クレジット:
UTokyo Online Education 知の構造化論　2020　美馬　秀樹

東京大学 THE UNIVERSITY OF TOKYO

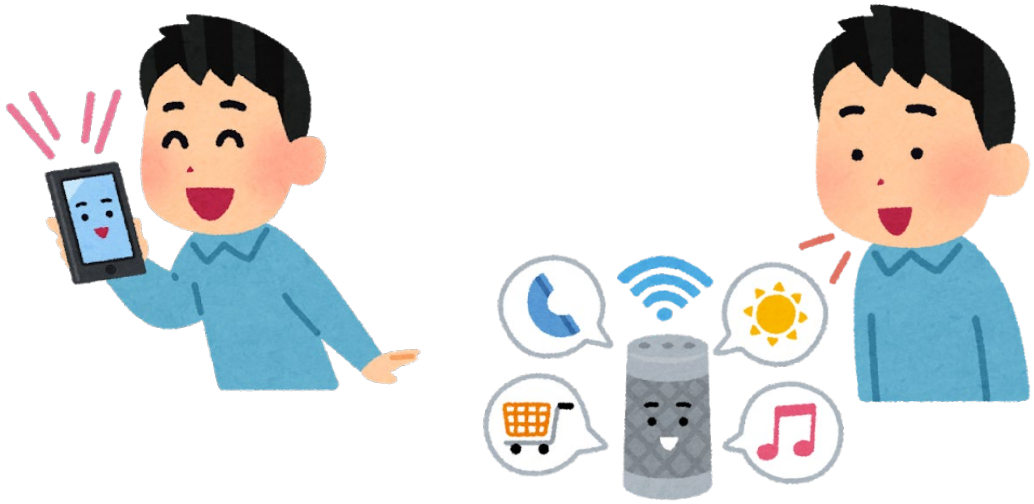
CC BY NC ND

自然言語処理の基礎
(+人工知能、機械学習との関連)

東京大学 工学系研究科／大学総合教育研究センター
美馬秀樹

自然言語処理で何ができる？

- Siri(Apple), Googleアシスタント, Alexa(Amazon)
  - スマートフォンやスマートスピーカーに話しかけ、会話、操作を行う
  - 人間の言葉を聞き、理解し、応答をする



イラスト：いらすとや

自然言語処理と人工知能

- 「本郷三丁目から駒場東大前までの電車」

現在地	行き先	手段
本郷三丁目	東大駒場前	電車

形態素解析

- 文を形態素（単語）に分け、品詞等の属性情報を同定する
  例：「本郷三丁目から駒場東大前までの電車」

形態素の表示				
表記	よみ	品詞	基本形	全情報
本郷三丁目	ほんごうさんちょうめ	名詞	本郷三丁目	名詞,地名,*,本郷三丁目,ほんごうさんちょうめ,本郷三丁目
から	から	助詞	から	助詞,格助詞,*,から,から,から
駒場東大前	こまばとうだいまえ	名詞	駒場東大前	名詞,地名,*,駒場東大前,こまばとうだいまえ,駒場東大前
まで	まで	助詞	まで	助詞,副助詞,*,まで,まで,まで
の	の	助詞	の	助詞,助詞連体化,*,の,の,の
電車	でんしゃ	名詞	電車	名詞,名詞,*,電車,でんしゃ,電車

構文解析の例

日本語パーザ オンラインデモ on Enju 2.4

東京大学 知の構造化センター美馬研究室で研究を進めている日本語CCGパーザのオンラインデモサイトです。

文を入力するとXML形式で解析結果を見ることができます。

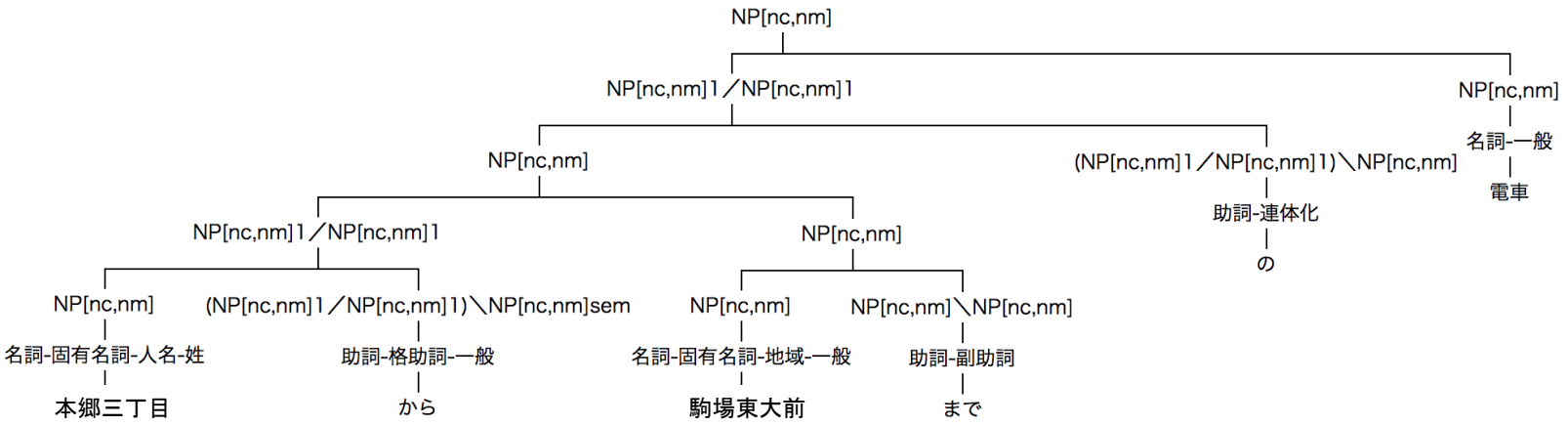
本郷三丁目から駒場東大前までの電車　解析　クリア

例文

- 気象庁は強風や高波に注意するよう呼びかけています。
- 前半のショートプログラムで首位に立った浅田選手は、２０日に行われたフリーで出場した１０人の最後に滑りました。

