1．自然言語処理とは

1.1 自然言語処理概略

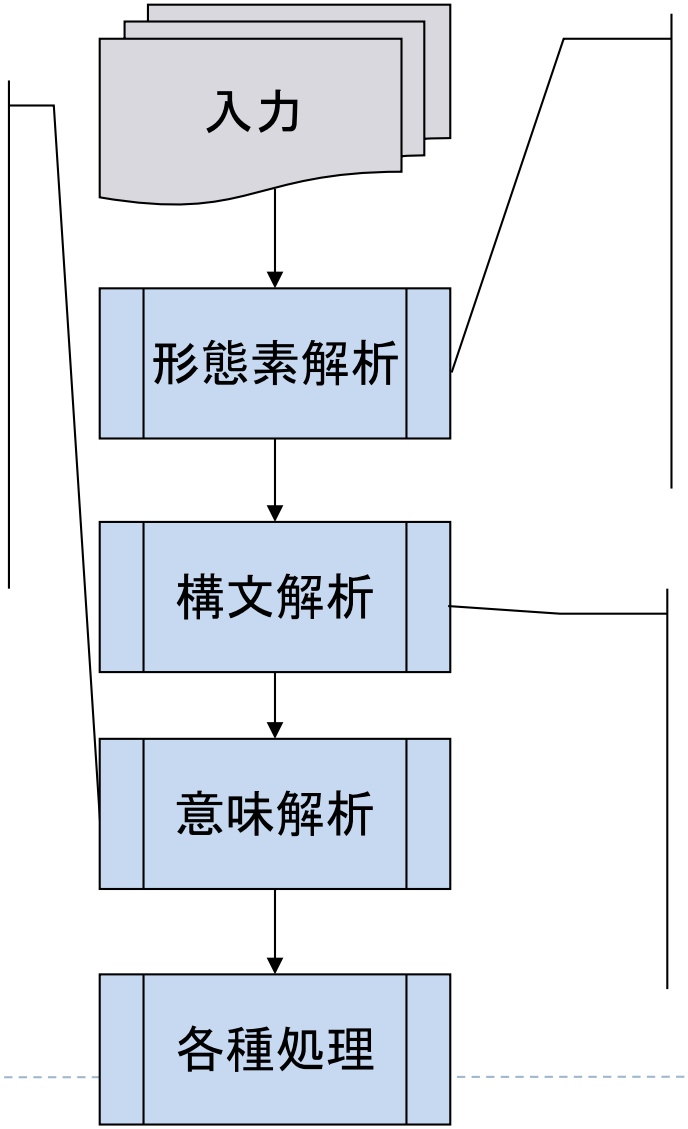
- 言語
  - プログラミング言語（人工言語）
  - 自然言語（ex. 日本語、英語、中国語、…）
- 自然言語処理の目的

  人間が使う言語を処理する
  - 話す言葉、音声言語
  - 書いた文,文章の解析
  - 文章からの情報抽出
  - 文,文章の生成
- 方法
  - 認知的原理、モデル、知見に基づくものもあるが。
  - 必ずしも認知的である必要はない。

自然言語処理過程の典型的モデル

一文について
- 格フレーム等により知識表現する。

文章全体について
- 一文の意味表現を文間の関係に応じて連結する。



1. 分かち書き（日本語）
2. 活用している時の原形、活用形、品詞（形態素）を同定
3. 必要なら語ごとに辞書から各種情報を得、次以降の過程で利用

- 品詞（カテゴリ）の並びについての文法
- 入力に対してどのように文法規則が適用されているのかを解析

文字の扱い
ASCII Code Table

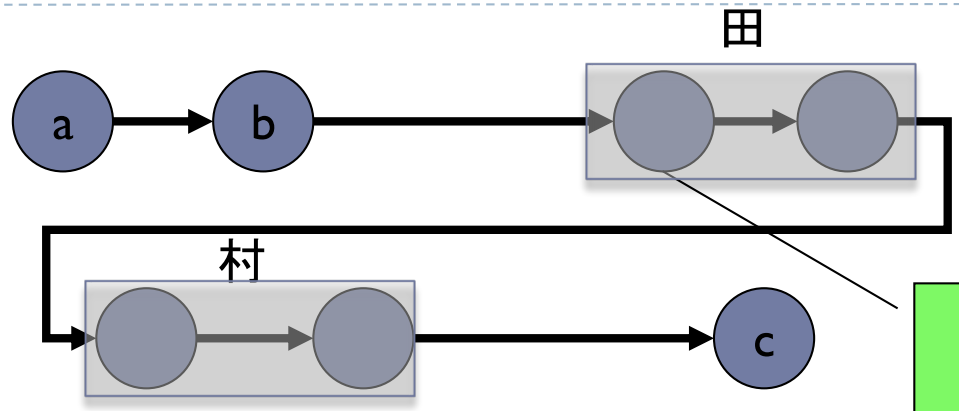
例　Book

文字	二進数	16進数
‘B’	1000010	42
‘o’	1101111	6f
‘o’	1101111	6f
‘k’	1101011	6b

b6 b5 b4	000	001	010	011	100	101	110	111
b3–b0								
0000	*nul*	*dle*		0	@	P	`	p
0001	*soh*	*dc1*	!	1	A	Q	a	q
0010	*stx*	*dc2*	″	2	B	R	b	r
0011	*etx*	*dc3*	#	3	C	S	c	s
0100	*eot*	*dc4*	$	4	D	T	d	t
0101	*enq*	*nak*	%	5	E	U	e	u
0110	*ack*	*syn*	&	6	F	V	f	v
0111	*bel*	*etb*	’	7	G	W	g	w
1000	*bs*	*can*	(	8	H	X	h	x
1001	*ht*	*em*	)	9	I	Y	i	y
1010	*lf*	*sub*	*	:	J	Z	j	z
1011	*vt*	*esc*	+	;	K	[	k	{
1100	*ff*	*fs*	,	<	L	¥	l	|
1101	*cr*	*gs*	−	=	M	]	m	}
1110	*so*	*rs*	.	>	N	^	n	~
1111	*si*	*us*	/	?	O	_	o	*del*

