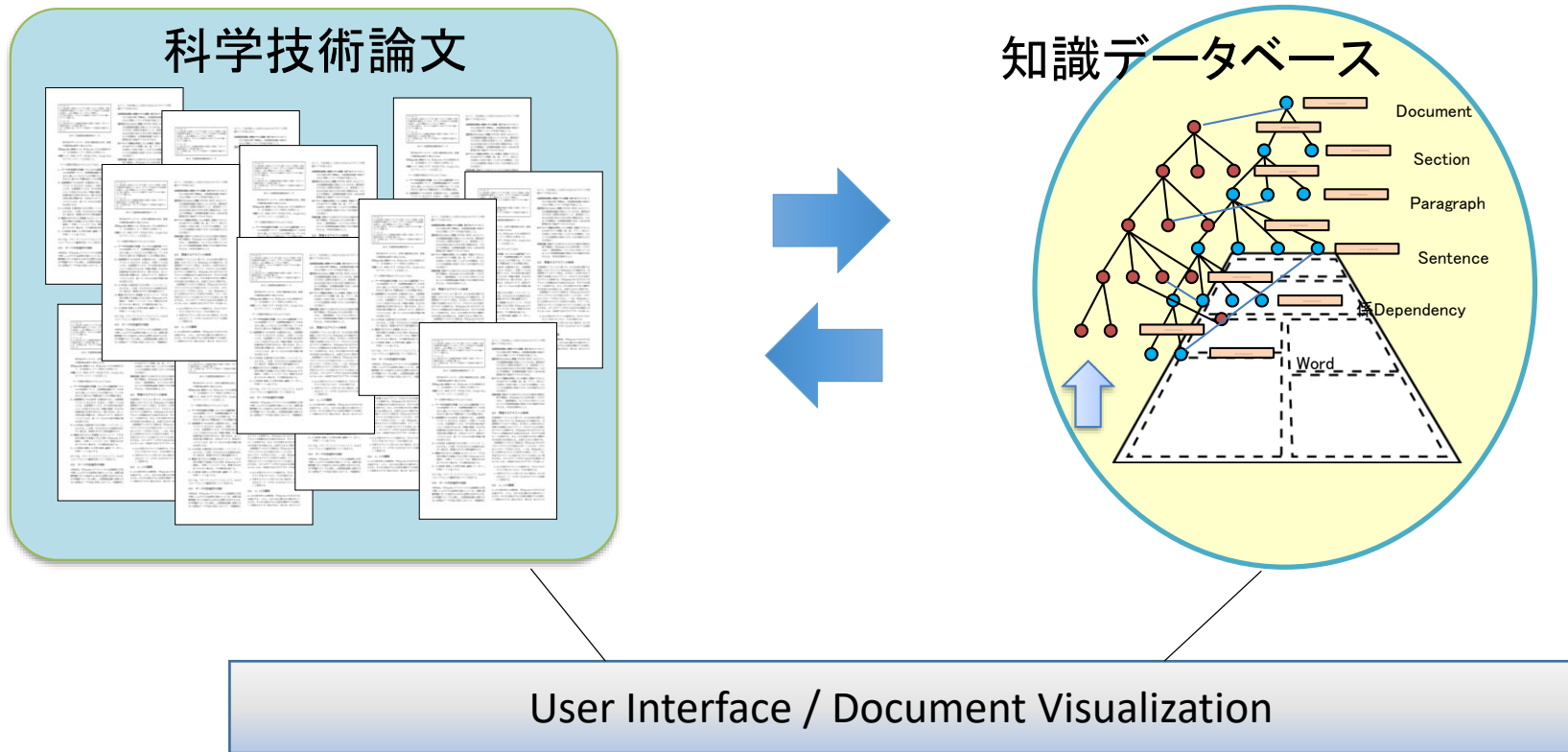
知識DBの高速な検索, フィルタリング, Deep Learning



NIPS2017 コンペティションで優勝

論文CREST

膨大な科学技術論文からの知識獲得・編集・検索



科学技術論文の解析

1992年から2014年までのarXiv投稿数



科学技術論文の解析

毎日20本の論文をチェックしないと追いつかない！



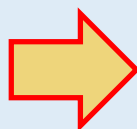
- arXiv (CL, ML): 1日20本程度
- BioRxiv: 1日20本程度

論文からの知識獲得

【従来】 人手による知識データベース構築



読解



知識抽出



データベース

Statistics of genus: KMAPsAick_v1.200.00					
SuperKingdom	Kingdom	Order	Family	# of genes	# of Metabolites
Archaea	****	Methanobacteriales	Methanobacteriaceae	1	7
	****	Methanococcales	Methanococcaceae	1	6
Bacteria	****	Actinomycetales	Micrococaceae	1	3
	****	Actinomycetales	Streptomyetaceae	1	69
	****	Bacillales	Alcalyflabaceae	1	10
	****	Bacillales	Bacillaceae	1	5
	****	Bacillales	Thapsylobacillaceae	1	1
	****	Bifidobacteriales	Caulobacteraceae	1	1
	****	Enterobacteriales	Enterobacteriaceae	3	19
	****	Flavobacteriales	Flavobacteriaceae	1	2
	****	Lactobacillales	Lactobacillaceae	1	1
	****	Mycrococcales	Cystobacteraceae	1	1
	****	Mycrococcales	Mycrococcaceae	1	13

データベースの例:

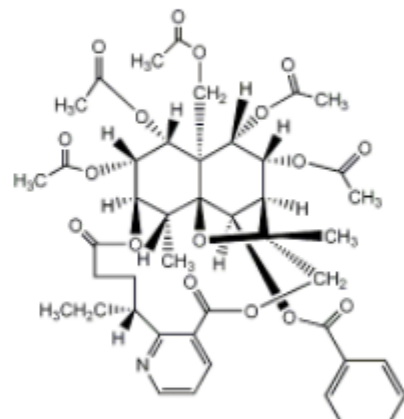

- KNApSAcK (バイオ)
- KEGG (バイオ)
- PolyInfo (材料科学)

1. KNApSAcK (バイオ)

植物⇔代謝物の関係データベース

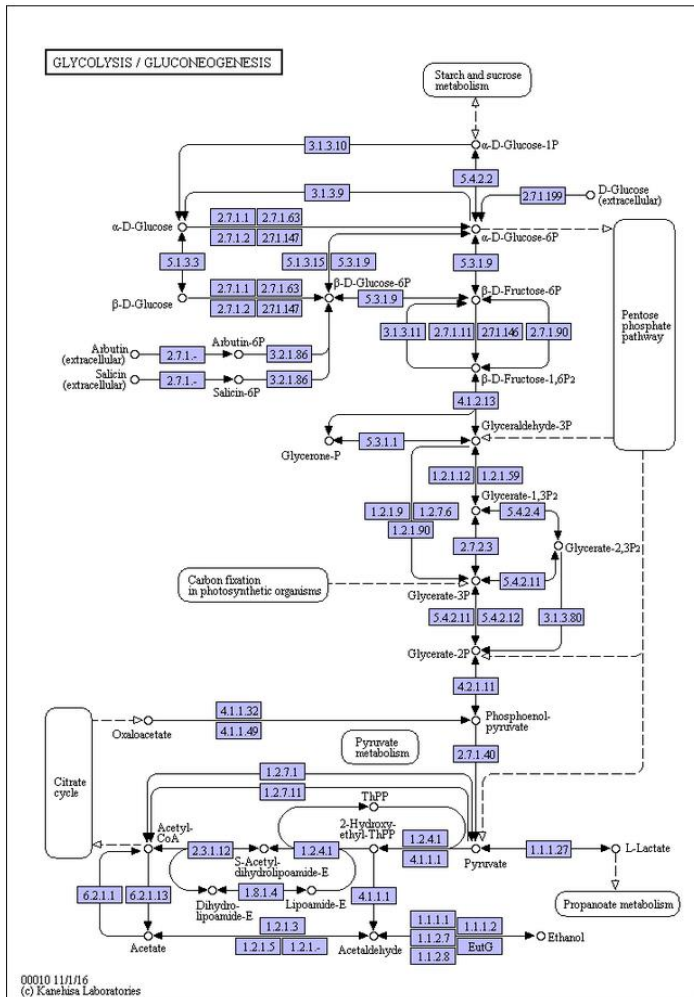


input word = C00001966

Metabolite Information					Structural formula
Name	Cassinine				 zoom in
Formula	C44H51NO17				
Mw	865.31569921				
CAS RN	62948-58-7				
C_ID	C00001966 				
InChIKey	DSIMUNJDGBYLQE-GUDTZNBYNA-N				
Organism	Kingdom	Family	Species	Reference	
	Plantae	Celastraceae	Cassine matabelica	Ref.	
	Plantae	Celastraceae	Cassine metabelica	Ref.	