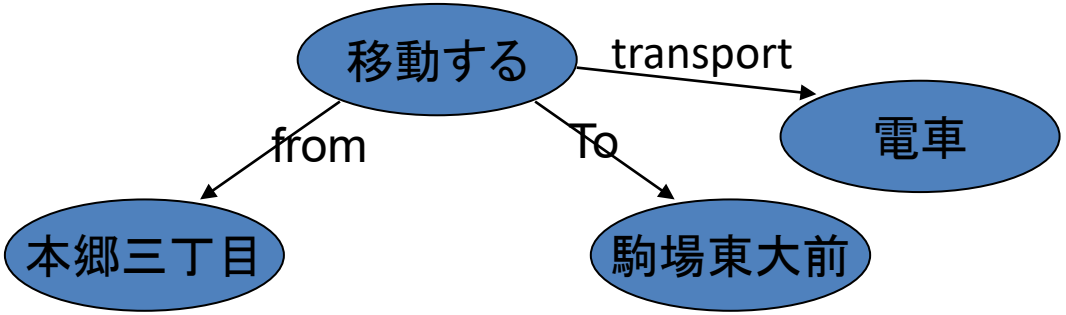
Parse status: success

Click to change the view.



意味解析

- 文の構造から意味表現を同定する（構造化）
- 文の意味とは
  - 意味ネットワーク



  - フレーム
    - (移動する
      (from X?)
      (to Y?)
      (transport Z?)
)

実体化 →

(移動する-001
(from 本郷三丁目-001)
(to 駒場東大前-001)
(transport 電車-002)
)

構造化

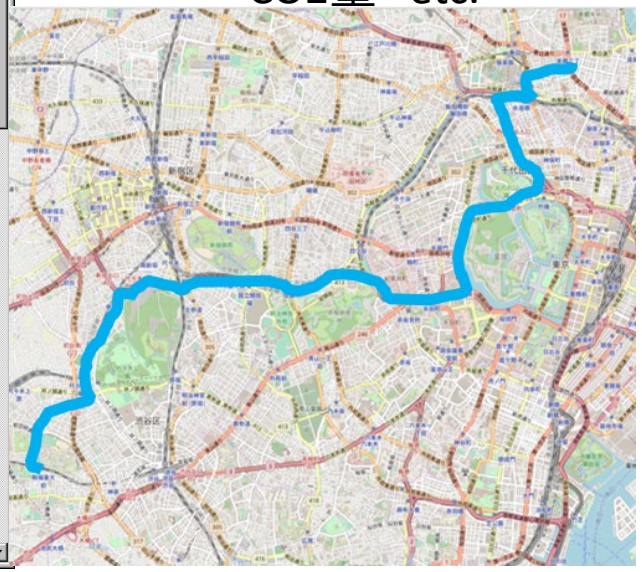
現在地	行き先	交通手段
本郷三丁目-001	東大駒場前-001	電車-002

実検索

最適化の例(本郷三丁目～駒場東大前)

- 経路探索
  可能なルートから最適なものを選ぶ
  - 評価軸
    - 運賃
    - 時間
    - 乗換回数
    - CO2量　etc.

