言語処理基盤としての言語資源

ータグ付きコーパス，生コーパス，そして新聞記事からウェブへー

河原 大輔
情報通信研究機構

言語処理学会第13回年次大会チュートリアル (2007/03/19)
http://www2.nict.go.jp/x/x161/member/kawahara/NLP2007tut.pdf

言語資源とは

The term language resources refers to sets of language data and descriptions in machine readable form, used specifically for building, improving or evaluating natural language and speech algorithms or systems, and in general, as core resources for the software localization and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users. Examples of linguistic resources are written and spoken corpora, computational lexicons, grammars, terminology databases, basic software tools for the acquisition, preparation, collection, management, customization and use of these and other resources.

Mark Liberman and Ron Cole, A Language Resources Primer
http://www.ldc.upenn.edu/myl/LR_background.html

言語資源とは

- 機械可読な言語データおよび言語の記述。音声・言語処理の構築・改良・評価のために用いられる。コーパス、辞書、文法、用語データベースのほか、基本的なツールも含む。

Mark Liberman and Ron Cole, A Language Resources Primer
http://www.ldc.upenn.edu/myl/LR_background.html

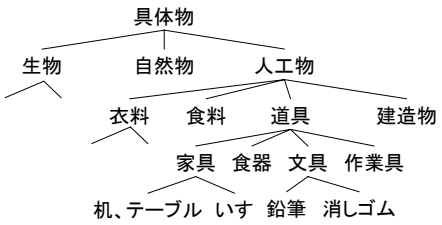
→ 本日の話題
- 言語処理のためのコーパス、辞書について

目次

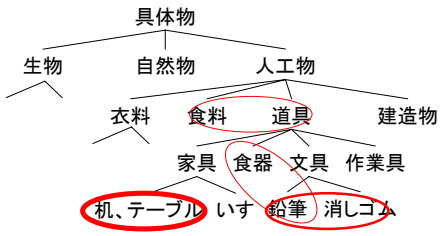
- シソーラス
- （単語）辞書
- タグ付きコーパス
- 生コーパス → 格フレーム自動構築

シソーラス

同義語・類似語・上位語・下位語などを体系的にまとめたもの

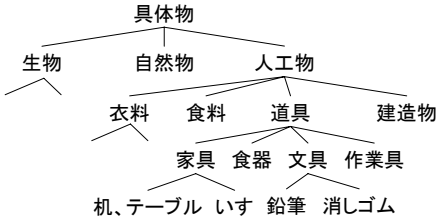


語と語の類似度の計算



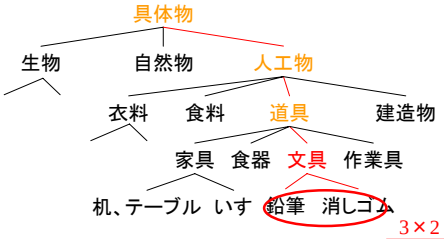
近ければ近いほど似ている

語と語の類似度の計算



$$\text{類似度} = \frac{(\text{共通の親の深さ}) \times 2}{\text{各語の深さの和}}$$

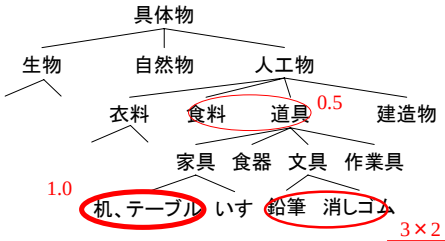
語と語の類似度の計算



$$\frac{3 \times 2}{4+4} = 0.75$$

$$\text{類似度} = \frac{(\text{共通の親の深さ}) \times 2}{\text{各語の深さの和}}$$

語と語の類似度の計算



$$\frac{3 \times 2}{4+4} = 0.75$$

$$\text{類似度} = \frac{(\text{共通の親の深さ}) \times 2}{\text{各語の深さの和}}$$

日本語のシソーラス

- 分類語彙表（国立国語研究所）
  - 約96,000語、6階層（語は最下層のみ）
- 角川類語新辞典
  - 約60,000語、3階層
- EDR概念体系
  - 約400,000概念、10階層程度
- 日本語語彙大系（NTT）
  - 一般名詞：2,715意味属性、12階層
  - 固有名詞：130意味属性、9階層
  - 用言：36意味属性、4階層、6000語

} 約300,000語

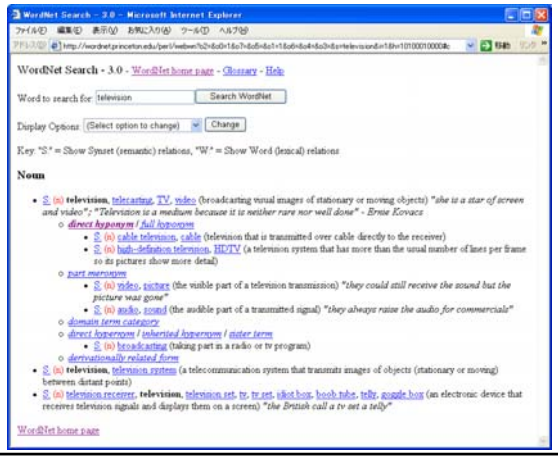
分類語彙表：例

- 原因（げんいん） 体,関係,類,因果
  - 1.1112,04,02,01
- 投票（とうひょう） 体,関係,作用,入り・入れ
  - 1.1532,16,03,02
- 投票（とうひょう） 体,活動,待遇,人事
  - 1.3630,17,01,01
- テレビ（てれび） 体,生産物,機械,電気器具・部品
  - 1.4620,02,01,03

http://www.kokken.go.jp/katsudo/kanko/goihyo/

英語のシソーラス

- Roget's thesaurus
  - Rogetによる最古のシソーラス
  - 1852年リリース: 1.5万語
- WordNet [Fellbaum, 1998]
  - 15万単語を11.5万個の「synset」にまとめたもの
  - synset間は様々な意味関係で結び付けられている
    - hypernym, hyponym, holonym, meronym, …
  - 多言語展開
    - ヨーロッパ系言語, 中国語, 韓国語, …



辞書

- 単語辞書
  - EDR単語辞書
    - 27万語
  - ipadic（茶筌辞書）
    - 24万語
  - JUMAN辞書
    - 基本3万語＋固有名詞3万語

EDR電子化辞書

- 単語辞書
  - 日本語27万語、英語19万語
- 対訳辞書
  - 日英23万語、英日16万語
- 概念辞書
  - 41万概念
- 共起辞書
  - 日本語90万ペア、英語46万ペア
- EDRコーパス（形態素・構文情報付き）
  - 日本語20万文、英語12万文

http://www2.nict.go.jp/r/r312/EDR/J_index.html

JUMANの辞書・文法

文法辞書　　形態素辞書

JUMAN.grammar（品詞分類）
JUMAN.katuyou（活用）
JUMAN.kankei（活用関係）
JUMAN.connect.c（連接規則：250）

ContentW.dic など
自立語：3万語
付属語：1,500語
固有名詞：3万語

コンパイル

jumandic.tab（連接対応表）
jumandic.mat（連接行列）

jumandic.dat（データベース）
jumandic.pat（インデックス）

http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

ContentW.dic（形態素辞書）

```
…
(名詞 (普通名詞 ((読み からくさ)(見出し語 唐草 (から草 1.6) (からくさ 1.6))(意味情報 "代表
      表記:唐草")))
(名詞 (普通名詞 ((読み からくち)(見出し語 辛口 (から口 1.6) (からくち 1.6))(意味情報 "代表
      表記:辛口")))
(副詞 ((読み からくも)(見出し語 辛くも からくも)(意味情報 "代表表記:辛くも")))
(名詞 (普通名詞 ((読み からくり)(見出し語 からくり)(意味情報 "代表表記:からくり"))))
(動詞 ((読み からす)(見出し語 枯らす からす)(活用型 子音動詞サ行)(意味情報 "代表表記:枯
      らす")))
(名詞 (普通名詞 ((読み からす)(見出し語 烏 カラス (からす 1.6))(意味情報 "代表表記:烏"))))
(名詞 (普通名詞 ((読み からだ)(見出し語 身体 体 (からだ 1.6))(意味情報 "代表表記:身体"))))
(名詞 (普通名詞 ((読み からだつき)(見出し語 体付き 体付 体つき (からだつき 1.6))(意味情報
      "代表表記:体付き"))))
(名詞 (普通名詞 ((読み からっかぜ)(見出し語 空っ風 (からっかぜ 1.6))(意味情報 "代表表記:
      空っ風")))
(副詞 ((読み からっきし)(見出し語 からっきし)(意味情報 "代表表記:からっきし")))
…
```