2. KEGG (バイオ)

代謝などのパスウェイに関するデータベース




■ : 遺伝子産物

○ : 化合物

- Relation (protein ネットワーク)
- Reaction (chemical ネットワーク)

3. PolyInfo (材料科学)

高分子の物性, 構造, 化学式に関するデータベース



[Polymer Search: Basic / Advanced / Text](#)
[Polymer Structure Search: by Elements / by Modeling](#)
[Easy Browse: Property table / Popular polymer / Plotted data](#)

[Monomer Search: Easy / Basic](#)
[Property Prediction: Group contribution](#)
[Nomenclature: IUPAC structure based name](#)
[NMR: NMR Database](#)

HELP: [Japanese](#) [English](#)

Polymer Structure Search [by Elements]

[Query by Template](#) [Query by Statement](#)

Template

Polymer Type:	not specified
Material Type:	not specified
Search option:	<input type="radio"/> Search polymers consisting of selected elements only. <input checked="" type="radio"/> Search polymers containing selected elements.
Element selection:	Set the numbers of elements included in the target polymer into the below blanks.

ex.1)	Specification of number	
	1	1-3
	>3	0
	even	odd
ex.2)	2;m	Containing two specified basic elements in the main chain.
ex.3)	>=3;s	Containing more than three specified basic elements in the side chain.

m=main chain, s=side chain [\[details\]](#)

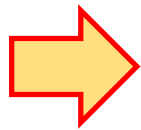
Chain Elements , [3-8,4-Membered Rings](#) , [5-Membered Rings](#) , [6-Membered Rings](#) , [5-8,5-\(Condensed, Spiro\)Membered Rings](#) , [5-8,6-\(Condensed, Spiro\)Membered Rings](#) , [6-8,6-\(Condensed, Spiro\)Membered Rings](#) , [6-8,6-8,6-Membered Rings](#) , [Others](#)

E101	E102	E103	E104	E130	E187	E129	E131
—H	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{—}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{—}$	=C=	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{CH}}}\text{—}$	—CH=	$\text{—CH}_2\text{—}$	—CH_3
E205	E297	E298	E202	E203	E204	E206	E201
$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{—}$ CH_2	—CH=CH—	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{=CH—}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{=C—}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{=C—}$ <i>cis</i>	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{=C—}$ <i>trans</i>	—CH=CH_2	$\text{—C}\equiv\text{C—}$
E191	E192	E188	E207	E208	E209	E211	E106
—CF_3	$\text{—CF}_2\text{—}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{CF}}}\text{—}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{—}$	—CHO	$\text{—C}\equiv\text{N}$	$\text{—}\overset{\text{I}}{\underset{\text{I}}{\text{C}}}\text{—}$	—O—

論文からの知識獲得

クライアントの持っているデータ

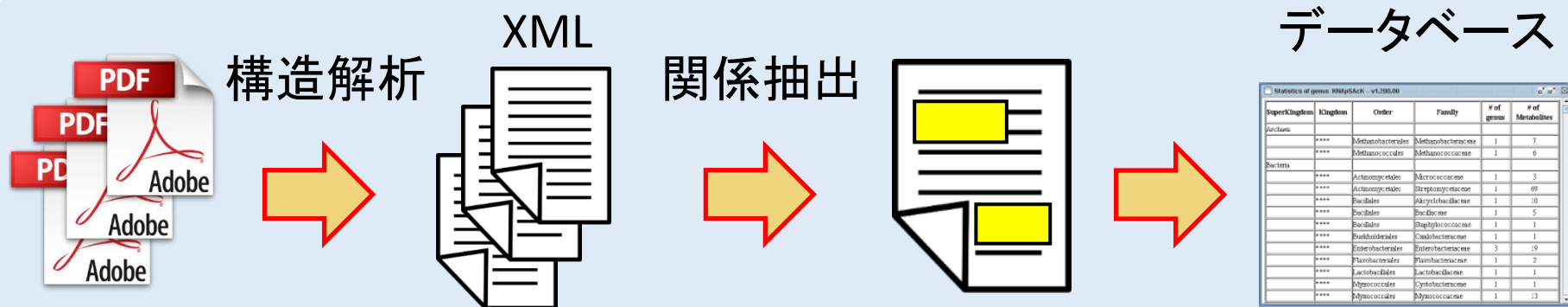
- 大量の論文(PDF)
- 論文から抽出したデータベース
- どの論文からどの関係を抽出したかという参照情報



既存のデータベースは弱い教師データとして使える

論文からの知識獲得

計算機による自動知識データベース構築



- どうやってPDFをXML化するか？
- XMLのタグ仕様は？
- 学習データをどうやって作るか？


etc.

PDFの中身を抽出する

- PDFを画像化してOCR
 - PDFを直接読み取ってテキスト化(XML, HTML)する
 - Popplerなどのツール
- 汚いテキストファイルが出来上がる
数式や表がグチャグチャで、本文と混ざっている, 等

PDFの中身を抽出する

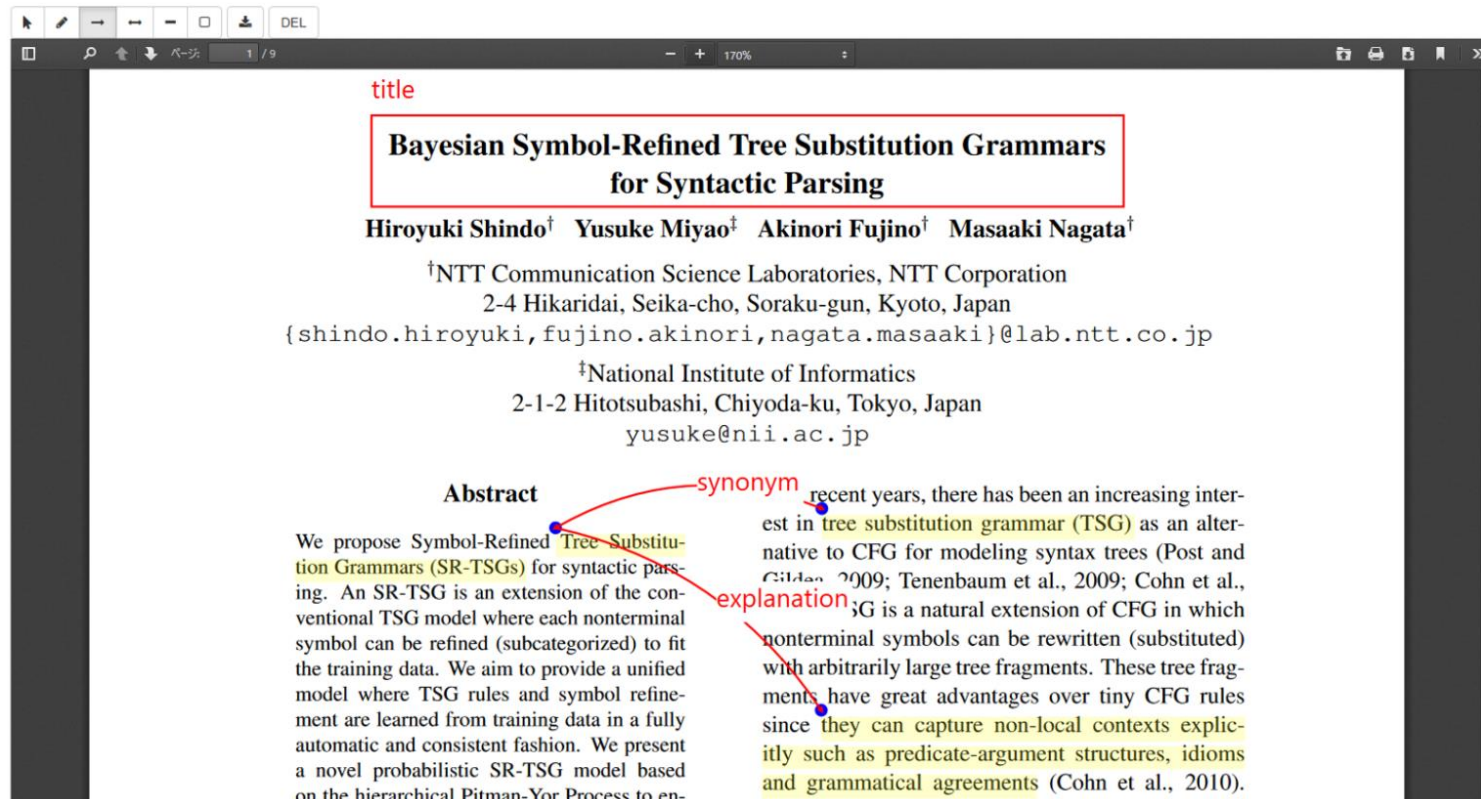
ACL Anthology Reference Corpus (2008)

- ACL系論文のPDFファイルをNuance Omnipage(OCRソフト)を使ってXML化したコーパス
 - 追加でデータのcleaningやエラー除去を行っているが、不十分
- コーパスにアノテーションを行い、機械学習の訓練データにする場合、コーパスが更新されても、アノテーションを移行できない 

PDFの中身を抽出する

PDFアノテーションツール (PDFAnno) の開発

PDFへ直接アノテーションを行ったほうが良い.