NAVITIME
https://www.navitime.co.jp/

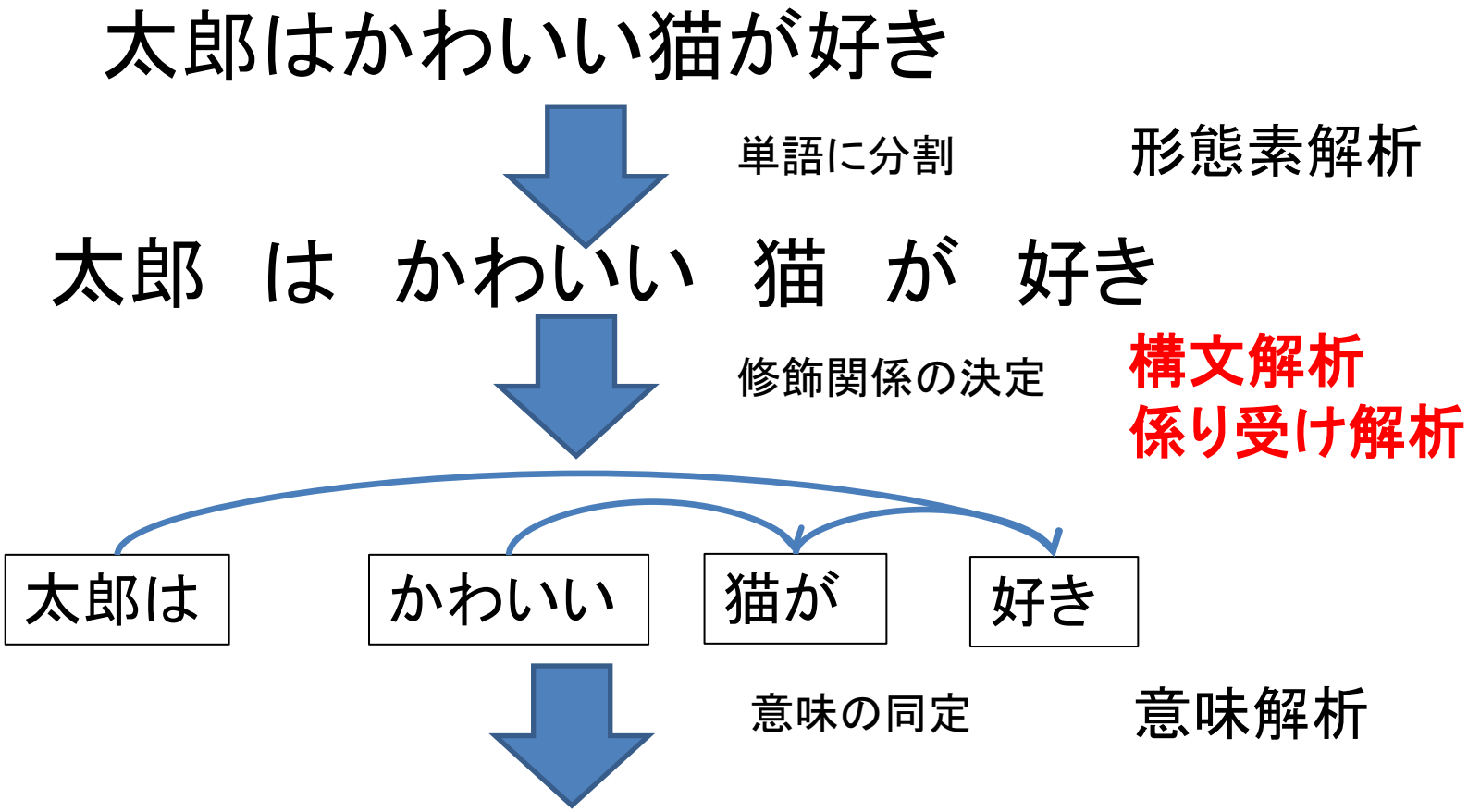


(c)OpenStreetMap contributors

「電車で駒場東大前まで本郷三丁目から行きたい」→　同じ結果

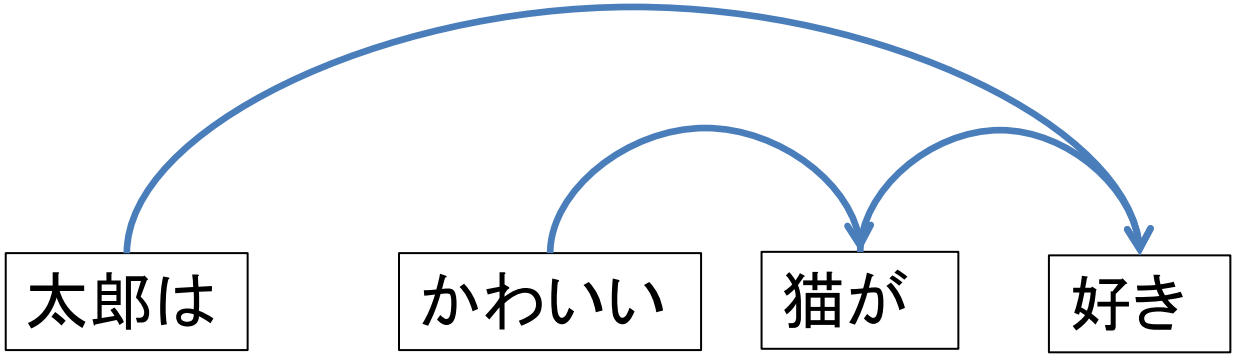
構文解析

自然言語処理の基礎技術



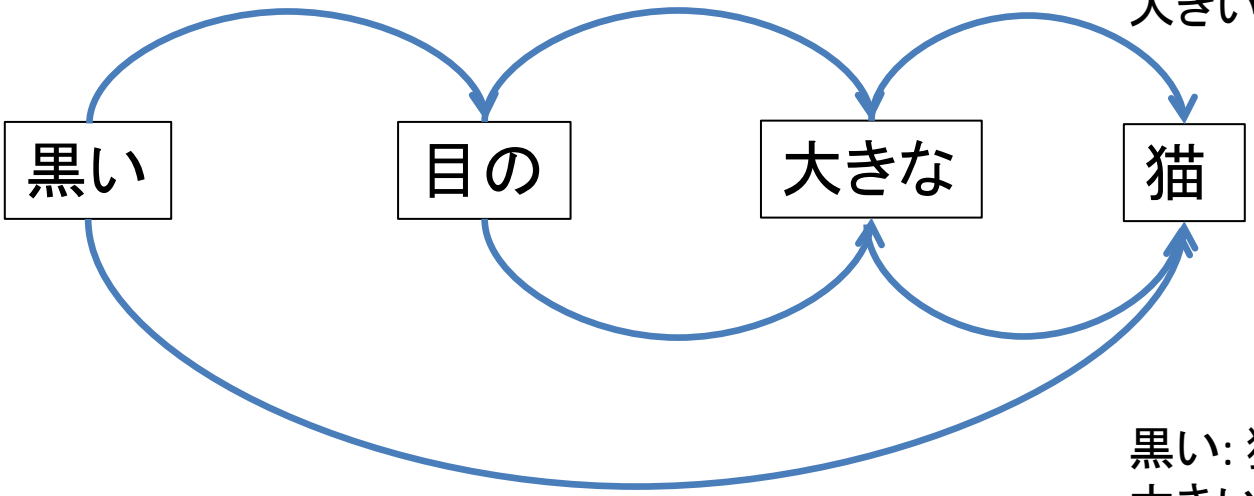
構文解析(係り受け解析)