JUMAN辞書に記述されている情報

- 代表表記
- 1文字漢字について、音・訓の区別
  例）字/心 → 音，字/あざ → 訓
- 可能動詞であることと、もとの動詞
  例）見える：可能動詞：見る
- 尊敬動詞・謙譲動詞であることと、もとの動詞
  例）おっしゃる：尊敬動詞：言う
- 動詞が付属動詞として振舞うかどうか
  例）合う

代表表記

```
子ども こども 子ども 名詞 普通名詞 * *
は は は 助詞 副助詞 * *
リンゴ りんご リンゴ 名詞 普通名詞 * *
が が が 助詞 格助詞 * *
すきだ すきだ すきだ 形容詞 * ナ形容詞 基本形
EOS

かぜ かぜ かぜ 名詞 普通名詞 * *
で で で 助詞 格助詞 * *
おくれた おくれた おくれる 動詞 * 母音動詞 タ形
EOS
```

代表表記

```
子ども こども 子ども 名詞 普通名詞 * * "代表表記:子供/こども"
は は は 助詞 副助詞 * *
リンゴ りんご リンゴ 名詞 普通名詞 * * "代表表記:林檎/りんご"
が が が 助詞 格助詞 * *
すきだ すきだ すきだ 形容詞 * ナ形容詞 基本形 "代表表記:好きだ/すきだ"
EOS

かぜ かぜ かぜ 名詞 普通名詞 * * "代表表記:風/かぜ"
@ かぜ かぜ かぜ 名詞 普通名詞 * * "代表表記:風邪/かぜ"
で で で 助詞 格助詞 * *
おくれた おくれた おくれる 動詞 * 母音動詞 タ形 "代表表記:送れる/おくれる 可能動詞:送る"
@ おくれた おくれた おくれる 動詞 * 母音動詞 タ形 "代表表記:遅れる/おくれる"
EOS
```

代表表記（同じ読み）

- 漢字と平仮名、送り仮名
  例）拳銃/けん銃/拳じゅう/けんじゅう　表す/表わす/あらわす　落とす/落す/おとす
- 漢字別表記
  例）狩人/猟人　色取る/彩る　綺麗だ/奇麗だ
- カタカナ表記
  例）大根/だいこん/ダイコン　餃子/ぎょうざ/ギョウザ/ギョーザ　溝/みぞ/ミゾ　眼鏡/めがね/メガネ

代表表記（異なる読み）

- 音便関係
  例）私/わたし/わたくし/あたし　皆/みな/みんな　旅客機/りょかくき/りょかっき　ふわり/ふんわり　とびきり/飛び切り/とびっきり/飛びっ切り
- カタカナ表記のバリエーション
  例）ソフトウエア/ソフトウェア　コンピューター/コンピュータ

タグ付きコーパス

- 形態素・構文情報付きコーパス
  - Penn Treebank
  - 日本語話し言葉コーパス
  - EDRコーパス
  - 京都テキストコーパス
- 述語項構造・共参照情報付きコーパス
  - FrameNet, PropBank, MUC, ACE
  - 京都テキストコーパス
  - NAISTテキストコーパス

構文情報付きコーパス

- 基本的な言語の構造を調査するために重要
- 構文解析器の学習・評価のために必要
  - 英語
    - Penn Treebank
  - 日本語
    - 日本語話し言葉コーパス
    - 京都テキストコーパス

Penn Treebank

[Marcus et al., 1993]

- 品詞・構文タグ付きコーパス
  - – Wall Street Journal（100万単語）
  - – Brown Corpus
  - – Switchboard Corpus（telephone conversation）
    - dysfluencyもタグ付けされている
  - – ATIS（Air Travel Information System）Corpus

Penn Treebankの品詞タグセット

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	(],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Penn Treebankの構文タグ付け例

```
( (S
    (NP-SBJ-1 (NNS Consumers) )
    (VP (MD may)
      (VP (VB want)
        (S
          (NP-SBJ (-NONE- *-1) )
          (VP (TO to)
            (VP (VB move)
              (NP (PRP$ their) (NNS telephones) )
              (ADVP-DIR
                (NP (DT a) (RB little) )
                (RBR closer)
                (PP (TO to)
                  (NP (DT the) (NN TV) (NN set) ))))))))
```

『日本語話し言葉コーパス』

Corpus of Spontaneous Japanese : CSJ

【プロジェクト】「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」

【研究組織】東工大、情報通信研究機構、国語研究所

【開発期間】1999~2003年度（5年間）

【一般公開】2004年度

（http://www.kokken.go.jp/katsudo/kenkyu_jyo/corpus/）

【開発目標】

- 音声認識用言語モデル・音響モデルの学習データ ⇒量
- 自発音声の言語学的研究のリソース ⇒質

研究用情報付与

多 ← 付与情報 → 少

	A:コア独話（177）	B:コア対話・再 朗読（18）	C:準コア(195)	D:コア以外(2906)
音声信号	○	○	○	○
転記テキスト	○	○	○	○
印象評定データ(単独)	○	○	○	○
形態論情報(短単位)	○(手作業)	○(手作業)	○(手作業)	○(自動)
形態論情報(長単位)	○(手作業)	○(手作業)	○(手作業)	○(自動)
分節音情報	○(手作業)	○(手作業)	×	×
韻律情報	○(手作業)	○(手作業)	×	×
節単位情報	○(手作業)	○(自動)	○(自動)	○(自動)
印象評定データ(集合)	○(手作業)	×	×	×
係り受け情報	○(手作業)	×	△(手作業:11)	△(手作業:11)
要約・重要文情報	○(手作業)	×	△(手作業:11)	△(手作業:11)
談話情報	△(手作業: 40)	×	×	×

（テストセット30講演：A（8講演）、C（11講演）、D（11講演））

目的： 話し言葉自動要約技術の開発
基礎データとしての日本語話し言葉コーパスの作成

講演データの収録，転記，形態素/韻律ラベリング（国立国語研究所）

形態素ラベルの自動付与
（661時間分，752万形態素）
・機械学習により高い精度で付与

話し言葉の節単位データ作成（約50万語，199講演）
・文末表現に囚われない，統語的・意味的に有用な単位への分割．
・重要文抽出・係り受け関係付与・談話構造付与のための共通の入力データ．

重要文抽出・文編集
文を単位に10%，50%要約データを作成

係り受け関係
文内での係り受け関係と言い直し関係の付与

談話構造
談話のセグメント分割，階層化，見出し付与（40講演）

NICT自然言語グループでのコーパス作成

2004年春 一般公開

自動要約技術の開発・日本語話し言葉の構文・談話解析

京都テキストコーパス

- 新聞記事に形態素・構文情報を付与
  - 毎日新聞1995年1月1日～17日（2万文）
  - 毎日新聞1995年1月～12月の社説（2万文）
- JUMAN/KNPで解析した結果を人手で修正