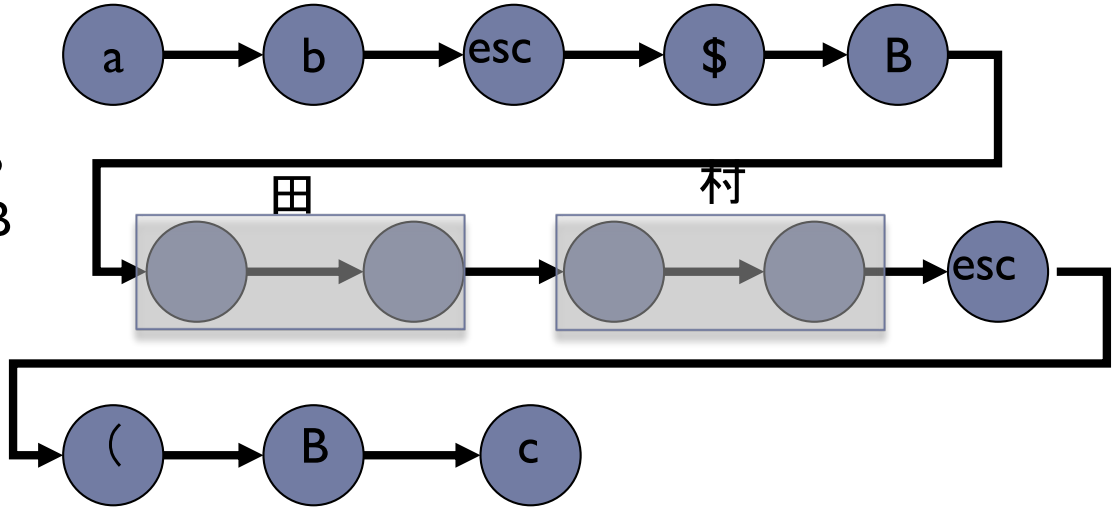
コード化（EUC, SJIS, JIS）
「ab田村c」の各コード

ShiftJIS

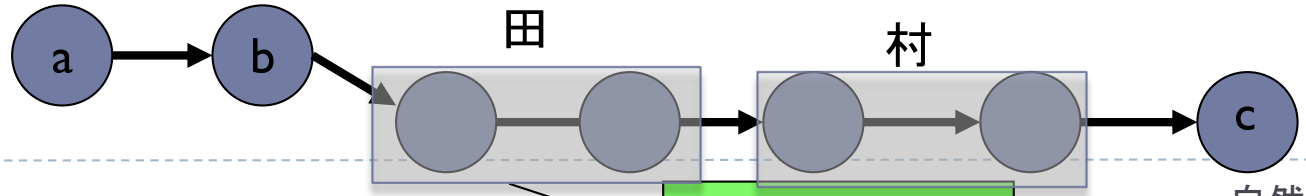


コード表の特別な位置

JIS
モード切り替え
漢字in : esc+$+B
漢字out:esc+(+B



EUC



8ビット目は1

文字集合の例（JIS X 0208)

文字集合(JIS X 0208)

- コンピュータで扱える文字の集合(Unicode)とその符号化(UTF-8)を
- 分離して考える。
- 一つの文字を区(row.1~94)点(cell.1~94)で表す。
  例：「医」＝16区69点（または16-69)
- 8,836(=94×94)字を扱うことが出来る。
- この構造は中国のGB2312、韓国のKS X 1001でも採用。

Row 16	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
00		亜	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪	握	渥	旭	葦	芦	鯵
20	梓	圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎	或	粟	袷	安	庵	按	暗	案	闇
40	鞍	杏	以	伊	位	依	偉	囲	夷	委	威	尉	惟	意	慰	易	椅	為	畏	異
60	移	維	緯	胃	萎	衣	謂	違	遺	医	井	亥	域	育	郁	磯	一	壱	溢	逸
80	稲	茨	芋	鰯	允	印	咽	員	因	姻	引	飲	淫	胤	蔭					

Row 17	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
00		院	陰	隠	韻	吋	右	宇	烏	羽	迂	雨	卯	鵜	窺	丑	碓	臼	渦	嘘
20	唄	欝	蔚	鰻	姥	厩	浦	瓜	閏	噂	云	運	雲	荏	餌	叡	営	嬰	影	映
40	曳	栄	永	泳	洩	瑛	盈	穎	頴	英	衛	詠	鋭	液	疫	益	駅	悦	謁	越
60	閲	榎	厭	円	園	堰	奄	宴	延	怨	掩	援	沿	演	炎	焔	煙	燕	猿	縁
80	艶	苑	薗	遠	鉛	鴛	塩	於	汚	甥	凹	央	奥	往	応					