- 文の統語的構造(係り受け)を同定する
  - 係り受け関係 : 分節間の修飾関係
- 例:太郎はかわいい猫が好き



係り受け解析

- 黒い目の大きな猫



黒い目の大きな、猫

黒い: 目
大きい: 目

黒い: 猫
大きい: 目

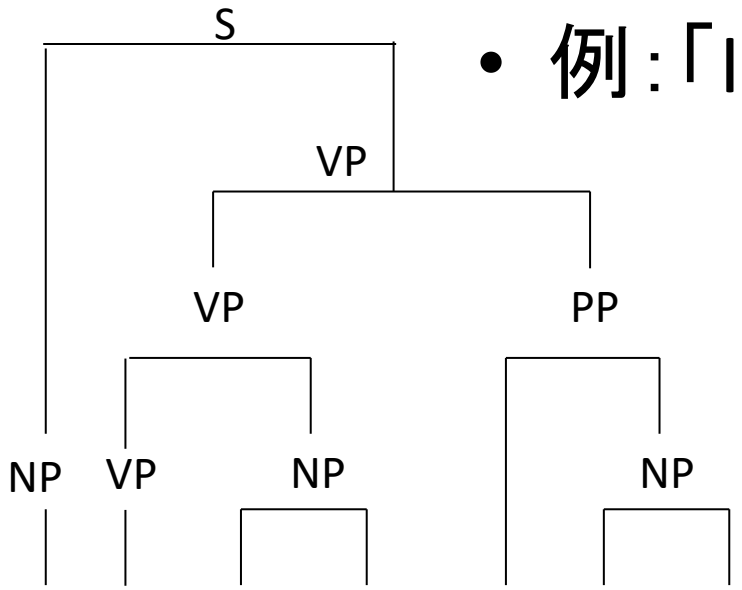
黒い、目の大きな猫

余談:「黒い目の大きな女の子」は
全部で11通りの解釈

写真:Pixabay

構文解析と文法

- 文の統語的構造を同定する
  - 文法の階層的なマッチング
    - = エキスパートシステムの知識と推論
- 例:「I saw a man in the park」