- さらに、格関係、名詞間の関係、共参照を付与
  - 後述

CoNLL-X shared task on Multilingual Dependency Parsing

language	treebank	words
Czech	Prague Dependency Treebank	1249
Arabic	Prague Arabic Dependency Treebank	54
Slovene	Slovene Dependency Treebank	29
Danish	Danish Dependency Treebank	94
Swedish	Talbanken05	191
Turkish	MetuSabancı treebank	58
German	TIGER treebank	700
Japanese	Japanese Verbmobil treebank	151
Portuguese	The Bosque part of the Floresta sintá(c)tica	207
Dutch	Alpino treebank	195
Chinese	Sinica treebank	337
Spanish	Cast3LB	89
Bulgarian	BulTreeBank	190

単位：キロ

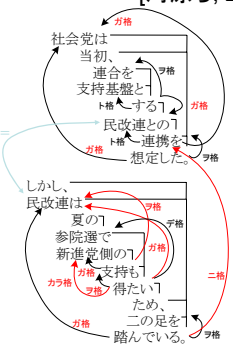
述語項構造・共参照情報付きコーパス

- 日本語
  - 京都テキストコーパス（5,000文）
  - NAISTテキストコーパス（4万文）
- 英語
  - 述語項構造
    - FrameNet
    - PropBank

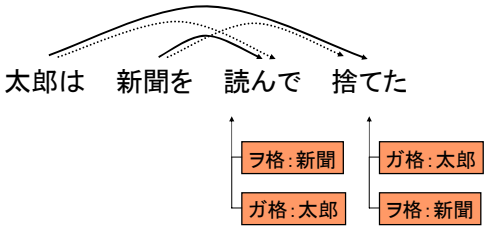
京都テキストコーパス

[河原ら, 2002]

- 形態素情報
- 係り受け情報
- 各種関係情報
  - 格関係
    例：新聞を　読む
  - 名詞間の関係
    例：太郎の　妹
  - 共参照
    例：太郎…　彼は…



格関係



名詞間の関係

太郎は背が低い。しかし、妹は背が高い。

太郎の 妹

共参照

太郎は太っている。彼はいつも何か食べている。

＝：太郎

奈緒美が来た。あの子はいつも長居する。

＝：奈緒美

車の販売台数をみると、自家用車は…

≒：車

上位/下位関係

NAISTテキストコーパス（http://cl.naist.jp/nldata/corpus/）

- 京都テキストコーパス全体を対象に述語項構造と共参照関係をタグ付与したコーパス
  - 述語項構造
    - 京都コーパス4.0と同様に表層格関係でタグ付与
    - 頻出するガ格，ヲ格，ニ格のみを対象に
  - 共参照関係
    - 実体と実体が厳密に同じものを指している場合に付与
    - 現状では，概念間の関係や実体とその属性の間には共参照関係を認めていない
- ダウンロード数: 169

タグの総数

- 京都テキストコーパス3.0 (2,929記事，38,384文)を対象に

	出現箇所	ガ格	ヲ格	ニ格
述語 106,628	同一文節内	177 (0.002)	60 (0.001)	591 (0.027)
	係り関係	44,402 (0.419)	35,882 (0.835)	18,912 (0.879)
	ゼロ照応(文内)	32,270 (0.305)	5,625 (0.131)	1,417 (0.066)
	ゼロ照応(文間)	13,181 (0.124)	1,307 (0.030)	542 (0.025)
	ゼロ照応(文章外)	15,885 (0.150)	96 (0.002)	45 (0.002)
	全体	105,915 (1.000)	42,970 (1.000)	21,507 (1.000)
事態性名詞 28,569	同一文節内	2,195 (0.077)	5,574 (0.506)	846 (0.436)
	係り関係	4,332 (0.152)	2,890 (0.263)	298 (0.154)
	ゼロ照応(文内)	9,222 (0.324)	1,645 (0.149)	586 (0.302)
	ゼロ照応(文間)	5,190 (0.183)	854 (0.078)	201 (0.104)
	ゼロ照応(文章外)	7,525 (0.264)	42 (0.004)	10 (0.005)
	全体	28,464 (1.000)	11,005 (1.000)	1,941 (1.000)
共参照関係		25,764		

タグの一致率

- 作業者2人に30記事を対象に作業を行ってもらった結果
  - 一人の作業結果を正解，もう一人の結果をシステムの出力として再現率，精度を求める

	再現率	精度
述語	0.921 (806/875)	0.944 (806/854)
ガ格	0.823 (683/830)	0.829 (683/824)
ヲ格	0.899 (329/366)	0.954 (329/345)
ニ格	0.724 (105/145)	0.890 (105/118)
事態性名詞	0.965 (247/256)	0.792 (247/312)
ガ格	0.735 (191/260)	0.743 (191/257)
ヲ格	0.827 (86/104)	0.869 (86/99)
ニ格	0.389 (7/18)	0.583 (7/12)
共参照	0.813 (126/155)	0.813 (126/155)

生コーパス

- バランスドコーパス
  - Brown Corpus, LOB Corpus
  - British National Corpus
  - 現代日本語書き言葉均衡コーパス
- その他の大規模生コーパス
  - 新聞記事コーパス
  - ウェブ

Brown Corpus

[Kucera and Francis, 1967]

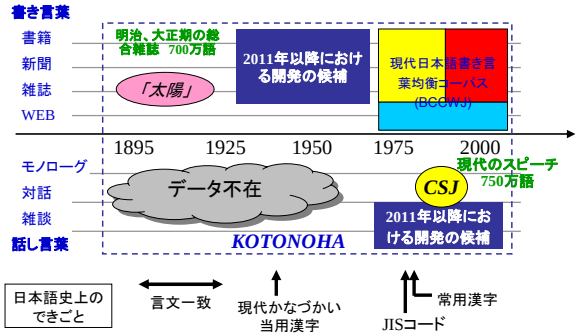
- アメリカの本、新聞、雑誌から収集
  - 15種のジャンル：science fiction, romance and love story, reportage, scientific writing, popular lore, …
- 2000単語×500テキスト＝100万単語
- 英国でも同様の試み（1971～1978年）
  - Lancaster-Oslo/Bergen (LOB) Corpus [Johansson et al., 1978]

British National Corpus

[Leech, 1992]

- 1億単語からなるバランスドコーパス
  - 書き言葉（90%）
    - 新聞、専門雑誌類、学術書、大衆小説、手紙やメモ類、エッセイなど
  - 話し言葉（残り10%）
    - 日常会話、会議、トークショーなど
- 自動で品詞タグが付与されている
  - CLAWS4 [Leech et al., 1994]

国立国語研究所のコーパス開発計画:KOTONOHA



- KOTONOHAとは？
  - 国立国語研究所が開発した、明治から現代までの近現代日本語を対象としたコーパスの集合体。
  - 「日本語話し言葉コーパス」(Corpus of Spontaneous Japanese)2004年公開：学会講演、模擬講演などを中心に現代の独話約750万語を収録。
  - 「太陽コーパス」2005年刊行：明治後期～昭和初期（原文一致完成期）の総合雑誌「太陽」から700万語を収録。研究論文集を同時に刊行。
  - 「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese)：2006年から構築を開始した現代語の書き言葉コーパス。書籍、雑誌、新聞等を中心に1億語を収録予定。完成は2010年。科学研究費特定領域研究に採択。

- 現代日本語書き言葉均衡コーパスの設計方針
- 多様な書き言葉をバランスよく反映したコーパス
  - いわば、現代日本語の縮図となるコーパス
  - 母集団からのランダムサンプリングを行い、統計的な代表性を確保することで母集団における諸特性の分布を高い精度で推測できるようにする。
- 幅広い目的に供するコーパス
  - 日本語学、日本語情報処理、日本語教育、国語教育、国語政策等で活用できるような設計。
  - 過去30年間の言語変化を観察できるようテキストを採録。
  - 文書構造の把握を考慮した設計。
- 公開可能なコーパス
  - 利用許諾が得られたテキストのみを収録。
  - 無償公開（オンライン）と有償公開（オンライン、DVD）を予定。
- 既存のコーパスとのスムーズな接続
  - 日本語話し言葉コーパス(CSJ)で採用した言語解析単位（短単位と長単位）を採用。