PDFの中身はどうなっているか

$$P^{\text{ru-cfg}}(\alpha | x) = \frac{1}{|x \Rightarrow \cdot|}$$

where $|x \Rightarrow \cdot|$ is the number of RU-CFG rules rooted with x . Overall, our hierarchical model encodes backoff smoothing consistently from the SR-TSG rules to the SR-CFG rules, and from the SR-CFG rules to the RU-CFG rules. As shown in (Blunsom and Cohn, 2010; Cohen et al., 2010), the parsing accuracy of the TSG model is strongly affected by its backoff model. The effects of our hierarchical backoff model on parsing performance are evaluated in Section 5.

4 Inference

We use Markov Chain Monte Carlo (MCMC) sampling to infer the SR-TSG derivations from parse trees. MCMC sampling is a widely used approach

抽出できる情報

- 文字と座標, 領域
- 直線・曲線の座標
- 画像の座標

PDFの中身はどうなっているか

- 単語や文, パラグラフの区切りはPDFに含まれていない.
スペース文字も含まれていない.
- 図や表などのフロート要素は, 本文に割り込んだ順番でPDFへ書き込まれる
- 図は, ベクタ形式の場合と画像形式の2通りある.
 - ベクタ形式: PDFでは線とテキストで図が表現されている.
 - 画像形式: 画像としてPDFへ埋め込まれる.

PDFの構造解析

文字系列から木構造への変換問題



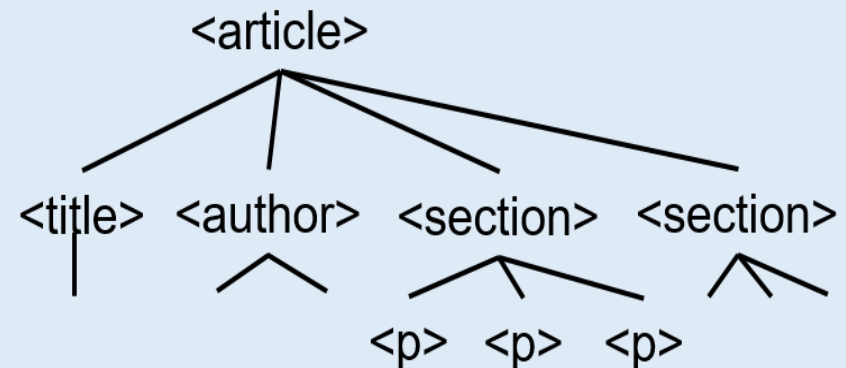
文字/図形描画命令	f_1	f_2	f_3	f_4	f_5	f_6
s	0.0	20.4	4.4	6.0	12.0	6.0
i	4.4	20.4	3.3	6.0	12.0	6.0
n	7.7	20.4	6.5	6.0	12.0	6.0
[MOVE_TO]	17.4	17.4				
[LINE_TO]	23.2	17.4				
θ	17.4	12.3	5.5	6.0	12.0	6.0
[MOVE_TO]	20.0	2.5				
[LINE_TO]	21.4	1.2				
[LINE_TO]	22.8	2.5				
2	17.4	28.6	5.9	6.0	12.0	6.0

文字(+座標)の系列

XML



解析



木構造

PDFの構造解析

どのようなXMLを出力すれば良いか？

- JATS (Journal Article Tag Suite): アメリカ国立医学図書館が開発した科学技術論文のためのXML
- セクション, 段落, 数式, 表, 図, 参考文献などのタグが定義されている
- ただし, 雑誌や分野特有のタグがあり, 曖昧性が高いため, これを簡略化したタグセットを用いる

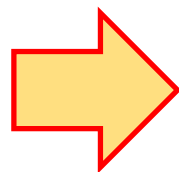


Ex.

`<abbrev>DASH</abbrev>`

PDFの構造解析

分野に依存しない JATS compatible な XML タグセット



機械学習

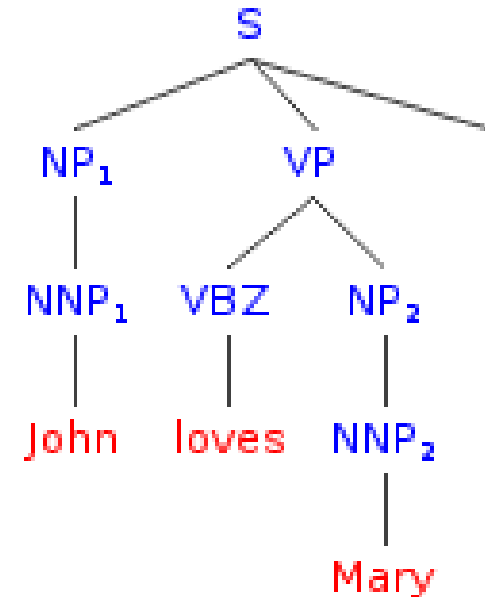
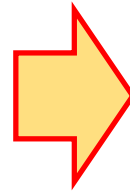
LatexをXMLへ変換することは
現時点でも限定的には可能

- `<head>`
 - `<title>`
 - `<author>`
 - `<abstract>`
- `<body>`
 - `<sec>`
 - `<p>`
 - `<disp-formula>`
- `<back>`
 - `<ref>`
- `<floats-group>`
 - `<fig>`
 - `<table>`

自然言語処理における構造解析

構文解析

John loves Mary .



構文解析では, 1文を入力 → 木構造を出力する問題

自然言語処理における構造解析

構文解析のアルゴリズムとモデル

アルゴリズム

- Shift-reduceアルゴリズム
- CKYアルゴリズム(動的計画法の一種)
- Easy-firstアルゴリズム

スコアモデル

- 線形モデル
- SVM
- ニューラルネットワーク

例

(初期状態)