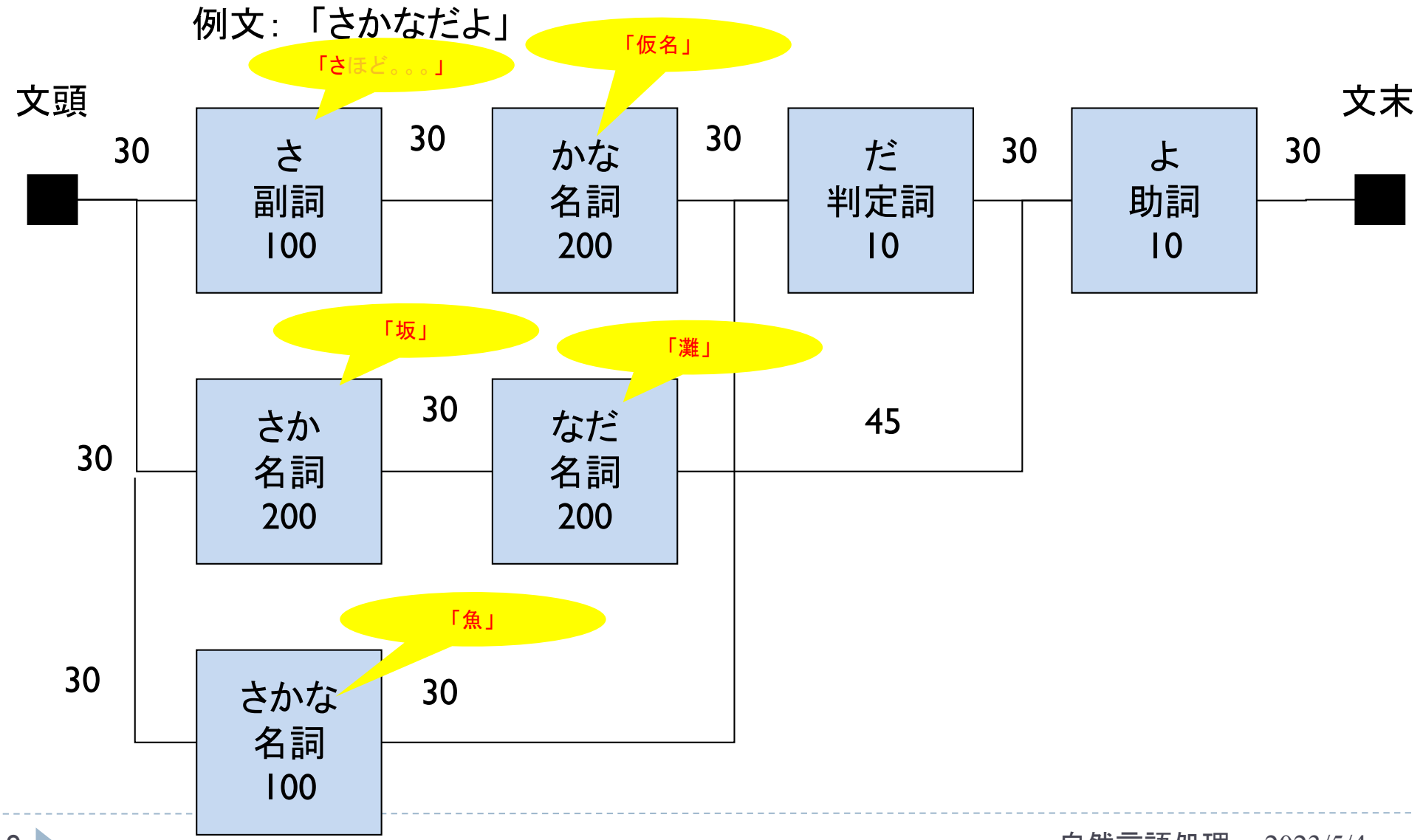
文字集合（Unicode）に対する文字符号化（UTF-8）

文字集合	文字符号化
U+0000~U+007F	00~7F (7bit)
(ASCIIと同一）	
U+0080~U+07FF	C280~DFBF (11bit)
(非漢字の一部）	
U+0800~U+FFFF	E08080~EFBFBF (16bit)
(残りの漢字）	
U+10000~U+1FFFFF	F0808080~F7BFBFBF (21bit)
(第3、第4水準漢字）	

形態素解析
解析結果の例

表層	読み	基本形	品詞	活用形	活用
いま	(いま)	いま	副詞		
米国	(べいこく)	米国	地名		
から	(から)	から	格助詞		
は	(は)	は	副助詞		
さまざまな	(さまざまな)	さまざまだ	形容詞	ナノ形容詞	ダ列基本連体
警報	(けいほう)	警報	普通名詞		
が	(が)	が	格助詞		
発せ	(はっせ)	発する	動詞	サ変動詞	文語未然形
られて	(られて)	られる	動詞性接尾辞	母音動詞	タ系連用テ形
いる	(いる)	いる	動詞性接尾辞	母音動詞	基本形
。	(。)	。	句点		
EOS					
商務	(しょうむ)	商務	普通名詞		
省	(しょう)	省	普通名詞		
の	(の)	の	接続助詞		
日本	(にっぽん)	日本	地名		
など	(など)	など	副助詞		
鉄鋼	(てっこう)	鉄鋼	普通名詞		
メーカー	(めーかー)	メーカー	普通名詞		
製品	(せいひん)	製品	普通名詞		
に	(に)	に	格助詞		
対する	(たいする)	対する	動詞	サ変動詞	基本形
ダンピング	(ダンピング)	ダンピング	カタカナ		
仮	(かり)	仮	普通名詞		
決定	(けってい)	決定	サ変名詞		
や	(や)	や	接続助詞		
上院	(じょういん)	上院	普通名詞		
…….					
EOS					