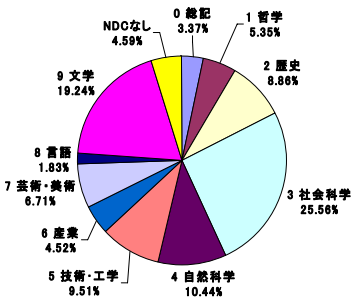
- 現代日本語書き言葉均衡コーパスの構成

生産実態サブコーパス	流通実態サブコーパス
書籍、雑誌、新聞 2001～2005年 約3500万語 固定長＋可変長	書籍 1976～2005年 約3000万語 固定長＋可変長

非母集団サブコーパス

白書、法律、国会会議録、ベストセラー、教科書、WWW上のテキスト …
1976～2005年　2001～2005年
3000万語強
可変長（一部 固定長＋可変長）

生産実態サブコーパスの「書籍」の構成比率



日本十進分類法(NDC)によって文字数の比率を示した。
対象は、2001年～2005年に発行された約32万冊。

大規模生コーパス（英語）

- English Gigaword
  - 4つの新聞記事から収集、18億単語
    - Agence France Press English Service (171M words)
    - Associated Press Worldstream English Service (540M words)
    - The New York Times Newswire Service (914M words)
    - The Xinhua News Agency English Service (132M words)
- Web 1T 5-gram（Google）
  - 1兆単語のウェブコーパスから1～5-gramを抽出
    - Number of tokens: 1,024,908,267,229
    - Number of sentences: 95,119,665,584
    - Number of unigrams: 13,588,391
    - Number of fivegrams: 1,176,470,663

Web 1T 5-gram: 例

3-gram

ceramics collectables collectibles	55
ceramics collectables fine	130
ceramics collected by	52
ceramics collectible pottery	50
ceramics collectibles cooking	45
ceramics collection ,	144
ceramics collection .	247
ceramics collection	120
ceramics collection and	43
ceramics collection at	52
ceramics collection is	68
ceramics collection of	76
ceramics collection |	59
ceramics collections ,	66
ceramics collections .	60
ceramics combined with	46
ceramics come from	69
ceramics comes from	660
ceramics community ,	109
ceramics community .	212

4-gram

serve as the incoming	92
serve as the incubator	99
serve as the independent	794
serve as the index	223
serve as the indication	72
serve as the indicator	120
serve as the indicators	45
serve as the indispensable	111
serve as the indispensible	40
serve as the individual	234
serve as the industrial	52
serve as the industry	607
serve as the info	42
serve as the informal	102
serve as the information	838
serve as the informational	41
serve as the infrastructure	500
serve as the initial	5331
serve as the initiating	125
serve as the initiation	63
serve as the initiator	81
serve as the injector	56
serve as the inlet	41
serve as the inner	87
serve as the input	1323

LDC Top 10

848	TIMIT Acoustic-Phonetic Continuous Speech Corpus
652	CELEX2
380	TIDIGITS
342	ECI Multilingual Text
296	NTIMIT
257	TIPSTER Complete
252	Treebank-3
248	YOHO Speaker Verification
230	Message Understanding Conference (MUC) 7
219	Web 1T 5-gram Version 1

大規模生コーパス（日本語）

- 新聞記事 (毎日, 読売, 朝日, 日本経済, …)
  - 1年分10～20万円程度
  - ドメイン: 新聞

→ ウェブ: 日本語10億ページ以上

- サーチエンジンで検索: 単語の出現ページ数

→ やっぱり言語処理したい

ウェブから収集したコーパス

- Webコーパスの提案 [関口, 山本, 2003]
  - 220MB
- NW1000G-04 [NTCIR-5, Webタスク]
  - 1億ページ
- 大規模Webコーパス [河原, 黒橋, 2006]
  - 5億文
- ウェブアーカイブ [田村, 喜連川, 1999～]
  - 9億ページ以上

Webコーパスの提案

[関口, 山本, 2003]

- HTMLタグや文字種の情報を用いて良質な日本語文を抽出
- 220MBのコーパスを構築（新聞10年分ぐらい?）

NW1000G-04 (NTCIR-5 Webタスク)

[Takaku et al., 2006]

- 1.4TB, 1億ページ
- 2004/01〜2005/01にかけて収集
- 主にjpドメインから収集: 50%がjpドメイン
  - 約60%が日本語ページ

ウェブコーパス

[河原, 黒橋, 2006]

- 1.3TB(圧縮), 4億ページ(各種言語)
- 2004年春に収集に1カ月で収集 (by 東大田浦研)
- 日本語文を抽出
  - 5億文

ウェブコーパスの作成手順

1. クローラによるページ収集(約4億ページ)
2. エンコーディング情報による日本語ページ候補抽出
   - charset属性, perl Encode::guess_encoding()関数
3. 言語情報 日本語ページ判定(2,300万ページ)
   -
4. ペ
5. 日
   - ひらがな, カタカナ,
6. 重複文の削除

→ 約5億日本語文

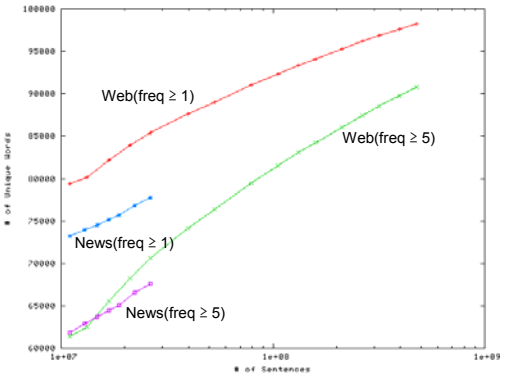
(日本語として妥当な文:995文/1,000文)

```
<HTML>
<HEAD>
<META HTTP-EQUIV="content-type"
        CONTENT="text/h  charset=ISO-2022-JP">
...
```

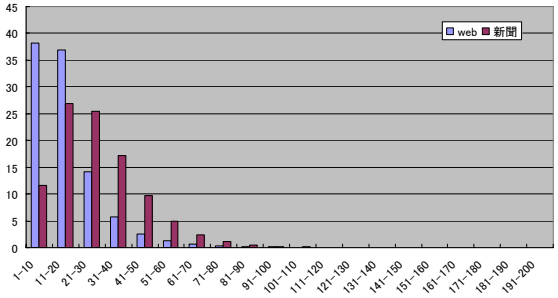
-
- <P>
- <PRE>中の改行

しょうがないので駅のレストランで食事をしようとした所、1日数本しかない山田線の存在に思い当たる。
もれなくプレゼント！
でも僕はTシャツの上に長袖のシャツ。
今回は某アイドルの高橋一也も参加したので客が若い。
団体Aが「まちづくり」をテーマにインターネット上で公開講座を開催しようとしている。
htaccessを置いたとたんそのディレクトリ以下で.
昨年の没後400年祭を機に復元した井戸を紹介する木下さん
恋は、真剣勝負。
ほめ言葉が多くって嬉しいですね。
開校式並びに入学式を挙行、初代校長佐治勝弥、職員10名沖館小学校校舎一部を併用す。
佐治勝弥校長青一中学校長に任命される。
いまだに言うでしょう。
「買いパラ」を見たと伝えれば、お買い上げ合計金額より5%引きいたします。
政治も危機的状況ですし、物資も不足しています。
そういう長期的な存在理由とか、長期的なビジョンとか、何故ここが国のお金で、我々の税金でやらなければいけないのか、その辺を評価する上で何かお考えになられていますでしょうか。
河北郡津幡町南中条・バリアフリー対応の学校案内や生徒会活動の紹介など思いやりのある優しい子に育ってネ。
工学的諸問題に対処する際に必要な, 線形代数・解析・確率・統計などの数学に関する知識を理解できること.
[illegible]