スタック

$$\sin \frac{\hat{\theta}}{2}$$

バッファ

s i n [-] θ [^] 2

例

SHIFT

スタック

S

バッファ

i n [-] θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

例

SHIFT

スタック

s i

バッファ

n [-] θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

例

SHIFT

スタック

s i n

バッファ

[-] θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

例

REDUCE

スタック

s <mi*>

i n

バッファ

[-] θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

REDUCE操作で木の一部を作る

例

REDUCE

スタック

<mi>

s <mi*>

i n

バッファ

[-] θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

REDUCE操作で木の一部を作る

例

SHIFT-REMOVE

スタック

<mi>

s <mi*>

i n

バッファ

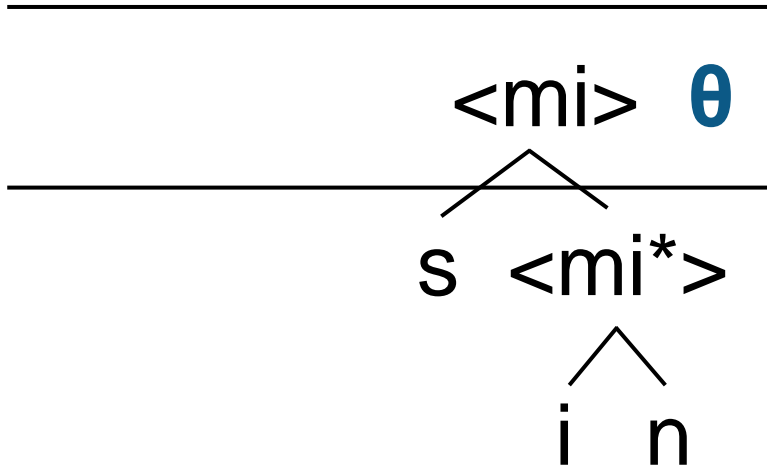
 θ [^] 2

$$\sin \frac{\hat{\theta}}{2}$$

例

SHIFT-COPY

スタック



$$\sin \frac{\hat{\theta}}{2}$$

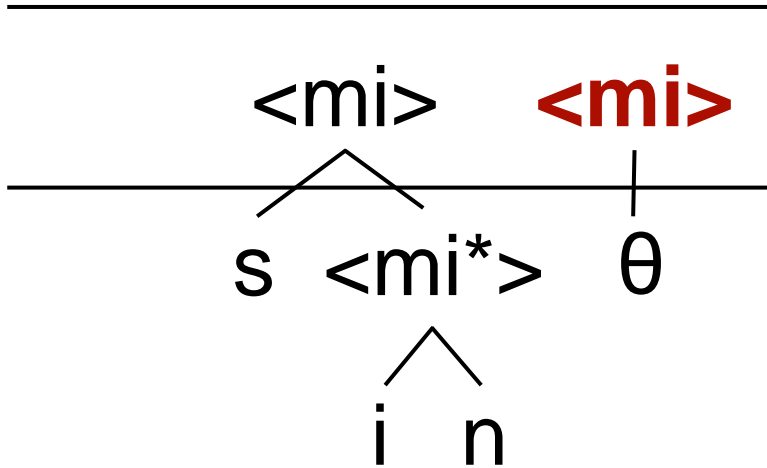
バッファ



例

REDUCE

スタック



$$\sin \frac{\hat{\theta}}{2}$$

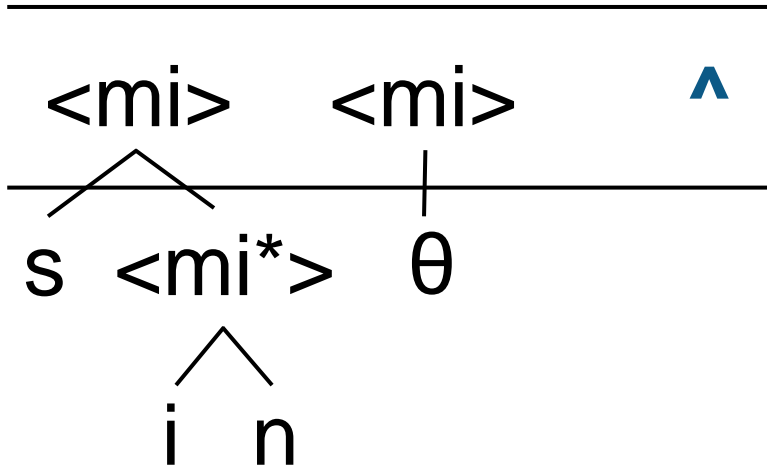
バッファ

[^] 2

例

SHIFT

スタック



$$\sin \frac{\hat{\theta}}{2}$$

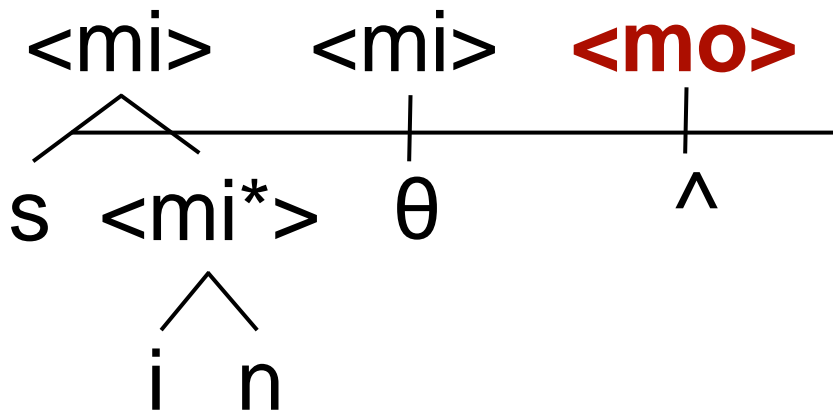
バッファ

2

例

REDUCE

スタック



$$\sin \frac{\hat{\theta}}{2}$$

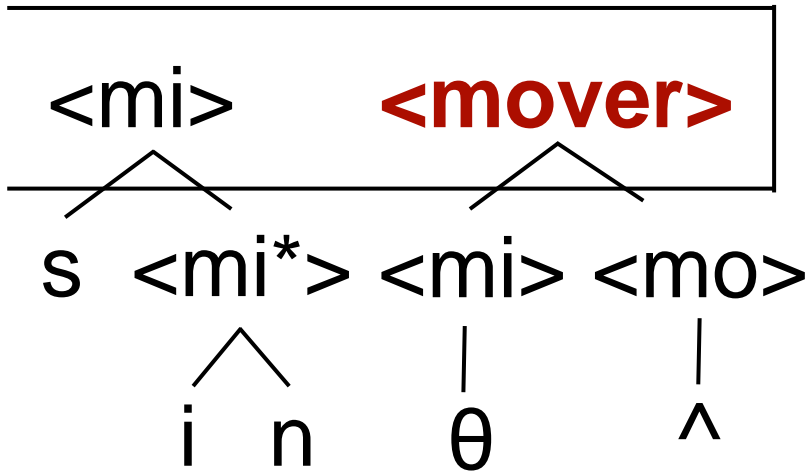
バッファ

2

例

REDUCE

スタック



バッファ

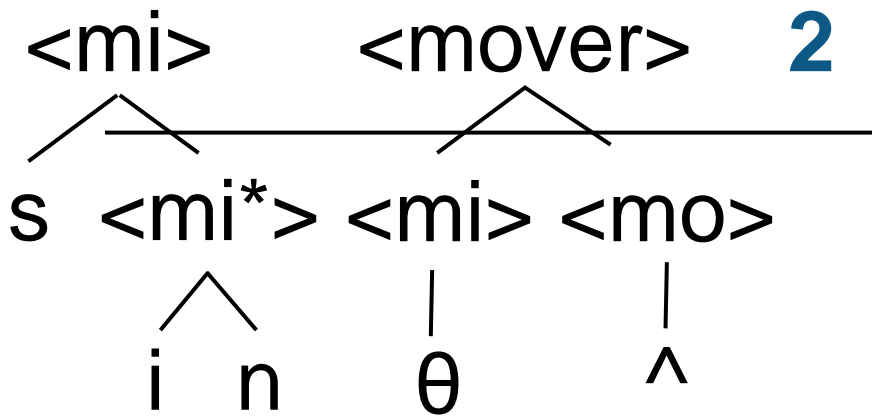
2

$$\sin \frac{\hat{\theta}}{2}$$

例

SHIFT

スタック



$$\sin \frac{\hat{\theta}}{2}$$

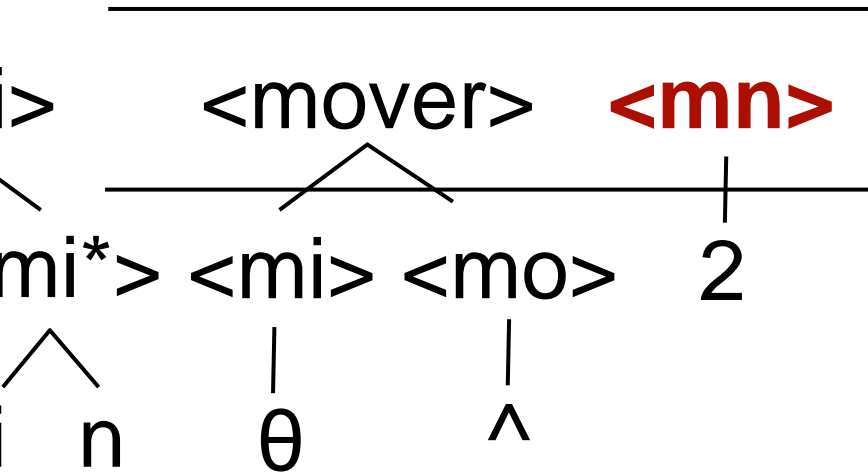
バッファ



例

REDUCE

スタック



$$\sin \frac{\hat{\theta}}{2}$$

バッファ



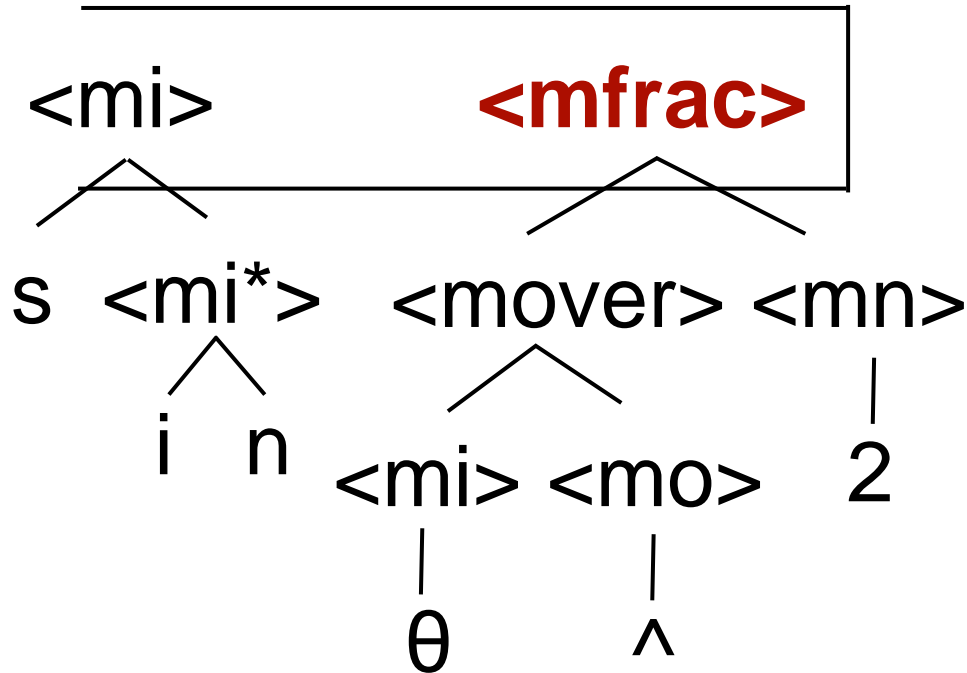
例

REDUCE

$$\sin \frac{\hat{\theta}}{2}$$

スタック