「茶筌」のコスト最小法


例文：「さかなだよ」
「さほど。。。」
「仮名」
文頭
文末
30
さ
副詞
100
30
かな
名詞
200
30
だ
判定詞
10
30
よ
助詞
10
30
「坂」
「灘」
さか
名詞
200
30
なだ
名詞
200
45
30
「魚」
さかな
名詞
100
30

文脈自由文法の例 $G_k=(V_k,\Sigma_k,P_k,文)$

Σ_k={きた, 文化, 伝わった, から, が}

V_k={文, 後置詞句, 動詞, 名詞, 助詞}

$\cup\ \Sigma_k$

P_k = { 文 → 後置詞句 文,

文 → 後置詞句 動詞,

後置詞句 → 名詞 助詞,

後置詞句 → 動詞 助詞,

名詞 → きた, 名詞 → 文化

動詞 → きた, 動詞 → 伝わった

助詞 → から, 助詞 → が }

導出の例

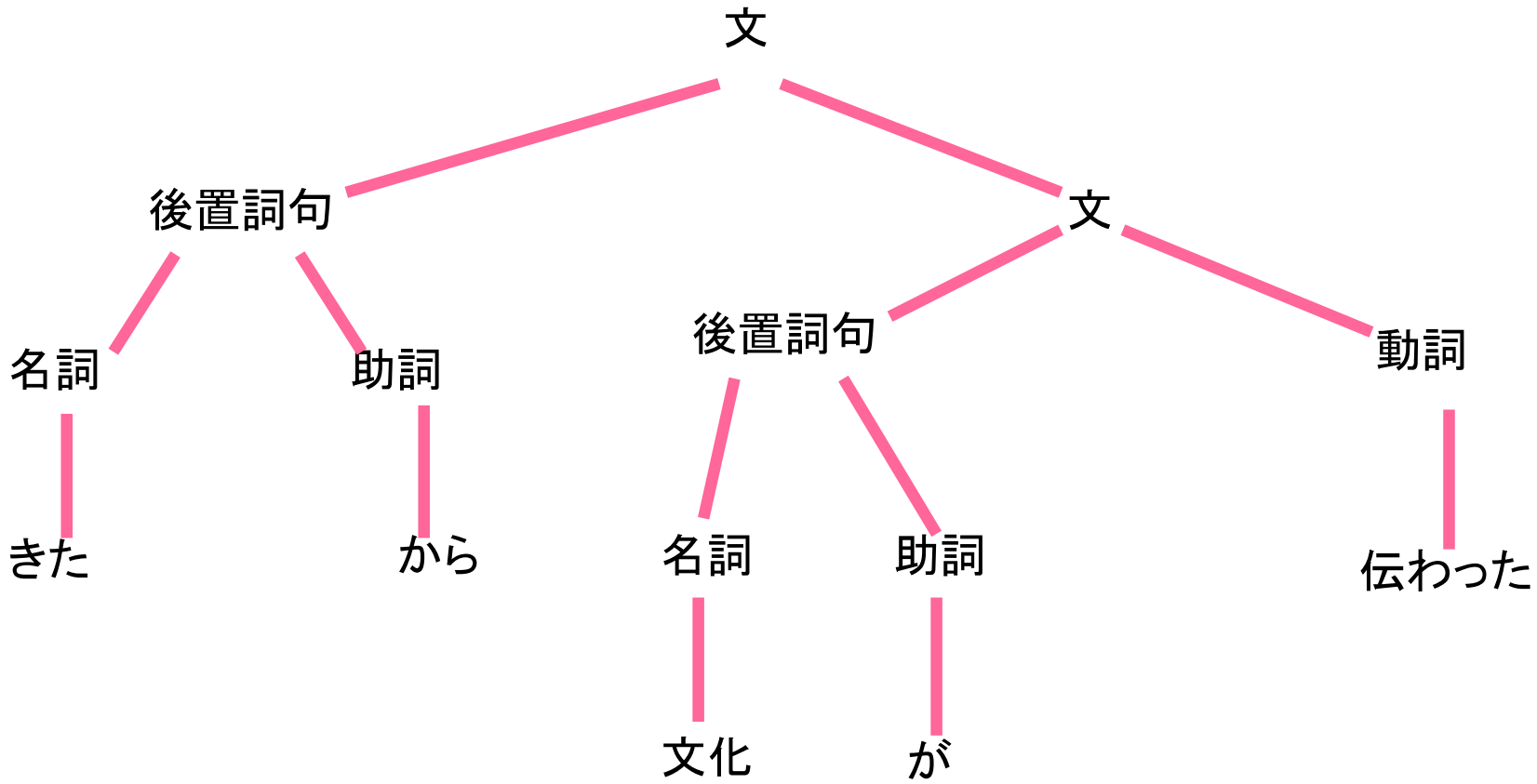
解析する文 w_k= きた　から　文化　が　伝わった

導出

文　⇒　後置詞句　文
⇒　名詞　助詞　文
⇒　きた　助詞　文
⇒　きた　から　文
⇒　きた　から　後置詞句　動詞
⇒　きた　から　名詞　助詞　動詞
⇒　きた　から　文化　助詞　動詞
⇒　きた　から　文化　が　動詞
⇒　きた　から　文化　が　伝わった