コーパスサイズと異なり単語数の関係



文長(単語数)の分布



新聞とウェブの比較

		新聞(報道)	新聞(社説)	ウェブ
文長		45文字	42文字	28文字
読点率(100文字あたり)		3.2個	3.0個	2.7個
名詞 文字種	漢字	80%	87%	69%
	カタカナ	10%	5%	14%
	ひらがな	4%	5%	7%
	かな漢字	3%	3%	2%

問題

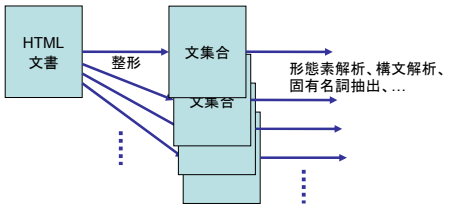
- 特殊文字
  - 参照
    - e.g., «
- 文区切り
  - 「!」の扱い
    - e.g., Yahoo! Japan・－ヘルプ
  - 謎なHTML
    - e.g, 地元活動は、安定的にミニ集会を連日とりお　こなうなど、いわ

配布、そして今後

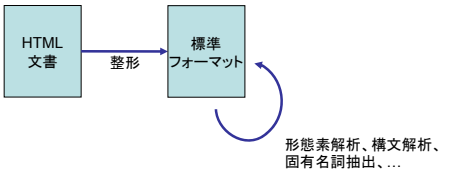
- ウェブコーパスの配布
  - 生, 構文解析済み, 格解析済み(部分的)
  - ハードディスク送付→コピー後返送
- コーパスのさらなる大規模化
  - NW1000G-04(NTCIRデータ)の利用
    - 日本語と判定されたページ: 5,500万ページ
- 標準フォーマット策定 [新里ら, 2007]
- 検索エンジン・API作成 [新里ら, 2007]

Webページ標準フォーマット Webテキスト整形ツール

[新里ら, 2007]



Webページ標準フォーマット Webテキスト整形ツール



```
<html><head><title>小泉総理プロフィール・信念</title>
<meta http-equiv="Content-Type"
  content="text/html;charset=Shift_JIS">
</head>
<body bgcolor="#ffffff" text="#000000">
<center>
<table cellpadding="0" cellspacing="0" width="610">
<tbody><tr>
<td valign="top" width="172">
<img src="sample_files/0_001.jpg"
  alt="無信不立" height="160" width="160">
</td>
<td valign="top" width="426"><b><font color="#0066ff">
▼座右の銘</font></b>
<br>
<img src="sample_files/1px_spacer.gif" alt="" height="5" width="426">
<br>
  小泉総理の好きな格言のひとつに「無信不立(信無くば立たず)」があります。論語の下篇「顔淵」の言葉で、弟子の子貢（しこう）が政治について尋ねたところ、孔子は「食料を十分にし軍備を十分にして、人民には信頼を持たせることだ」と答えました。<br>
</td>
</tr>
</tbody></table>
</center>
</body></html>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat
  Url="http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html"
  OriginalEncoding="Shift_JIS" Time="2006-08-14 19:48:51">
  <Text Type="dafault">
    <S Id="1" Length="70" Offset="525">
      <RawString>小泉総理の好きな格言のひとつに「無信不立」があります。
      </RawString>
    </S>
    <S Id="2" Length="160" Offset="595">
      <RawString>
        論語の下篇「顔淵」の言葉で、弟子の子貢（しこう）が政治について尋ねたところ、
孔子は「食料を十分にし軍備を十分にして、人民には信頼を持たせることだ」と答えました。</RawString>
    </S>
  </Text>
</StandardFormat>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat
  Url="http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html"
  OriginalEncoding="Shift_JIS" Time="2006-08-14 19:48:51">
  <Text Type="dafault">
    <S Id="1" Length="70" Offset="525">
      <RawString>小泉総理の好きな格言のひとつに「無信不立」があります。
      </RawString>
      <Annotation Scheme="KNP">
        <![CDATA[# S-ID:1 KNP:2006/08/10
* 1D <文頭><サ変><人名><助詞><連体修飾><体言><係:ノ格><区切:0-
4><RID:1056>
小泉 こいずみ 小泉 名詞 6 人名 5 * 0 * 0 NIL <文頭><漢字><かな漢字><名詞相当
語><自立><タグ単位始><文節始><固有キー>
 ...中略...
ます ます ます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 NIL <
表現文末><かな漢字><ひらがな><活用語><付属><非独立無意味接尾辞>
。。。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS]]>
      </Annotation>
    </S>
    <S Id="2" Length="160" Offset="595">
  …
```

ディープNLPサーチエンジン基盤 TSUBAKI
[新里ら, 2007]

- 日本語ウェブ文書5,000万件を対象とした開放型検索エンジン基盤
  - 高度ウェブ処理用標準フォーマットによりウェブ文書を管理
  - 構造的言語処理によるインデクシング
  - 無制限に利用可能なAPI
  - 透明性・再現性のある検索結果

格フレーム

?

クロールで 泳いでいる女の子を見た
望遠鏡で　泳いでいる女の子を見た

格フレーム

{人,子,…}が {クロール,平泳ぎ,…}で {海,大海,…}を**泳ぐ**	{人,者,…}が {双眼鏡,望遠鏡,…}で {姿,人,…}を**見る**

格フレーム（英語）

- subcategorization framesの学習
  [Brent, 1993] [Ushioda et al., 1993]
  [Manning, 1993] [Briscoe and Carroll, 1997]…
- FrameNet [Baker et al., 1998]
- PropBank [Palmer et al., 2005]

subcategorization framesの学習

- 動詞とフレームの組をコーパスから抽出し、その組が有意かどうかを頻度に基づく仮説検定により判定

  e.g., She greeted me.
  - NP(subj) greet NP(obj)

  e.g., She gave him a book.
  - NP(subj) give NP(obj) NP(obj)

subcategorization framesの学習

	# of SCFs	# of verbs	corpus size	Acc
[Brent, 1993]	6	63	1.2M	85a
[Ushioda et al., 1993]	6	33	0.3M	86b
[Manning, 1993]	19	200	4.1M	82b
[Ersan & Charniak, 1996]	16	30	36M	70a
[Caroll & Rooth, 1998]	15	100	30M	77a
[Briscoe & Caroll, 1997]	161	7	1.2M	81b
[Sarkar & Zeman, 2000]	137	914	0.3M	88b

a: type F-measure, b: token recall

FrameNet

[Baker et al., 1998]

- frame-by-frame basis
  - 625個のフレーム
  - 6,100個の語（動詞、形容詞、名詞）が関連付けられている
  - フレームごとに、項（frame element）が定義されている（coreとnon-coreの2種類）
- British National Corpusから抽出した文にタグ付け
  - 135,000文にタグ付け

FrameNet：例

- Frame: Commerce
  - Buyer
  - Goods
  - Seller
  - Payment
  - Rate/Unit, …

[$_{\text{Buyer}}$ Chuck] bought [$_{\text{Goods}}$ a car] [$_{\text{Seller}}$ from Jerry] [$_{\text{Payment}}$ for $1000].

[$_{\text{Seller}}$ Jerry] sold [$_{\text{Goods}}$ a car] [$_{\text{Buyer}}$ to Chuck] [$_{\text{Payment}}$ for $1000].