バッファ



例

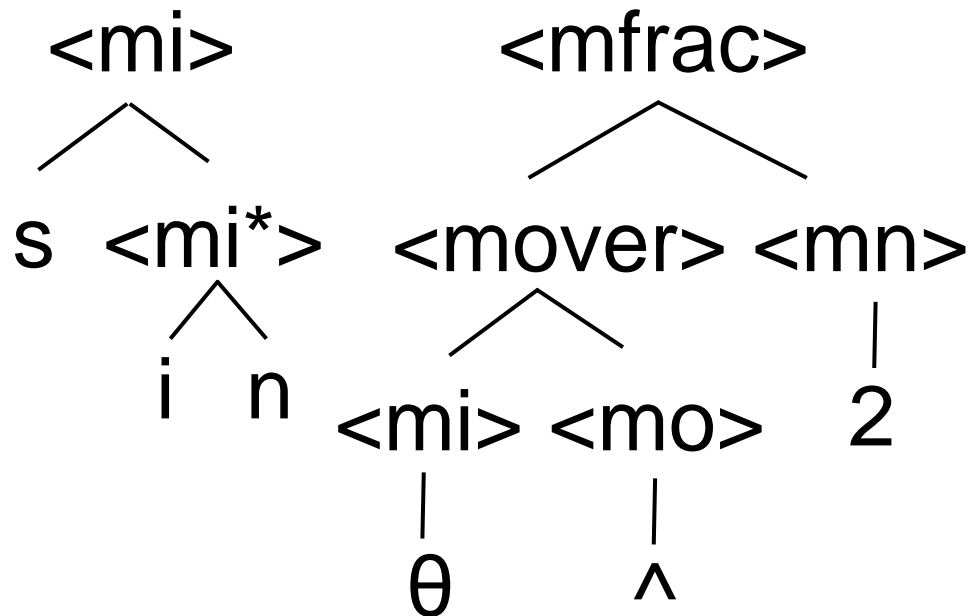
REDUCE

$$\sin \frac{\hat{\theta}}{2}$$

スタック

バッファ

<math>



自然言語処理における構造解析

構文解析と論文PDF解析の違い

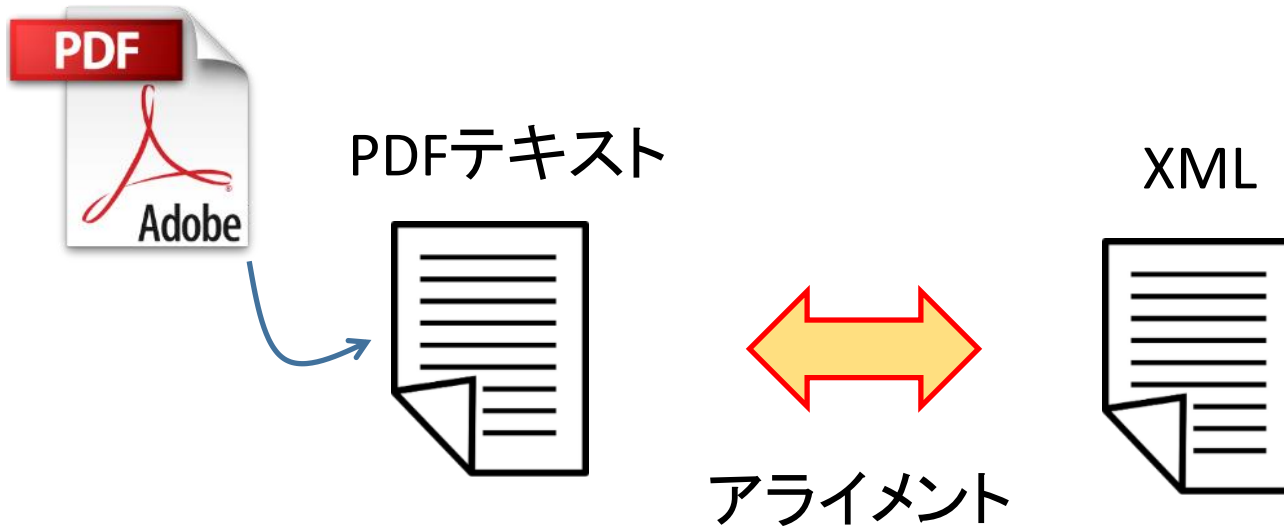
- 論文PDFは, 入力系列の長さが数万
 - ⇔ 構文解析では, 入力は20~30単語を想定している
 - 線形時間で動作するアルゴリズムでないと, 実質的に解析できない
- 数式や図表は, 本文テキストに割り込む.
- 表の中に数式が含まれることや, 数式の中に表が含まれる.
 - 複数の文法セットを用いた解析が必要
 - ⇔ 構文解析では, 単一の文法ルールにしたがって解析

学習データの構築

- PMC（生物医学系の論文）
 - PDFとXML(JATS)のペアを入手可能
 - 文献数： 200万件以上
- arXiv（計算機科学, 物理などの論文）
 - PDFとLaTeXソースのペアを入手可能
 - 文献数： 100万件以上
 - LaTeXはXMLへ変換できる

PDF2XML

学習データの構築



PDFとXMLのテキストは完全には一致しない

- 複数の文字を合成して一文字にする場合がある(合字)

例: **Refined**

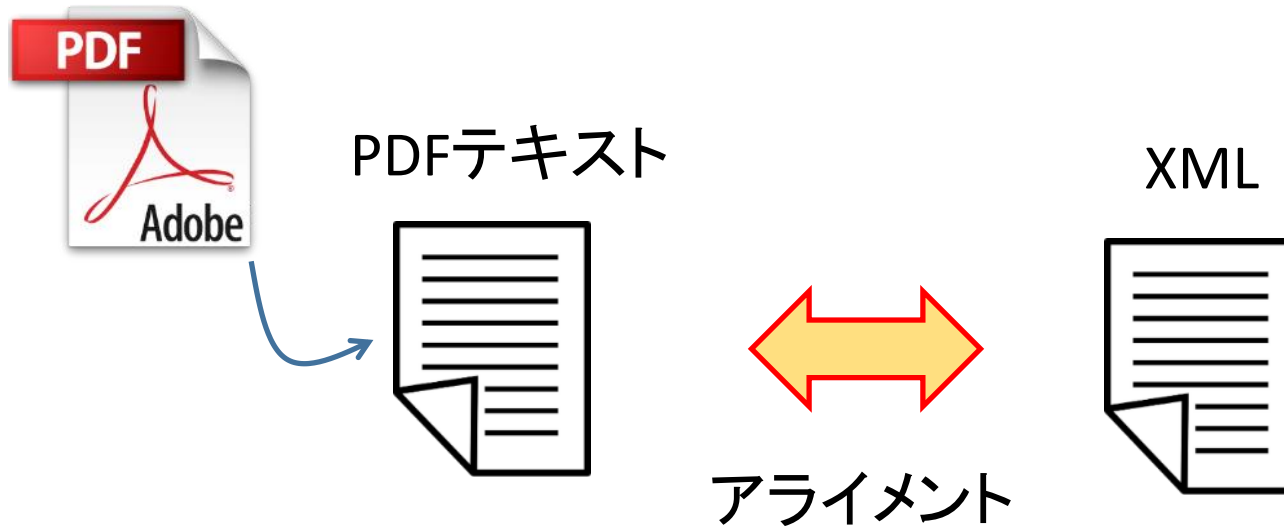
- 行末にハイフンが挿入される場合がある

We propose Symbol-Refined Tree Substitution Grammars (SR-TSGs) for syntactic parsing. An SR-TSG is an extension of the con-

- PDFから文字コードを読み取れない場合がある
 - グリフ(字形)と文字コードの対応が無い

PDF2XML

学習データの構築



系列アライメント

1. 動的計画法
 - 編集距離を求めるアルゴリズム
2. Diffのアルゴリズム
 - Myers' algorithm (1986)
 - Wu's algorithm (1989)
3. ゲシュタルトパターンマッチング
 - Pythonのdifflibで採用

系列アライメント

- 動的計画法

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0



系列長が数万なので、計算が
終わらない

系列アライメント

- Diffのアルゴリズム
 - Myers' algorithm (1986)
 - Wu's algorithm (1989)
- アルゴリズムがやや複雑
- 飛び飛びのマッチングになることがある



系列アライメント

- ゲシュタルトパターンマッチング
 - 最長共通部分文字列(連続する共通部分文字列の中で最長なもの)を再帰的に求めてマッチングする

G	E	E	K	S	F	O	R
G	E	E	K	S	Q	U	I

マッチした部分の右と左の文字列を切り出して繰り返し

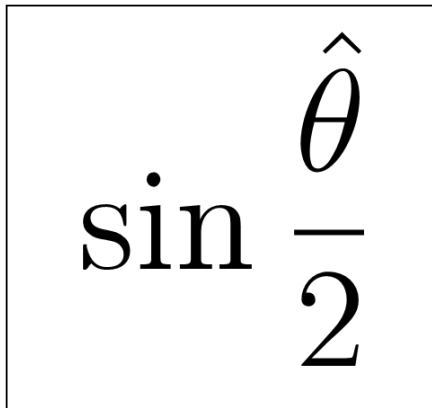
二つの文字列を連結してSuffix Arrayを構築すると、線形時間で求めることができる