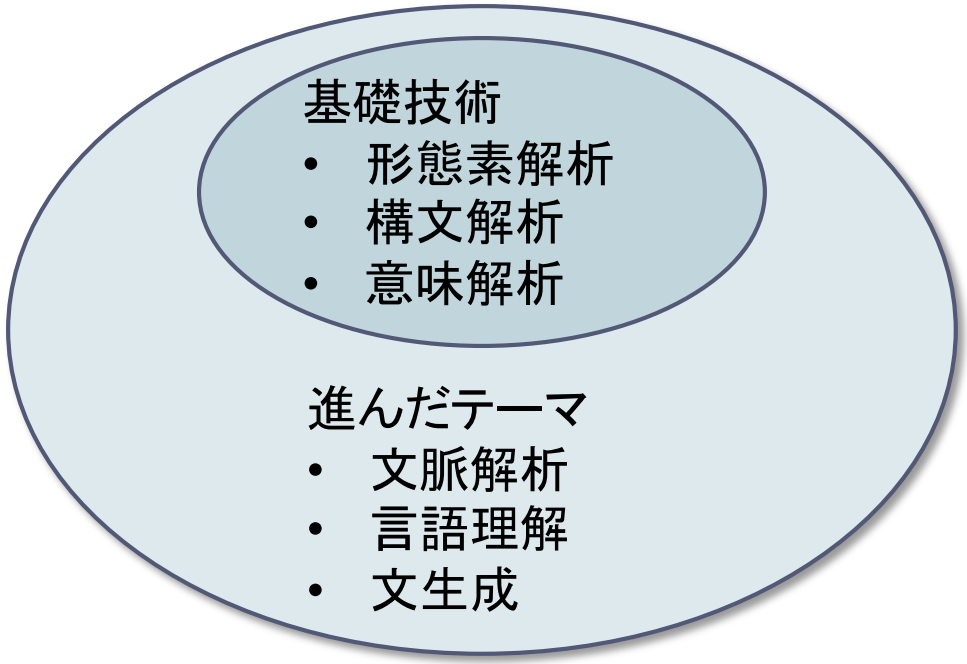
構文木（例）

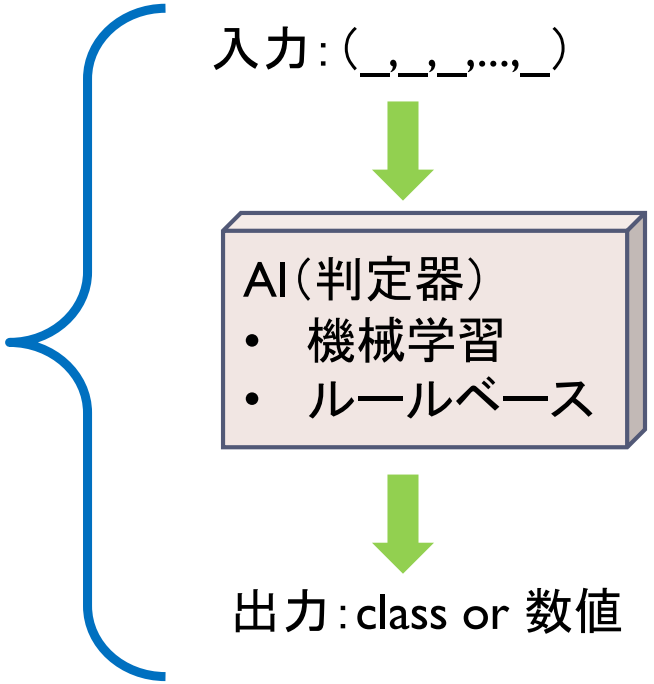
w_k = きた　から　文化　が　伝わった

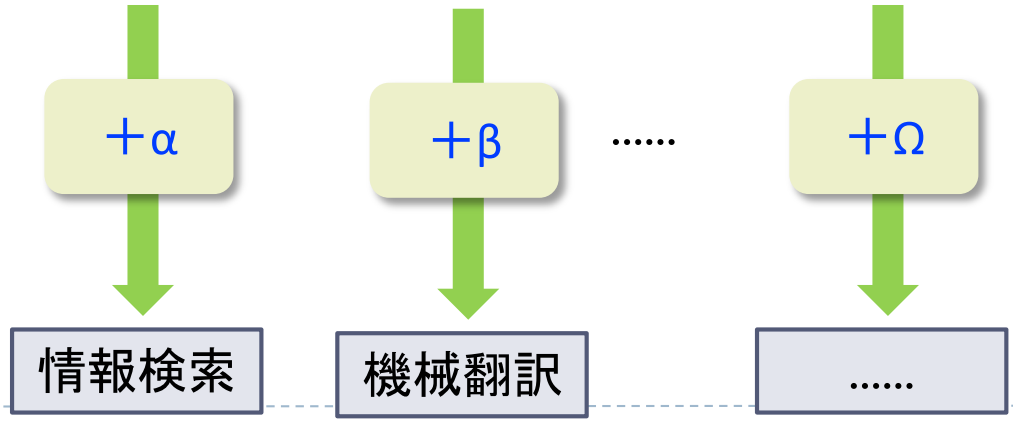


構文木（生成木、導出木）による文の構文構造の表現

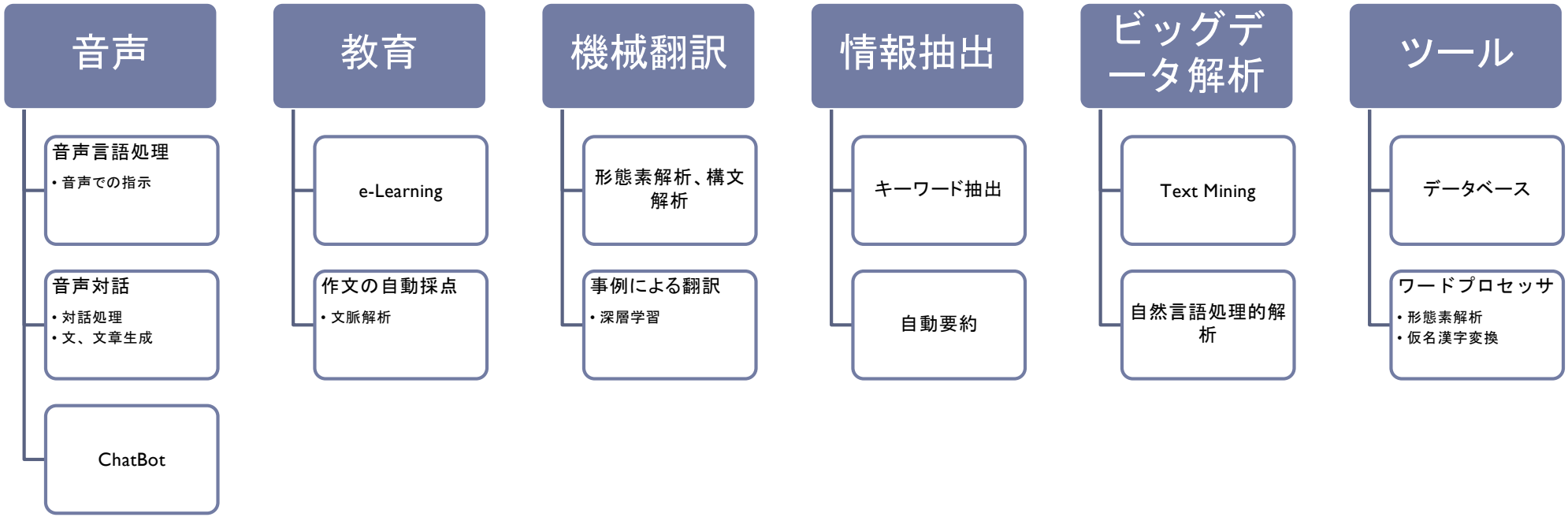
「自然言語処理」に関連する各種技術
～～～本講義で扱う各技術の位置づけ～～～







自然言語処理の応用分野



1.2 自然言語処理関連の歴史

1940年代、W.Weaver, A.D.Booth

- 機械翻訳（英←→仏（France, 法））
- 単語の置き換えによる

1954年、ジョージタウン大、IBM

- ロシア語→英語

- 直接翻訳方式。英語、ドイツ語、フランス語間である程度成功

  step1) 単語レベルの置換。単語の活用形を全て展開

  step2) 目的言語で語順入れ替え

1957年、N.Chomsky、“Syntactic Structure”

- 句構造文法
  - 文：句構造（木構造）、変形により疑問文、受け身文などを文生成

1960年代

- 辞書、形態素解析、文法の導入・・・→ 構文解析

予想分析法（predictive analyzer）

- 文を途中まで聞いて構文構造を理解。予想しながら聞いている。

1961年、C.Hockett、Grammar for the Hearer

- Rhodes、ロシア語の解析文法
- 久野暲、英語文法

このような解析手法の研究が進んだが、構文的曖昧性が問題

They are flying planes.

Time flies like an arrow.

He saw a woman in the garden with a telescope.

日本

1955年～、九州大学、電気試験所

- 英独（徳）日間の機械翻訳

1959年、機械翻訳「やまと」（英→日）、九州大学（英独日）

- 漢字を扱えない→仮名
- 形態素解析が難しい　－－－－－－＞　ワープロ（1978年）出現まで

1966年、米（美）、ALPAC報告書

- 機械翻訳は当分無理
- もっと基礎研究が必要

1960~1970前半、暗黒の時代

事務処理

1959年、IBM KWIC (H.P.Luhn)、検索結果の表示法

1964年、MEDLARS、医学文献の情報検索システム

1950年～1960年、オートマトン

意味解析の重要性

構文解析→精密

⇒ 構文的曖昧性→増加

⇒ 意味解析→重要

1968年、C.Fillmore、格文法(case grammar)

He opened the door by a key

A key opened the door

行為者(agent):he、対象(object):door、道具(instrument):key

それぞれの動詞について意味素(数十～数百個)