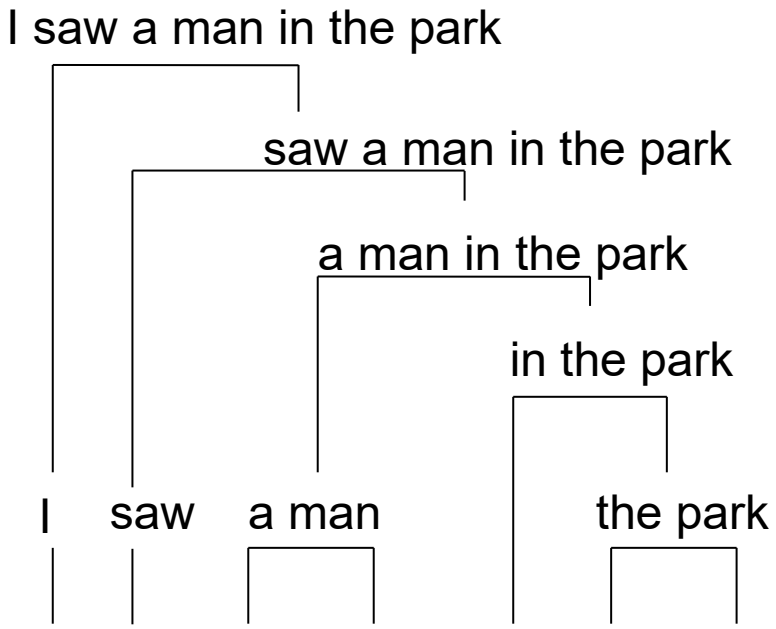


文法ルール

S ← NP VP
VP ← VP PP
VP ← VP NP
VP ← V
NP ← det N
NP ← N
PP ← P NP

構文解析とあいまい性

- 文の統語的構造を同定する
- 例:「I saw a man in the park」

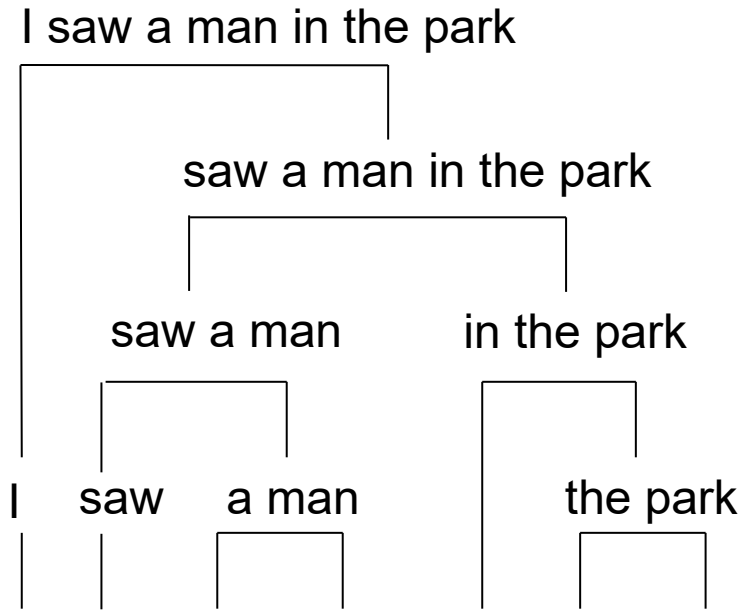


公園にいる男を見た

I saw a man in the park

構文解析とあいまい性

- 文の統語的構造を同定する
- 例:「I saw a man in the park」



公園で男を見た

I saw a man in the park

意味解析

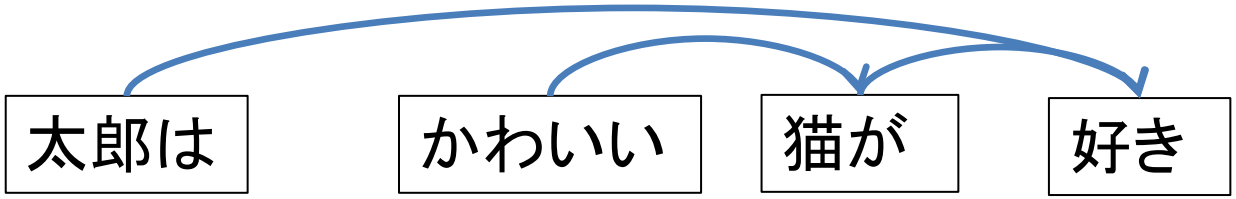
自然言語処理の基礎技術

太郎はかわいい猫が好き

単語に分割　　形態素解析

太郎　は　かわいい　猫　が　好き

修飾関係の決定　　構文解析
係り受け解析



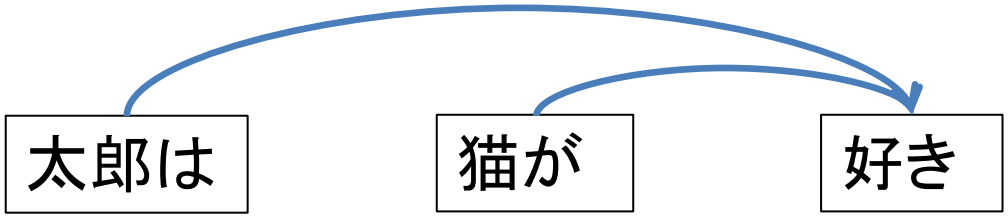
意味の同定　　**意味解析**

好き: agent-太郎 object-猫

意味解析

- 文の構造から意味を同定する

「太郎は猫が好き」

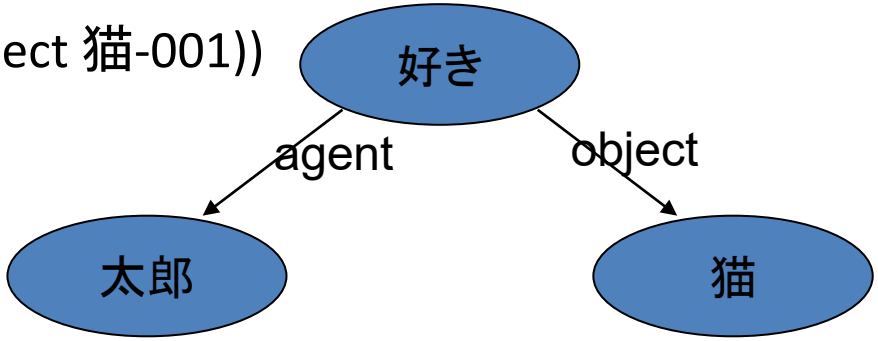


– フレーム

(好き (agent X?) (object Y?))

実体化

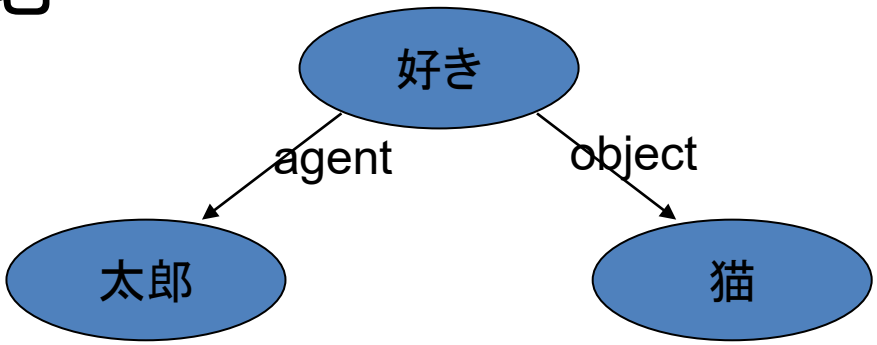
(好き-001 (agent 太郎-001) (object 猫-001))