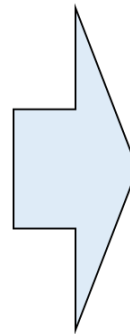
数式・表の解析

数式の解析

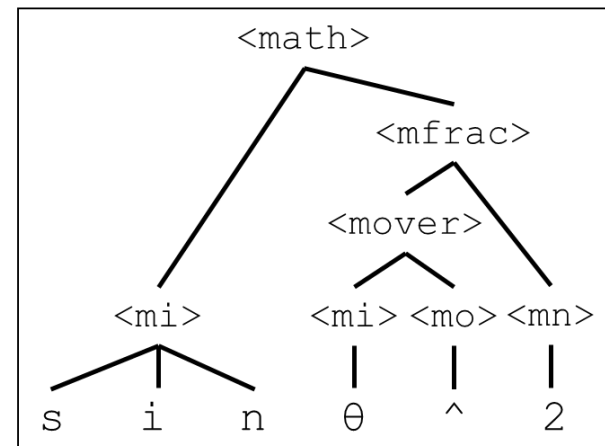
MathML: 数式構造のXML表現

IN: Math Expression in PDF


$$\sin \frac{\hat{\theta}}{2}$$



OUT: MathML



数式の解析

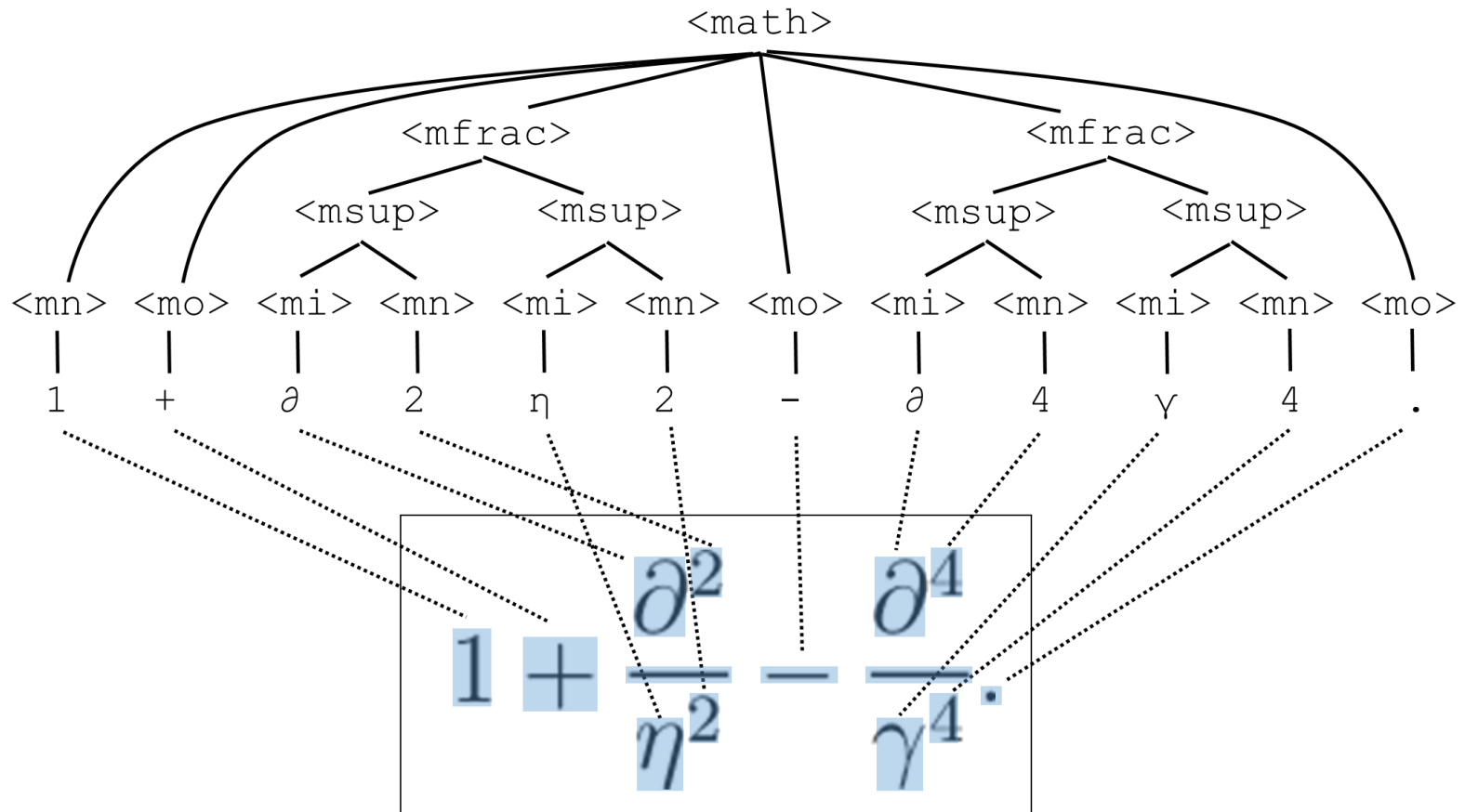
MathML: 数式構造のXML表現

数式解析の問題

- MathMLは曖昧性が高い
同じ数式を表現するために何通りもの書き方がある
例: 同じ数式をWord, Mathematica, MATLABで入力すると, 全て異なるMathMLが生成される場合がある. → 正規化が必要
- LaTeXからcompileしたPDFと, XMLをcompileしたPDFでは, 数式がPDFに書き込まれる順番が異なる.
→ 並べ替え＋構造解析の問題になる 😞