PropBank

[Palmer et al., 2005]

- verb-by-verb basis
  - 3,300個の動詞
  - 4,500個のフレームが関連付けられている
  - 項は次の2種類
    - role：Arg0, Arg1, Arg2, …（フレームごと）
    - 任意的なrole：ArgM-LOC, ArgM-TMP, …（共通）
- Penn Treebankの文に対してタグ付け
  - 85,000文にタグ付け

PropBank：例

- Frameset: buy.01
  - Arg0: buyer
  - Arg1: things bought
  - Arg2: seller
  - Arg3: price paid
  - Arg4: benefactive
- Frameset: sell.01
  - Arg0: seller
  - Arg1: things sold
  - Arg2: buyer
  - Arg3: price paid
  - Arg4: benefactive

[$_{\text{Arg0}}$ Chuck] bought [$_{\text{Arg1}}$ a car] [$_{\text{Arg2}}$ from Jerry] [$_{\text{Arg3}}$ for $1000].

[$_{\text{Arg0}}$ Jerry] sold [$_{\text{Arg1}}$ a car] [$_{\text{Arg2}}$ to Chuck] [$_{\text{Arg3}}$ for $1000].

人手による格フレーム（日本語）

- IPAL
  - 約1,000語の和語動詞、形容詞について、とりうる格と用例を記述
- EDR動詞共起パターン副辞書
  - 約5,000語の動詞について、とりうる格と概念を記述
- NTT構文体系
  - 約6,000語の動詞,形容詞について、とりうる格と意味属性を記述
- その他関連研究
  - 語彙概念構造（LCS）辞書
  - JCASRプロジェクト

人手による格フレーム（日本語）：例

- IPAL
  - N1がN2に/と会う
    - N1: 彼, N2: 友達、恩師
  - N1がN2に会う
    - N1: 彼／会社／船, N2: 嵐、雨、夕立、火事／災難、怖い目、ひどい目、事故、盗難、反対、抵抗
- NTT構文体系
  - N1がN2に会う
    - N1: 主体, N2: 抽象
  - N1がN2に/と会う
    - N1: 主体/動物, N2: 主体/動物

語彙概念構造(LCS)辞書

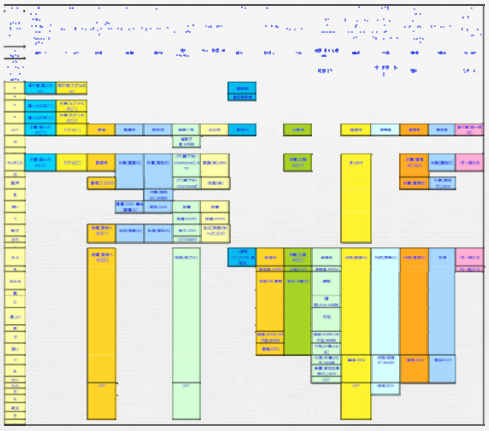
- 研究組織(科研費による研究プロジェクト)
  - 奈良先端大，岡山大，名古屋大
- 目標
  - 日本語の動詞間の含意関係を項を含めて記述
  - 言語処理に有効な辞書の構築
- 現状
  - NTTコミュニケーション科学基礎研究所のLexeedの語義単位に対して付与
  - 約5700語義について人手で付与
  - 語義，語釈，例文，必須格，深層情報，動作主性，状態変化性，状態性，言語テスト結果(一部)
  - excel の表形式で記述

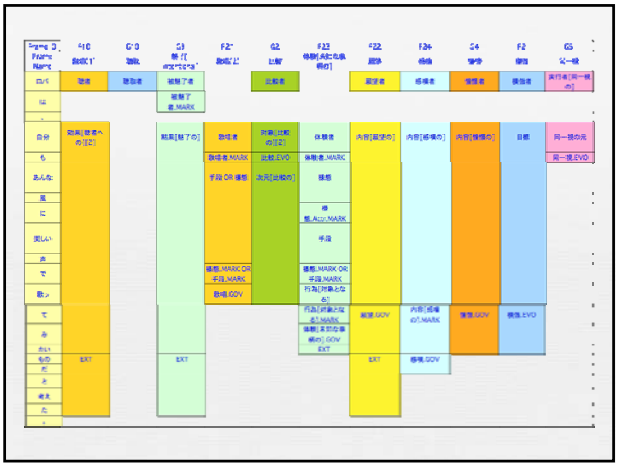
LCS辞書(一部)

語の情報			タイプ	意味特性と検査						
語	必須格	例文		動作主	検査結果 a1	検査結果 a2	状態性	検査結果 s1	検査結果 s2	…
入籍する	ガ，ト	彼が入籍する	関係の変化	y	N/A	y	n	n	結果残存	
登録する	ガ，ヲ，ニ	彼の電話番号を登録した	位置関係の変化	y	N/A	Y	n	n	結果残存	

JCASR プロジェクト

- JCASR = Japanese Corpus Annotated for Semantic Roles
- 複層意味フレーム分析 (MSFA) を使ってコーパスに意味フレーム (=状況)とフレーム要素 (=状況レベルの粒度の細かい “意味役割”) の情報をタグづけしたもの
  - Berkeley FrameNet (Baker et al. 1998) から派生





フレームの例

- <歌唱>フレームの要素
  - 歌唱者，聴者[+potential]，効果[+potential]など
- <比較>フレームの要素
  - <比較者>，<比較の対象>，<比較の基準>など
- <模倣>フレームの要素
  - <模倣者>，<模倣の対象>，<目標>など
- ほとんどのフレームは FrameNet から独立して探索的に認定されている

方向性の比較

	大規模な辞書構築	注釈指向	コーパス規模	不連続な喚起パターンの記述	記述粒度
JCASR	No	Yes	Small	Yes	Very Fine
Berkeley FrameNet	Yes	No	?	No	Fine
Penn PropBank	?	Yes	Large	No	Rough

格フレームの自動構築（日本語）

- 構文情報付きコーパスからの格フレーム自動構築
  - [Haruno, 1995]
    - 11個の動詞について、人手で構文解析した新聞記事1年分を用い、情報量尺度に基づき格フレームを構築
  - [Utsuro et al., 1997]
    - 最大エントロピー法によってEDRコーパスから格フレームの確率モデルを学習
- 生コーパスからの格フレーム自動構築
  - [河原，黒橋，2005]

日本語格フレーム自動構築

[河原，黒橋，2005]

人手では難しい

- 「数万語の自立語 × 多義性」を幅広くカバーすることが難しい
- 「がをに」の単純なパターンばかりではない
  e.g. ～によって，～について，二重主語構文
- 新語・専門用語
  e.g. ググる，サチる

構築する格フレームの例

用言	格	用例
買収する:1	ガ格	企業,会社,銀行,…
	ヲ格	会社,企業,工場,球団,ホテル…
買収する:2	ガ格	企業,ソフトバンク,キャピタル,JT,…
	ノ/カラ格	ナビスコ,会社,薬品,企業…
	ヲ格	部門,事業
買収する:3	ガ格	政権,顧問,～者
	ヲ格	議員,役人
…		

コーパスから自動構築する方法

- 構文情報に基づいて格フレームを自動構築
  e.g. 数人の男たちが大きなボートに荷物を積んでいる。

コーパスから自動構築する方法

- 構文情報に基づいて格フレームを自動構築
  e.g. 数人の男たちが大きなボートに荷物を積んでいる。

コーパスから自動構築する方法

- 構文情報に基づいて格フレームを自動構築
  e.g. 数人の男たちが大きなボートに荷物を積んでいる。
  – 構文情報が付与されたコーパスから構築
    - EDRコーパス（20万文）
    - 京都テキストコーパス（4万文）

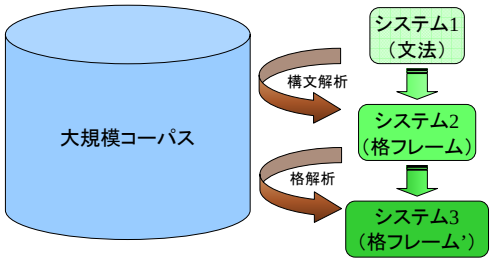
    ⇒量的に不十分
  – 構文情報が付与されていないコーパス（生テキスト）から構築