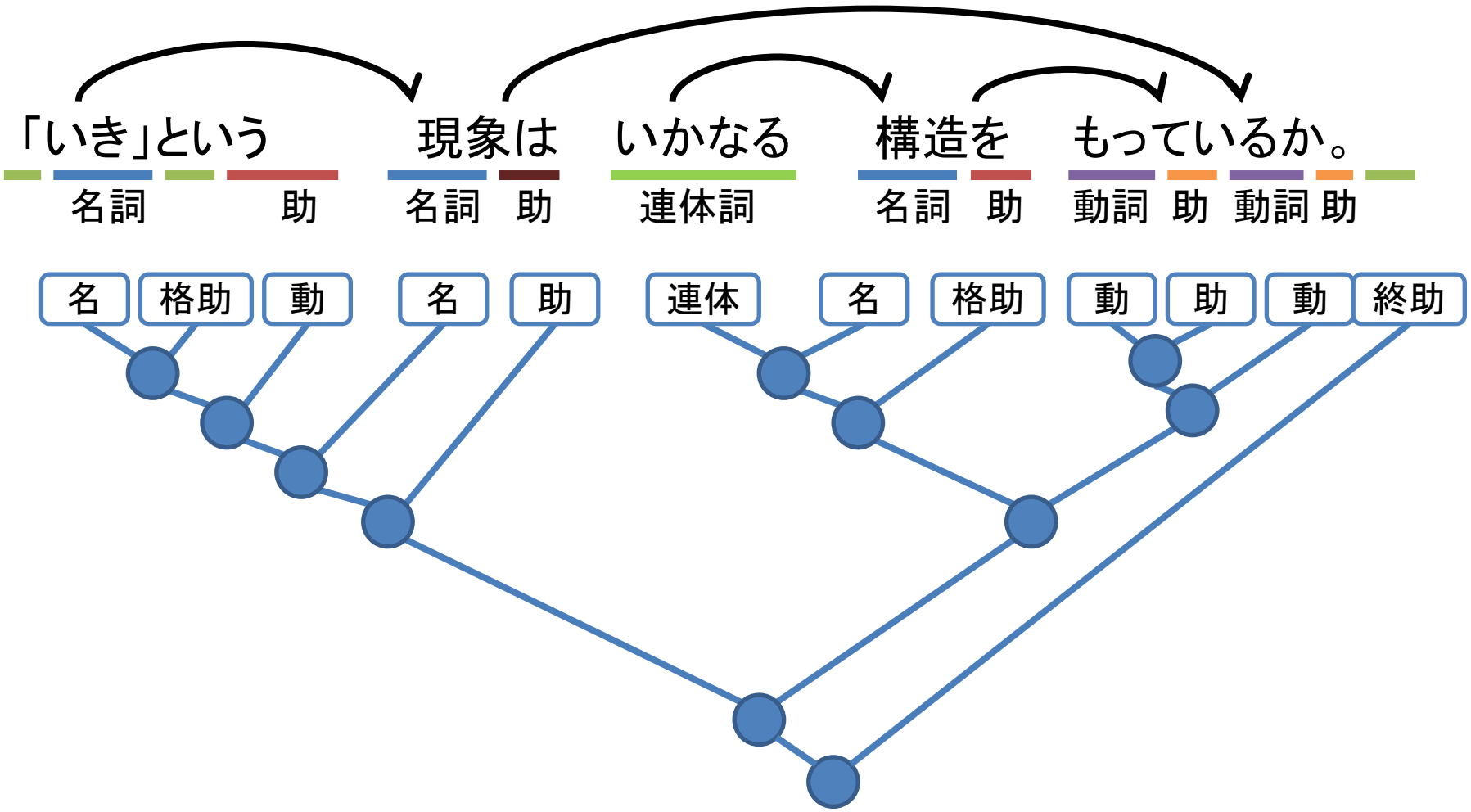
意味解析

- 太郎は猫が好き
- 猫が太郎は好き
- 太郎が好きなのは猫だ
- 猫が好きな太郎は・・・

↑すべて同じ意味を含む



構文意味解析



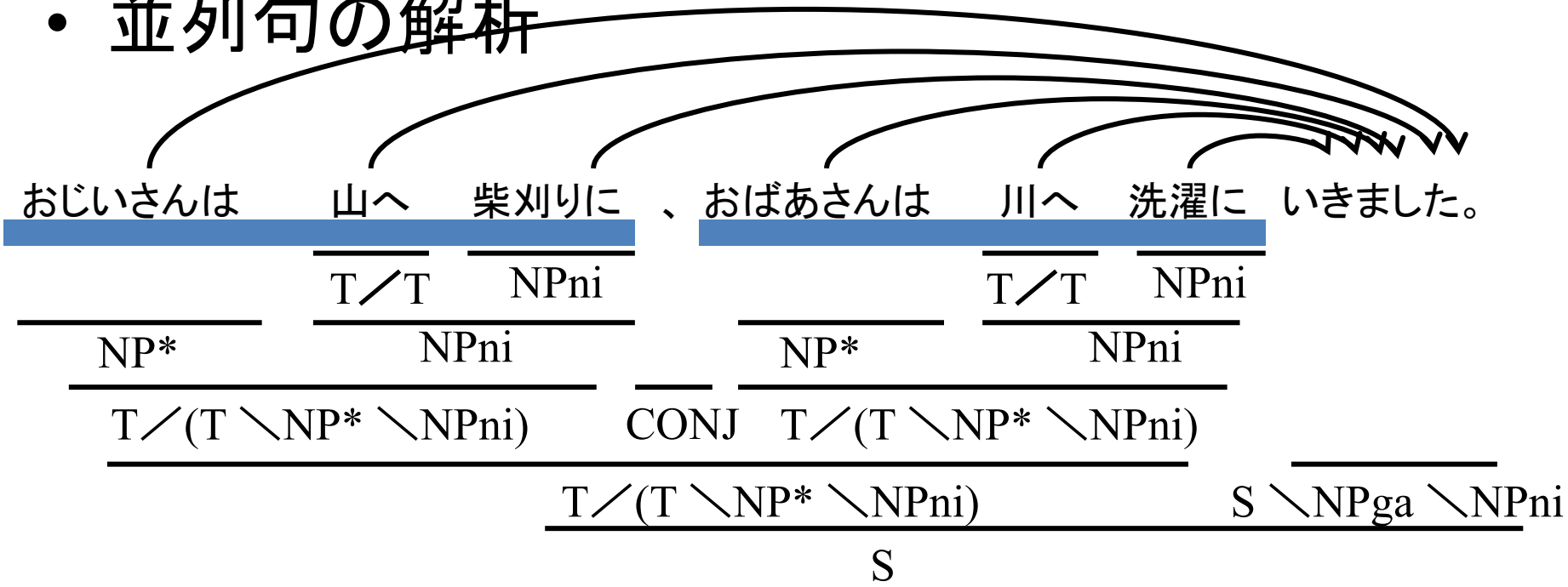
CCG構文解析？

- Combinatory Categorial Grammar とは
  - 言語を数学的に記述する枠組み (cf. 文脈自由文法)
  - 文脈自由文法と比べて
    少ない組み合わせ規則（6~8個程度）
    単語に対する詳細な記述
    (「動詞」→「ガ格とヲ格をとる動詞の過去接続形」)

    を使ってより精密な解析を可能にする
- CCG文法で文の構造を解析すると何がうれしいのか
  - 構造解析の利点の例:　並列句の解析
  - CCG文法による利点の例:　格の解析