- 各動詞がどのような意味の名詞を格にとるか

⇒ 構文的曖昧性→小

情報検索

- 書誌的事項（本、論文の表題、著者名）→論文の抄録作成
- 1971年、MEDLINE（医学分野の世界最大のDB）

1976年～、日本科学技術情報センター

- 本文文書全体を検索語（書誌的事項、キーワード）でscan
- 全文検索（full text search）

人工知能（artificial intelligence）

1966年、ELIZA（対話システム）

⇒ 知識が重要

場面の分割が重要（T.Winograd）

代名詞の照応、曖昧さの解消、... 場面の認識により解消

計算言語学

1984年、M.Kay、functional unification grammar

- 隣接する2つの要素（単語）を一つ上の要素とする

LFG, GPSG, HPSG

LFG

- 2，3の結合規則のみ　＋　単語に対して豊富な属性情報

言語理論の限界　⇒　大量の文章データ

1985年頃、英語フランス語の対訳を確率的に解析

→　文法的、語彙的情報の抽出

1文単位の解析では不十分

- 文脈情報、場面知識

まず、隣接する文相互間の関係、解析

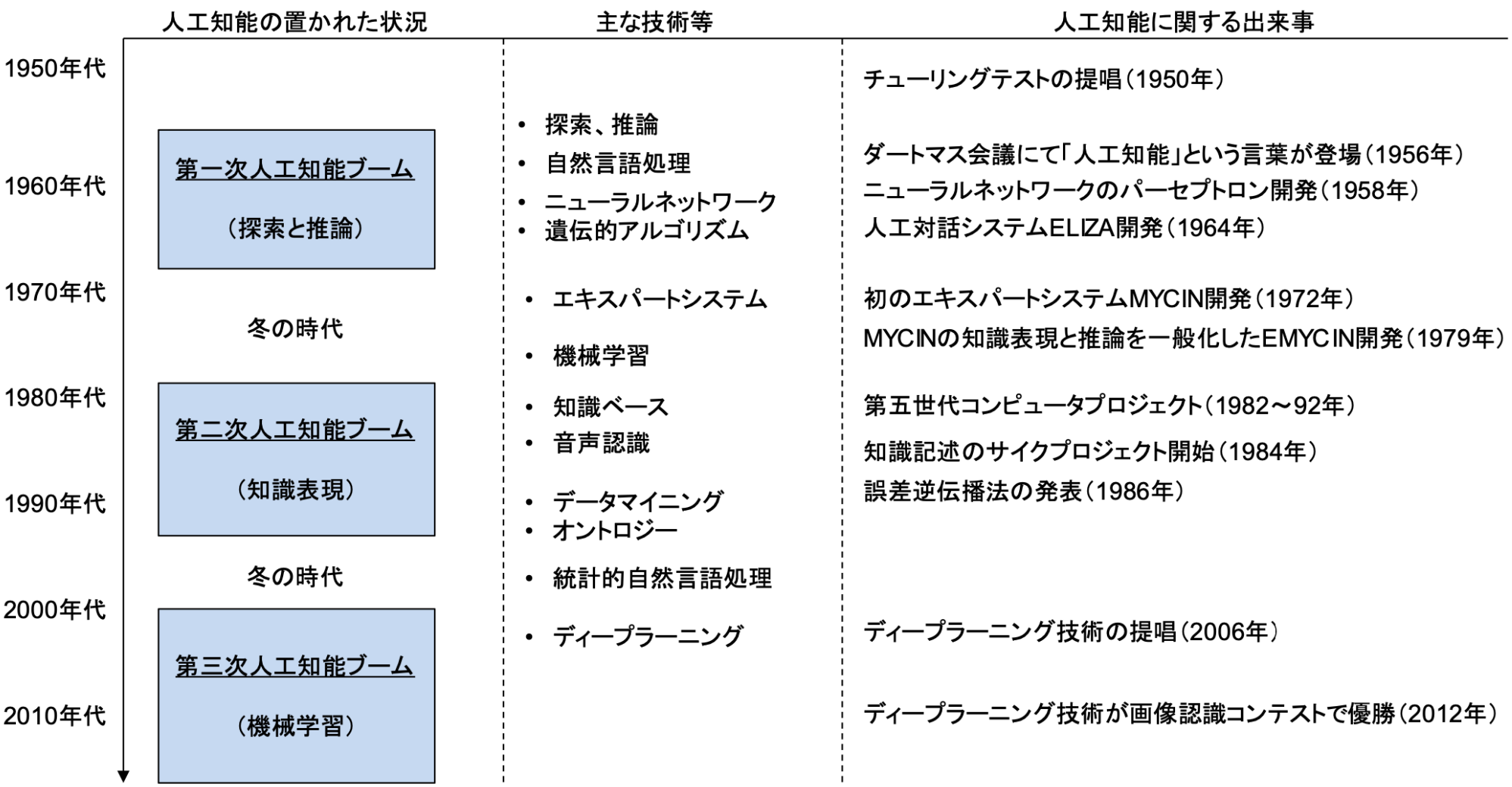
- 代名詞の照応、省略、話題、主題、焦点、
- 隣接する文間の関係（並列、理由、説明、．．．）