    ⇒大量の生テキストを自動解析し、その解析結果の信頼度の高い部分から格フレームを構築できる

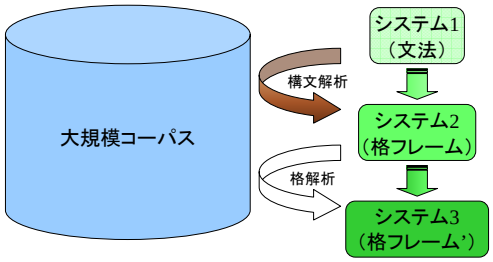
本研究のアプローチ

- 大量の生テキストから格フレームを自動構築
  – ウェブコーパス（5億文）
- 自動解析結果に含まれる構文的曖昧性に対処
  – 信頼度の高い部分のみを用いる（精度：98.3%）
- 漸次的格フレーム構築
  – 様々な関係を段階的に得る

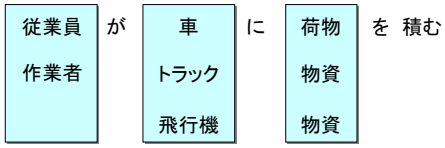
格フレームの自動構築



格フレームの自動構築



格フレームの自動構築(1)



従業員 作業者 が

車 トラック 飛行機 に

荷物 物資 物資 を 積む

格フレーム自動構築の問題

- 構文的曖昧性
  - 生コーパスを自動解析するので、解析誤りが含まれる
- 意味的曖昧性
  - 同じ表記の用言でも複数の意味をもつ
    - 荷物を積む
    - 経験を積む

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

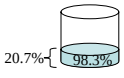
- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …

問題1：構文的曖昧性

- 確実な部分だけを集める
  - ひどい 風邪を 引いたので、家で 寝ていた。
  - 被害者を 早く 救い出すべきだ。
  - 火の 回りが 早く 救いだせなかった。
  - その 議員は 法案を 提出した。
  - 議員が 提出している 法案は …



係り受けの20.7%について
98.3%の精度で抽出

問題2：意味的曖昧性

従業員	が	車	に	荷物	を 積む
作業員	が			荷物	を 積む
		飛行機	に	荷物	を 積む
彼	が	車	に	物資	を 積む
		トラック	に	物資	を 積む
従業員	が			経験	を 積む
選手	が			経験	を 積む

直前格ごとにまとめる

従業員	が	車	に	荷物	を 積む
作業員	が			荷物	を 積む
		飛行機	に	荷物	を 積む
彼	が	車	に	物資	を 積む
		トラック	に	物資	を 積む
従業員	が			経験	を 積む
選手	が			経験	を 積む

クラスタリング

従業員	が	車	に	荷物	を 積む
作業員	が			荷物	を 積む
		飛行機	に	荷物	を 積む
彼	が	車	に	物資	を 積む
		トラック	に	物資	を 積む
従業員	が			経験	を 積む
選手	が			経験	を 積む

格フレーム間の類似度

＝ 格の一致度 × 用例の類似度

[作業員]が$_1$ [倉庫]で$_2$ [車,飛行機]に$_{3\ 2}$ [荷物]を$_8$ 積む

[従業員,彼]が$_{1\ 1}$ [トラック]に$_5$ [物資]を$_{10}$ 積む

格フレーム間の類似度

＝ 格の一致度 × 用例の類似度

[作業員]が$_1$ [倉庫]で$_2$ [車,飛行機]に$_{3\ 2}$ [荷物]を$_8$ 積む

[従業員,彼]が$_{1\ 1}$ [トラック]に$_5$ [物資]を$_{10}$ 積む

$$\text{格の一致度} = \sqrt{\frac{1+(3+2)+8}{1+2+(3+2)+8} \times \frac{(1+1)+5+10}{(1+1)+5+10}}$$

格フレーム間の類似度

＝ 格の一致度 × 用例の類似度

[作業員]が$_1$ [倉庫]で$_2$ [車,飛行機]に$_{3\ 2}$ [荷物]を$_8$ 積む

[従業員,彼]が$_{1\ 1}$ [トラック]に$_5$ [物資]を$_{10}$ 積む

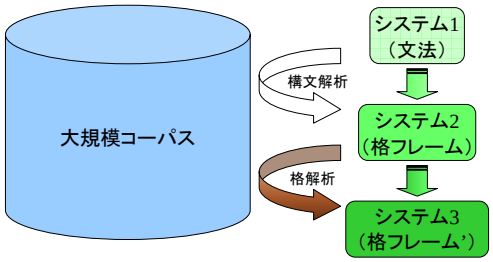
1.0, 0.73 / 1.0, 0.85 / 1.0

$$\frac{1.0\times1+1.0\times1+0.73\times1}{1+1+1} = 0.91 \qquad \frac{1.0\times3+0.85\times2+0.85\times5}{3+2+5} = 0.90 \qquad \frac{1.0\times8+1.0\times10}{8+10} = 1.0$$

$$\text{格の一致度} = \sqrt{\frac{1+(3+2)+8}{1+2+(3+2)+8} \times \frac{(1+1)+5+10}{(1+1)+5+10}}$$

$$\text{用例の類似度} = \frac{\sqrt{1+1}\times\sqrt{1}\times0.91+\sqrt{3+2}\times\sqrt{5}\times0.90+\sqrt{8}\times\sqrt{10}\times1.0}{\sqrt{1+1}\times\sqrt{1}+\sqrt{3+2}\times\sqrt{5}+\sqrt{8}\times\sqrt{10}}$$

格フレームの自動構築



格フレームの自動構築(2)

- この法案はA議員が提出した。
  [議員，委員．．]が [法律，案．．]を 提出する
- その車はエンジンがよい。 ⇒ 二重主語構文
  [エンジン]が よい
- 法案を提出した議員．．．
  [議員，委員．．]が [法律，案．．]を 提出する
- 法案を提出する見通し．．．⇒ 外の関係
  [議員，委員．．]が [法律，案．．]を 提出する

格フレームの自動構築(2)

- 外の関係の整理
  - 業務を営む免許 ⇒ 「業務を営む」固有の外の関係
  - 業務を営む見通し
  - 法案を提出する見通し ⇒ 一般的な外の関係
  - 衆院を通過する見通し
  - ．．．

可能性、必要、結果、方針、ケース、考え、予定、見込み．．．

格フレームの自動構築(2)

- 格変化を扱う
  - 同じ意味を表すためにも、さまざまな格が用いられる
    - …の経緯の説明を求めた
    - …の経緯について説明を求めた

格フレームの格間の類似度をとって、似ている格をマージする

格フレームの自動構築(2)

「説明を 求める」に対応する格フレーム

求める	
ガ格	委員会,団,氏,…
ヲ格	説明,釈明
ニ格	政府,学会,社長,…
について	経緯,実態,状況,…
ノ格	経緯,理由,内容,…

類似度 0.94

→ 格をマージする

必須格の選択

- ガ格については、すべての用言がとると考え、用例が1つでもあれば選択する
- ガ2格は、格フレームに必須と考えられるので、用例が1つでもあれば選択する
- 外の関係は、必須的ではないが格フレームに固有と考えられるので、用例が1つでもあれば選択する
- 直前格の用例頻度 mf に対して、用例数が $2\sqrt{mf}$ より多い格を選択する

高性能計算機グリッドを用いた格フレーム構築

- 形態素・構文解析（JUMAN/KNP）
  解析スピード：20文/s

→ 解析に10ヶ月；クラスタリングに7年

- 350CPUの高性能計算機グリッドを使用（by グリッド環境用シェルGXP）

→ 解析に1日；クラスタリングに7日

構築した格フレームの例

用言	格	用例
焼く(1)	ガ格	私:18, 人:15, 職人:10, …
	ヲ格	パン:2484, 肉:1521, ケーキ:1283, …
	デ格	オーブン:1630, フライパン:1311, …
焼く(2)	ガ格	先生:3, 政府:3, 人:3, …
	ヲ格	手:2950
	ニ格	攻撃:18, 行動:15, 息子:15, …
焼く(3)	ガ格	メーカー:1, ディストリビューター:1, …
	ヲ格	データ:178, ファイル:107, コピー:9, …
	ニ格	R:1583, CD:664, CDR:3, …
…		