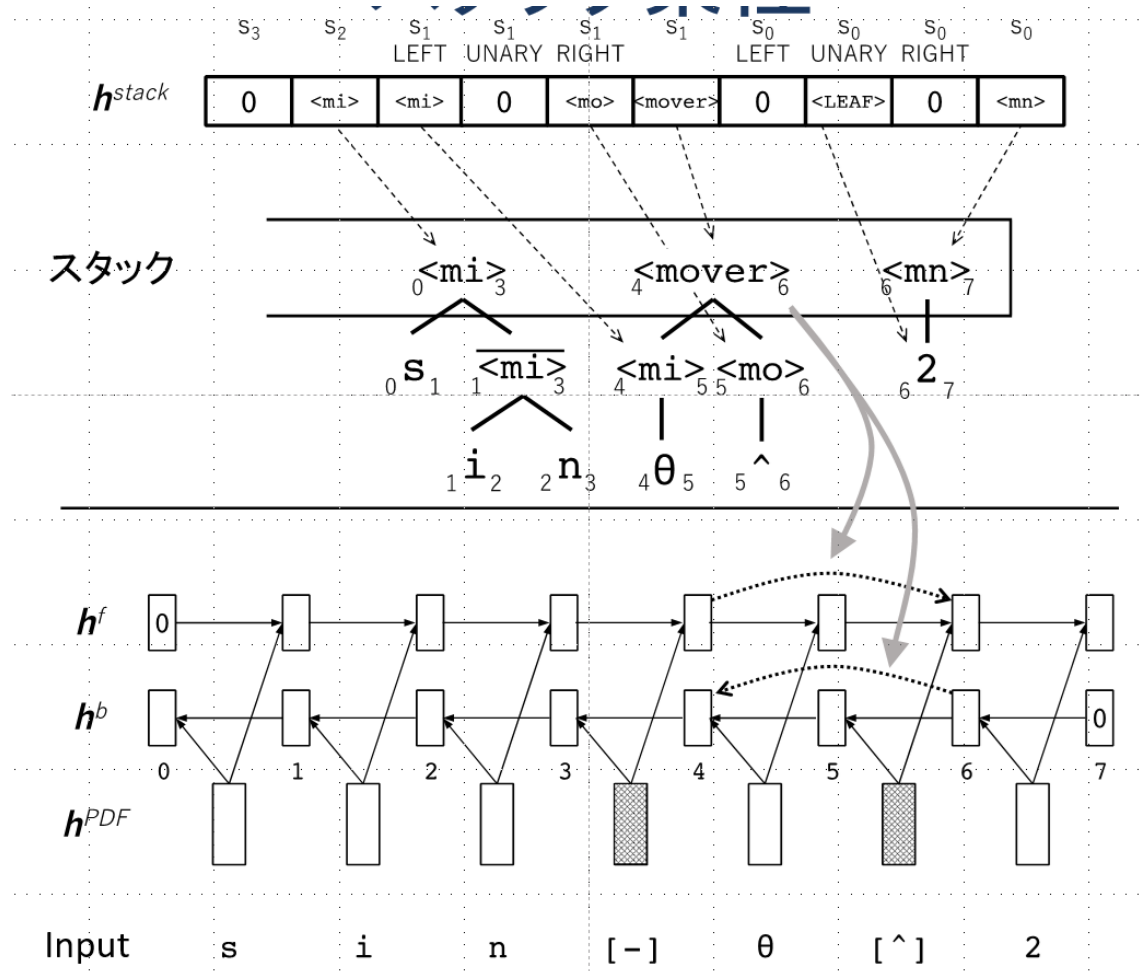
数式の解析

数式PDFとMathMLは、ほぼ1対1対応になっている



数式の解析

ニューラルネットワークでモデル化



数式の解析

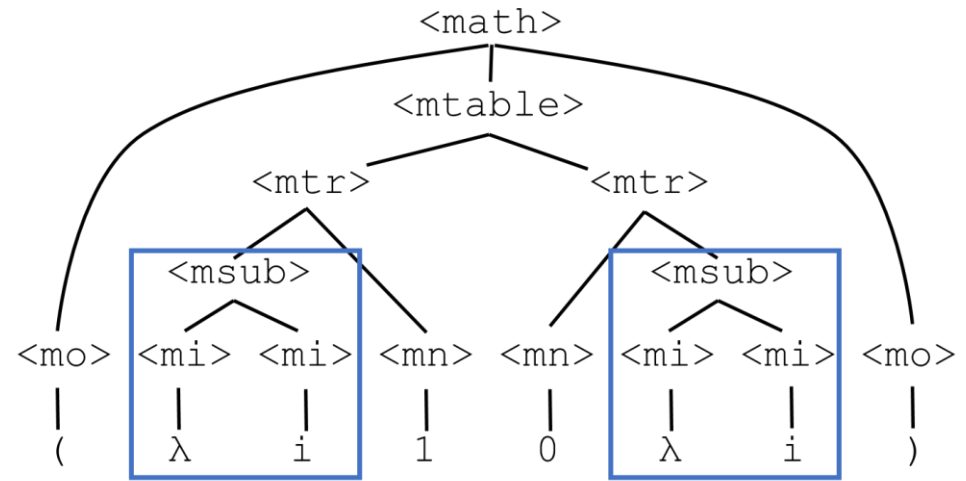
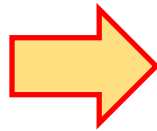
結果

比較手法	系列 (BLEU)	句構造木 (PARSEVAL)		
		適合率	再現率	F ₁ 値
ベースライン: ENCDEC-IMG-XML	81.73	0.790	0.804	0.797
提案手法1: ENCDEC-PDF-XML	81.05	0.812	0.841	0.826
提案手法2: ENCDEC-PDF-ACTION	82.53	0.790	0.838	0.814
提案手法3: FFNN-PDF-ACTION	88.96	0.912	0.893	0.902

数式の解析

解析が困難な例

$$\begin{pmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{pmatrix}$$



数式の解析

今後の改善策

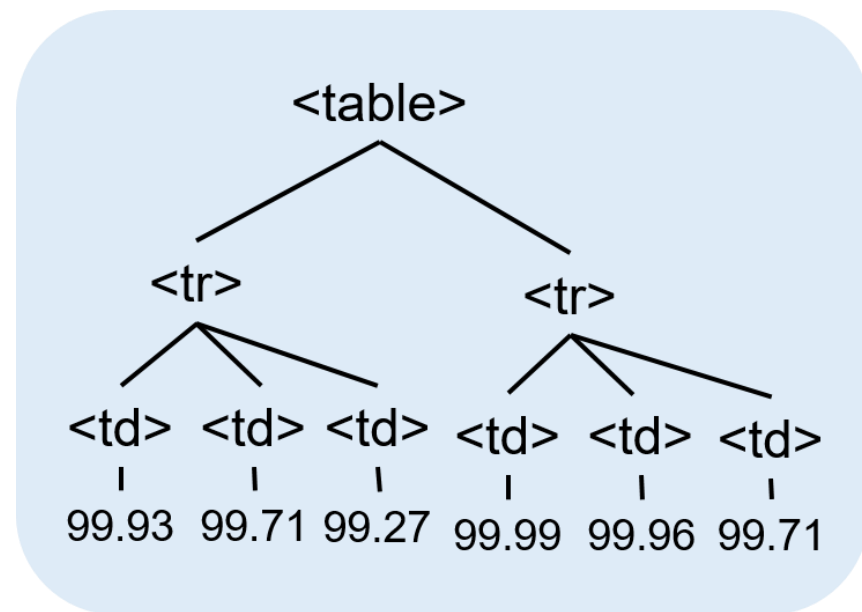
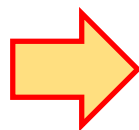
Data Augumentation

- 数式のLatex, XMLからPDFを生成することができる
→ データをいくらでも人工的に生成できる

表の解析

数式と同じ手法で解析できる

Threshold	Train	Oracle Dev	Test
$T=0.05$	99.93	99.71	99.27
$T=0.0005$	99.99	99.96	99.71
$T=0.00005$	100.0	99.98	99.83