例文主導翻訳、実例型機械翻訳、用例翻訳

種々の特有の表現を例文（例句）として集め、規則化

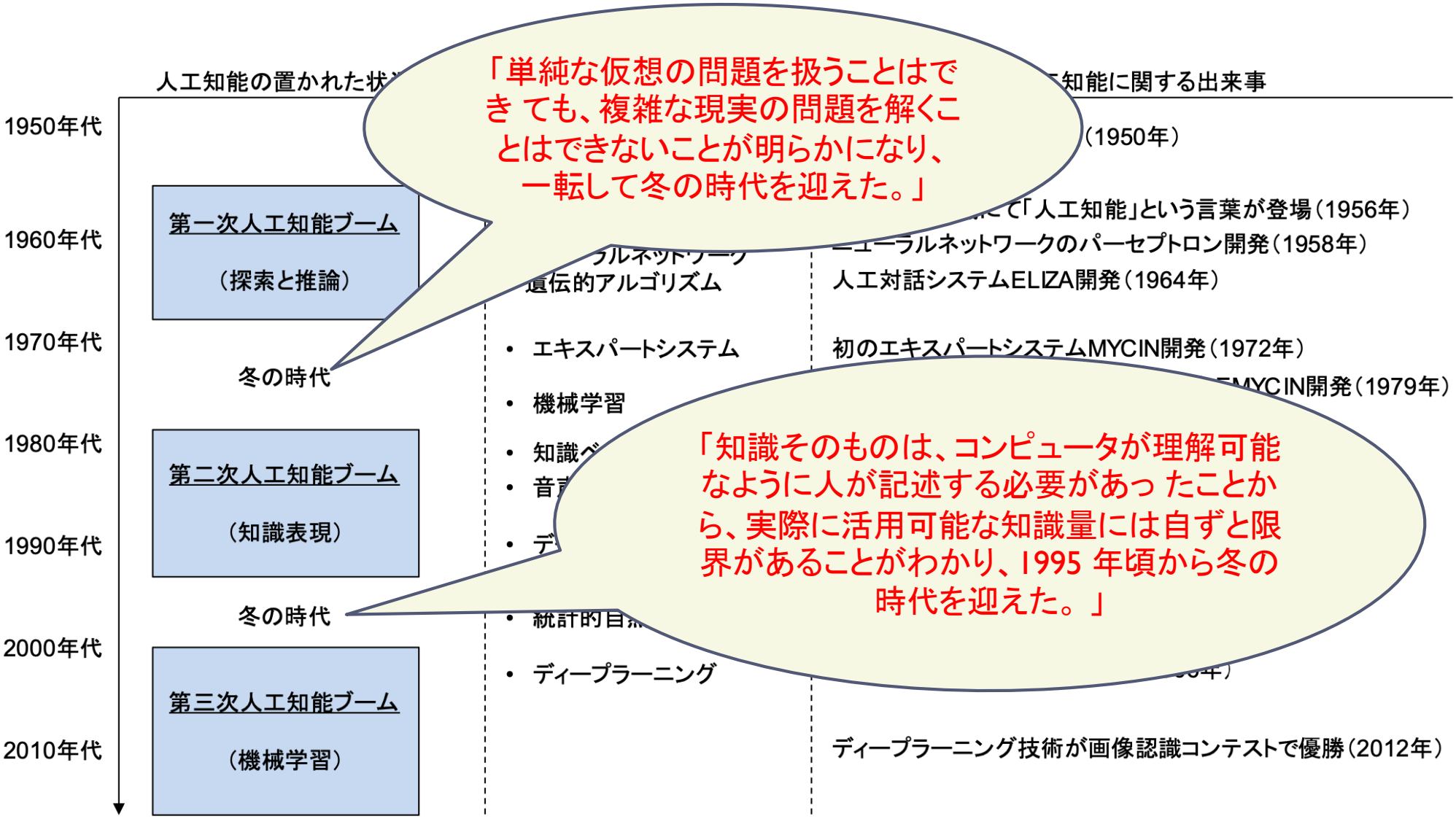
AIの歴史

総務省「ITCの進化が雇傭と働き方に及ぼす調査研究（H28）



AIの歴史

総務省「ITCの進化が雇傭と働き方に及ぼす調査研究（H28）



歴史的な対話システム（AIの歴史）

- SHRDLU（1970年頃、テリー・ウィノグラード）
  積み木の世界での積み木いじり操作に
  限定することにより、人間とコン
  ピュータとの会話システムを構築した。
  映像（SHRDLU in Action, by Alec Julien
  https://www.youtube.com/watch?v=bo4RvYJYOzI）
- ELIZA（1965年頃、MITのジョセフ・ワイゼンバウム）
  DOCTORという来談者中心療法のセラピストを
  シミュレート。
  Emacs上で、M-x doctor
- MYCIN（1970年頃、スタンフォード大）
  伝染性の血液疾患を診断し、抗生物質を
  推奨するエキスパートシステム

まとめ

1. 自然言語処理の概要
    1. 自然言語処理の典型的モデル
    2. コンピュータ内での文字の扱い
    3. 形態素解析
    4. 構文解析
    5. 各技術の位置づけ
2. 自然言語処理の歴史

課題

算術式を記述できる文法G_hにより、次の式を構文木で表せ。

$G_h=(\{S,T,F\}\cup\Sigma_h,\ \Sigma_h, P_h, S)$

$\Sigma_h=\{id, +, -, \times, \div, (,)\}$

$P_h=\{S\rightarrow S+T,\ S\rightarrow S-T,$

$S\rightarrow T,$

$T\rightarrow T\times F,\ T\rightarrow T\div F,$

$T\rightarrow F,$

$F\rightarrow id,$

$F\rightarrow (S)\}$

1. id + id +id
2. id × id + id
3. id × (id + id)