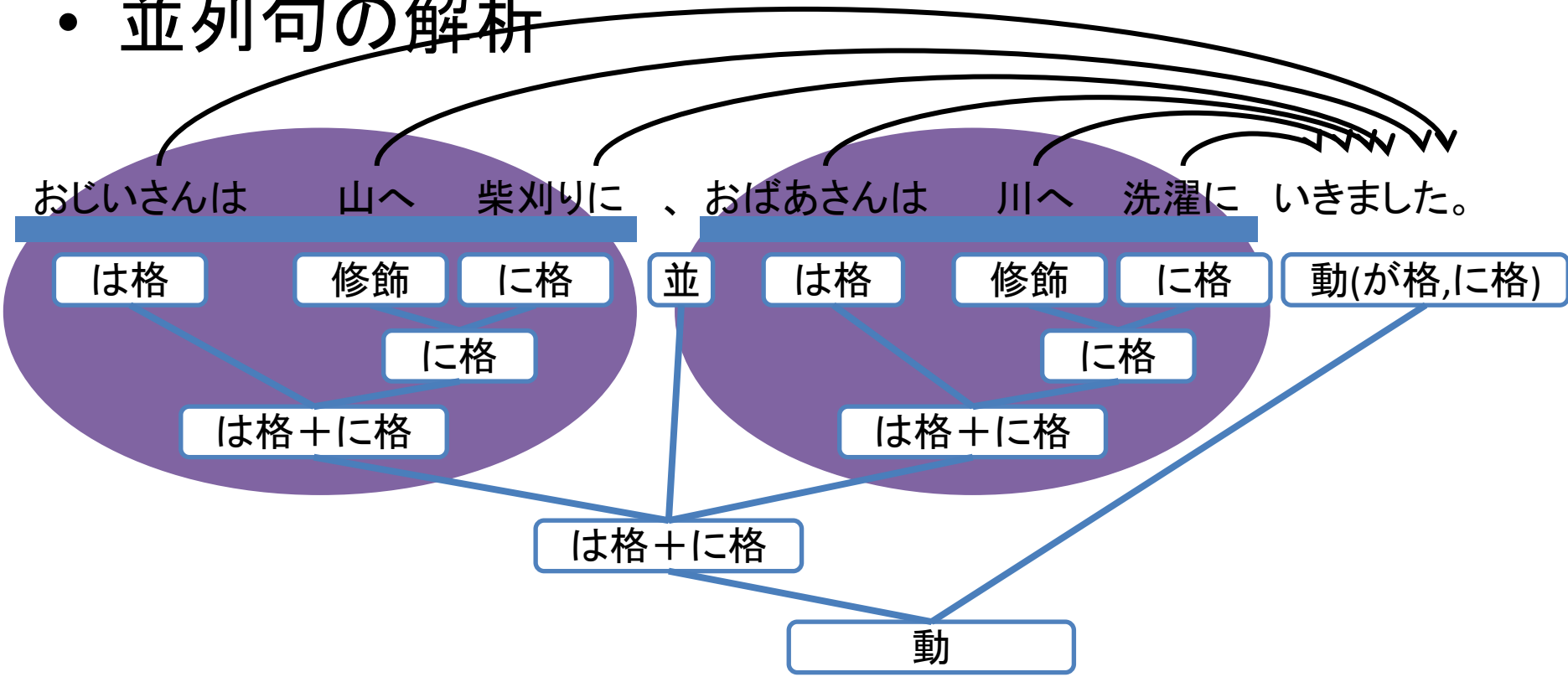
並列句の係り受けと構文構造

- 並列句の解析



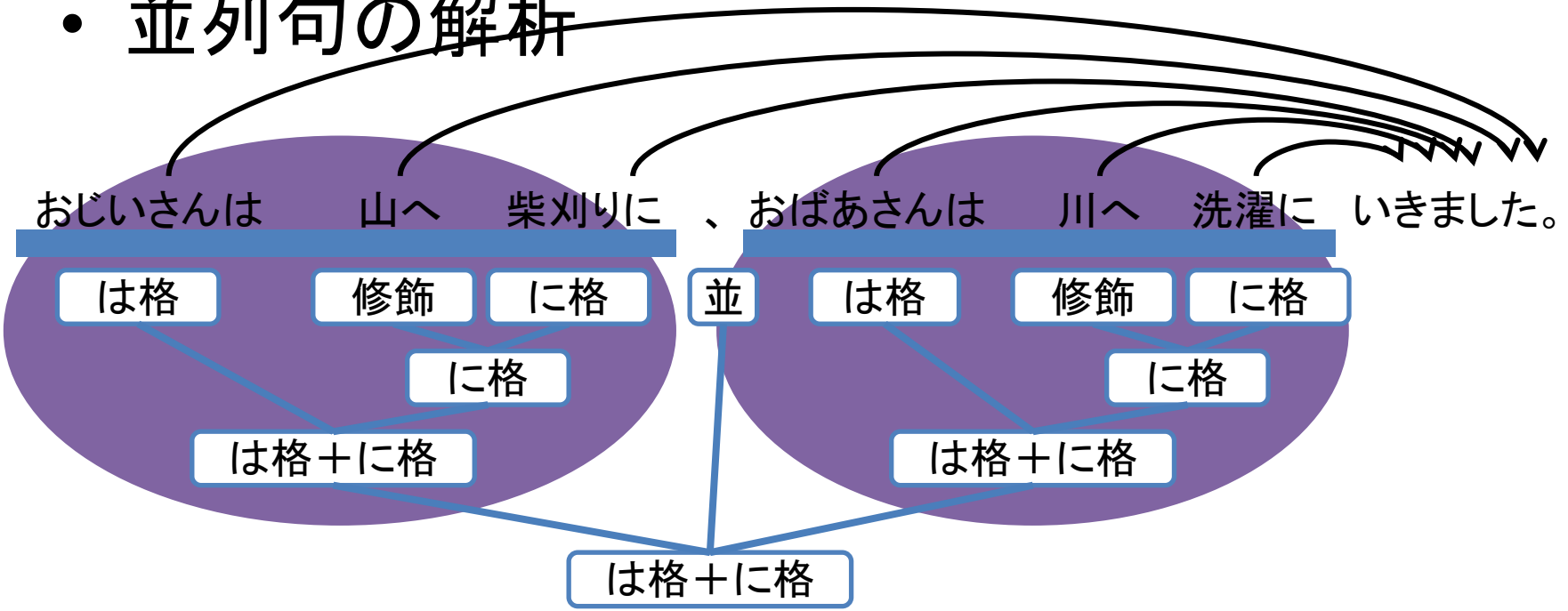
並列句の係り受けと構文構造

- 並列句の解析