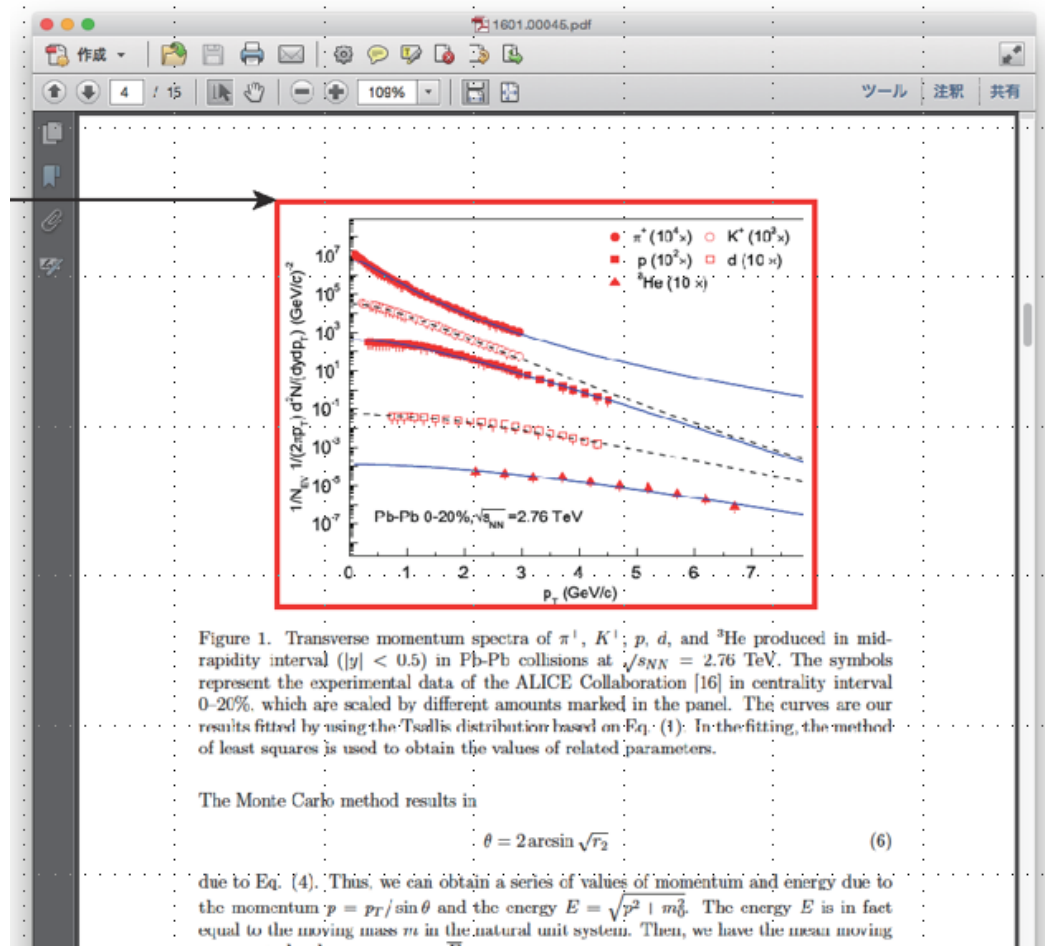
図の解析

図の解析

- PDF中の図の範囲同定
- 図の中身の解析

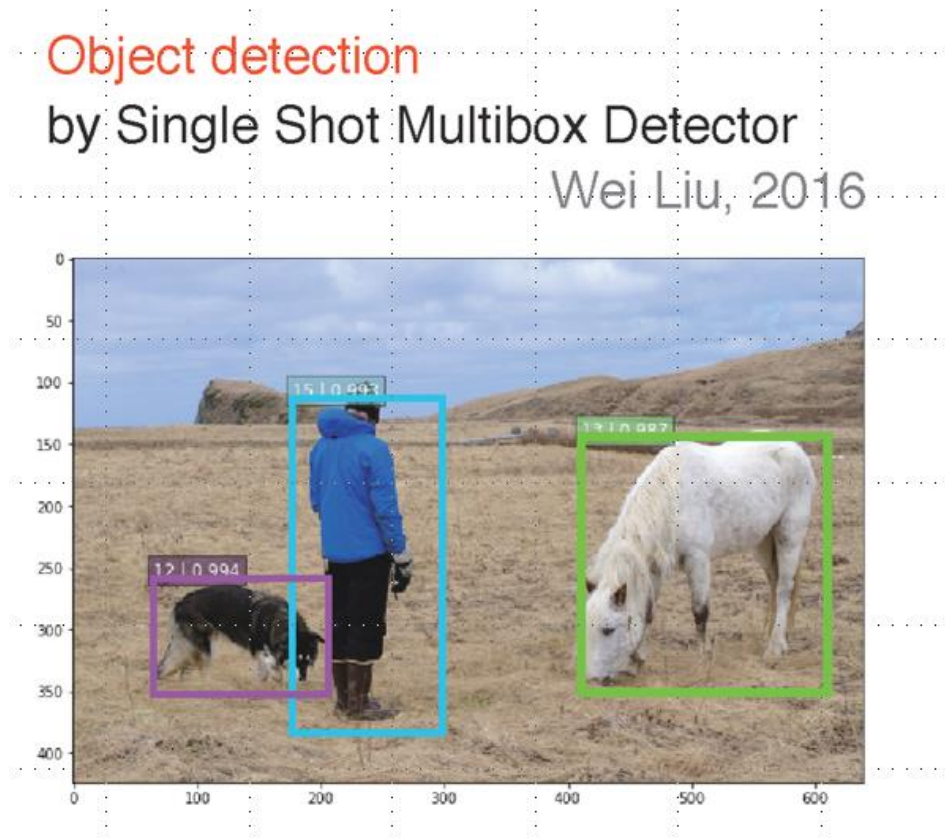
図の解析

- PDF中の図の範囲同定



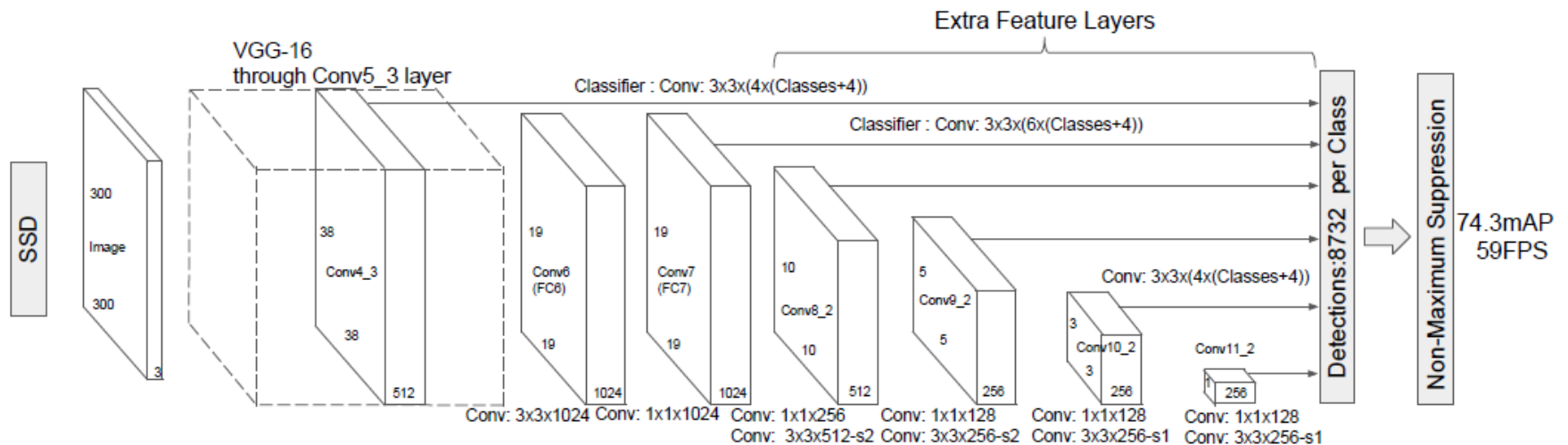
図の解析

- 一つのやり方は, PDFを画像化して, 画像認識(オブジェクト同定)する



図の解析

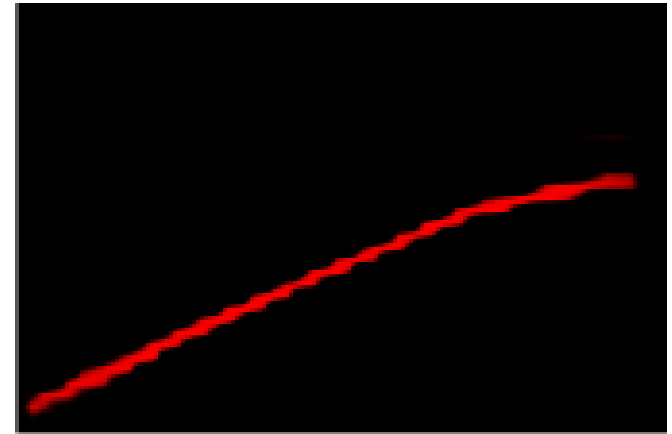
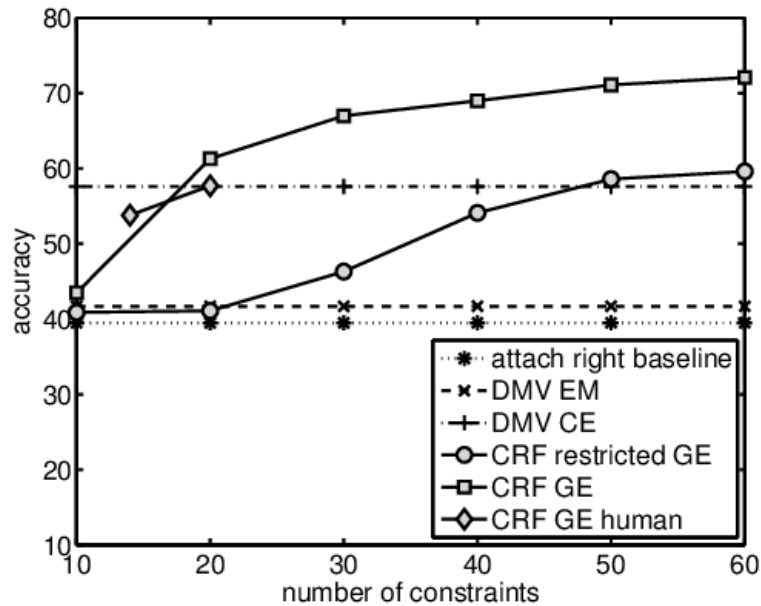
- 一つのやり方は、PDFを画像化して、画像認識(オブジェクト同定)する



- 精度は7~8割で、思ったより低い

Line Chart 解析

折れ線グラフの自動読み取り

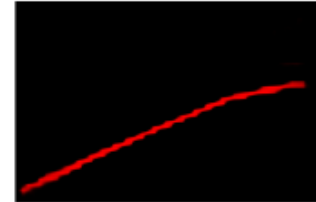
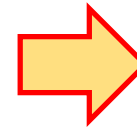
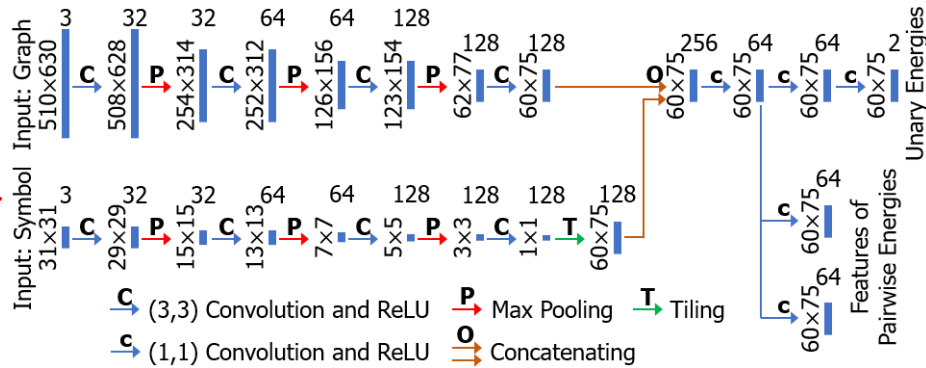
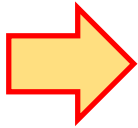
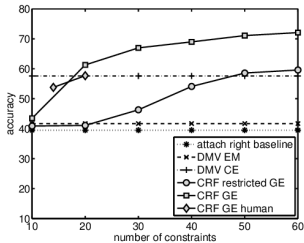


Statistics of genes: KNApSACN - v1.200.00

SuperKingdom	Kingdom	Order	Family	# of genes	# of Metabolites
Archaea					
	****	Methanobacteriales	Methanobacteriaceae	1	7
	****	Methanococcales	Methanococcaceae	1	6
Bacteria					
	****	Actinomycetales	Micrococaceae	1	3
	****	Actinomycetales	Streptomyetaceae	1	69
	****	Bacillales	Akrycylbactilaceae	1	10
	****	Bacillales	Bacillaceae	1	5
	****	Bacillales	Diaphylococcaceae	1	1
	****	Burkholderiales	Caulobacteriaceae	1	1
	****	Enterobacteriales	Enterobacteriaceae	3	19
	****	Flavobacteriales	Flavobacteriaceae	1	2
	****	Lactobacillales	Lactobacillaceae	1	1
	****	Mycococcales	Cystobacteriaceae	1	1
	****	Mycococcales	Mycococcaceae	1	13

Line Chart 解析

畳み込みニューラルネットワーク



結果

Method	Accuracy [%]	F1 score
FigureSeer	19.4	0.462
Proposed	50.5	0.838
Unary	35.4	0.785
Small	19.7	0.642

まとめ

- PDFを構造解析することは、論文からの知識獲得に必要不可欠
- 数式, 表の解析はある程度上手いく.
 - これから様々な分野で性能評価を行う予定
- 図の解析はまだまだ難しい