構築した格フレームの例

用言	格	用例
泳ぐ	ガ格	イルカ:142, 生:50, 魚:28, …
	ヲ格	海:1188, 水中:281, 海中:101, …
	デ格	クロール:86, 平泳ぎ:49, 泳法:24, …
磨く	ガ格	私:4, 男性:4, 人:4, おれ:4, …
	ヲ格	歯:5959, 奥歯:27, 前歯:12
	デ格	ブラシ:38, 塩:13, 粉:12, …
録画する	ガ格	旦那:4, 妹:2, 知人:2, 友人:2, …
	ヲ格	番組:1435, 放送:521, 特番:26, …
	ニ格	ビデオ:3753, ディスク:256, …

格フレームの規模

5億文

2600万文
(26年分)

	Web	新聞
用言数	89243	18246
(内訳) 動詞	40860	12641
形容詞	4121	991
名詞+判定詞	44262	4614
用言あたりの平均格フレーム数	34.3	17.5
格フレームあたりの格の平均数	3.2	2.4
格あたりの平均用例数	72.9	29.8
格あたりの平均異なり用例数	26.9	4.2

Web格フレームのカバレッジ

- テスト文に対して対応する格フレームができているかどうかを調査
  - Webのテキスト 650文
  - 国語辞典定義文 1,000文

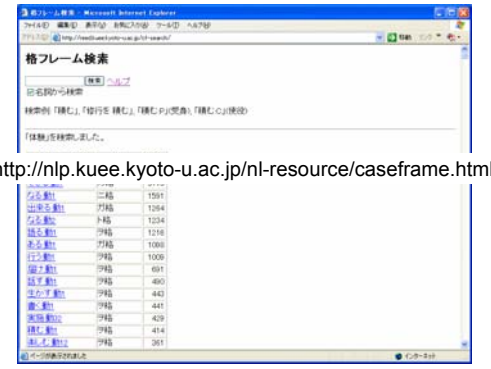
対象	CF	一致	一致+類似
Webテキスト	WebCF	0.628 (510/812)	0.777 (631/812)
	新聞CF	0.459 (373/812)	0.648 (526/812)
定義文	WebCF	0.599 (269/449)	0.835 (375/449)
	新聞CF	0.481 (216/449)	0.726 (326/449)

一致: 直前格要素が一致している格フレームが存在する
類似: 直前格要素が類似している格フレームが存在する

格フレームの人手評価

- 誤り原因
  - シソーラスの問題
    - {感, 下落}ガ {景気, 心理, マインド}ヲ 冷やす
  - イディオムの問題
    - {<数量>人, 少年,..}ガ {探し, 調整,..}ニ {頭, 首,..}ノ {骨, ろっ骨,..}ヲ 折る

格フレーム検索



http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html

構文・格解析の統合的確率モデル

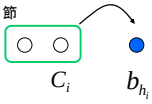
入力文 S，構文構造 T，格構造 L

$$(T_{best}, L_{best}) = \arg\max_{(T,L)} P(T, L \mid S)$$

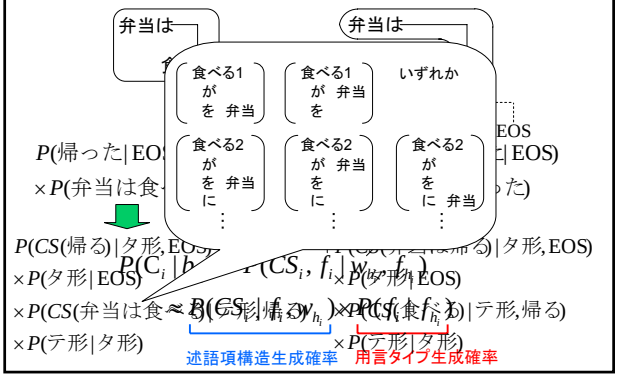
$$= \arg\max_{(T,L)} \frac{P(T, L, S)}{P(S)}$$

$$= \arg\max_{(T,L)} \underbrace{P(T, L, S)}_{\text{def}\,\|}$$

$$\prod_{C_i \in T} P(C_i \mid b_{h_i})$$



構文・格解析の統合的確率モデル



実験

- 各パラメータ推定のためのリソース
  - 格フレーム：Webテキスト5億文から自動構築
  - 構文・格解析済みデータ：Webテキスト1億文の解析結果
- Webテキスト759文を用いた評価実験
  - 構文解析の実験
    - 文節ごとに係り先を評価
      （文末から2番目の文節を除く）
  - 並列構造検出の実験

実験結果

構文解析実験の結果

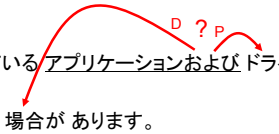
	構文	構文+格	構文+格+並列
すべて	0.864 (3833/4436)	0.868 (3852/4436)	**0.878** (3893/4436)
体言→用言	0.850 (1637/1926)	0.864 (1664/1926)	**0.874** (1684/1926)
体言→体言	0.908 (1032/1136)	0.906 (1029/1136)	**0.913** (1037/1136)
用言→用言	0.800 (654/817)	0.792 (647/817)	**0.807** (659/817)
用言→体言	0.916 (510/557)	0.919 (512/557)	**0.921** (513/557)

並列構造検出の結果

	ベースライン	提案手法
適合率	0.796 (366/460)	**0.830** (361/435)
再現率	**0.819** (366/447)	0.808 (361/447)
F	0.807	**0.819**

正解例

水が 高い ところから 低い ところへ 流れる。

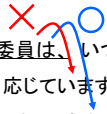


プレインストールされている アプリケーションおよび ドライバとの

競合により 動作しない 場合が あります。

誤り原因

- 意味的には係り先はどちらでもよい場合



行政相談委員は、 いつでも 自宅で みなさんからの

ご相談に 応じていますが、 この期間中は 次の ところで

行政相談所を 開きます。

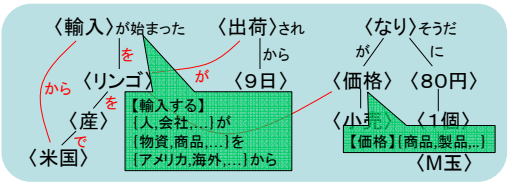
ポイント

- タグ付きコーパスから学習できることの明確化
  - 適正な利用
- 大規模生コーパスの利用
  - 量がすべてというところで利用
  - ＋言語の精緻な扱い
  - すべての知識がウェブに書いてある？

まとめと今後の展開

- まとめ
  - タグ付きコーパスの作成
  - 超大規模生コーパスの構築
  - 大規模格フレームの自動構築
- 今後の展開

  文章理解の高精度化：文脈の表現



その他の言語資源

- 意味情報付きコーパス
  - 語義曖昧性解消（Senseval, SemEval）
- 意見、評判情報付きコーパス
  - 情報の取捨選択、信頼性評価
- 対訳辞書・コーパス
  - 使える機械翻訳に向けて
- 異なるドメインの言語資源
  - GENIA（bioNLP）

お世話になった方々

- 飯田 龍氏（奈良先端大）
- 乾 健太郎氏（奈良先端大）
- 内元 清貴氏（NICT）
- 黒田 航氏（NICT）
- 黒橋 禎夫氏（京都大）
- 新里 圭司氏（京都大）
- 竹内 孔一氏（岡山大）
- 山崎 誠氏（国研）

更新した配布資料：

http://www2.nict.go.jp/x/x161/member/kawahara/NLP2007tut.pdf