並列句の係り受けと構文構造

- 並列句の解析

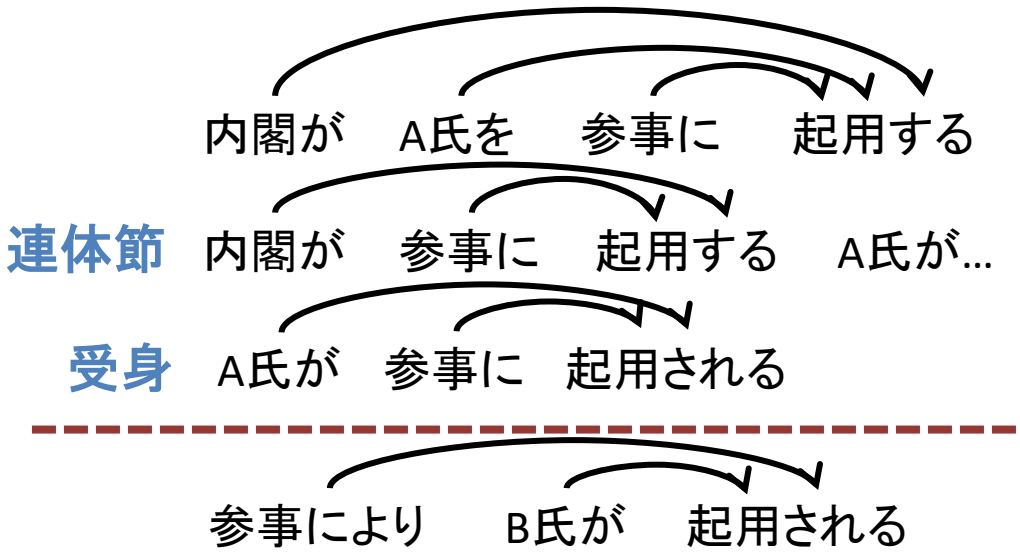


文中で「柴刈り」と並列された語句

語句	頻度
洗濯	48
薪	10

格の解析

- 例えば受身表現・連体節　の「正規化」



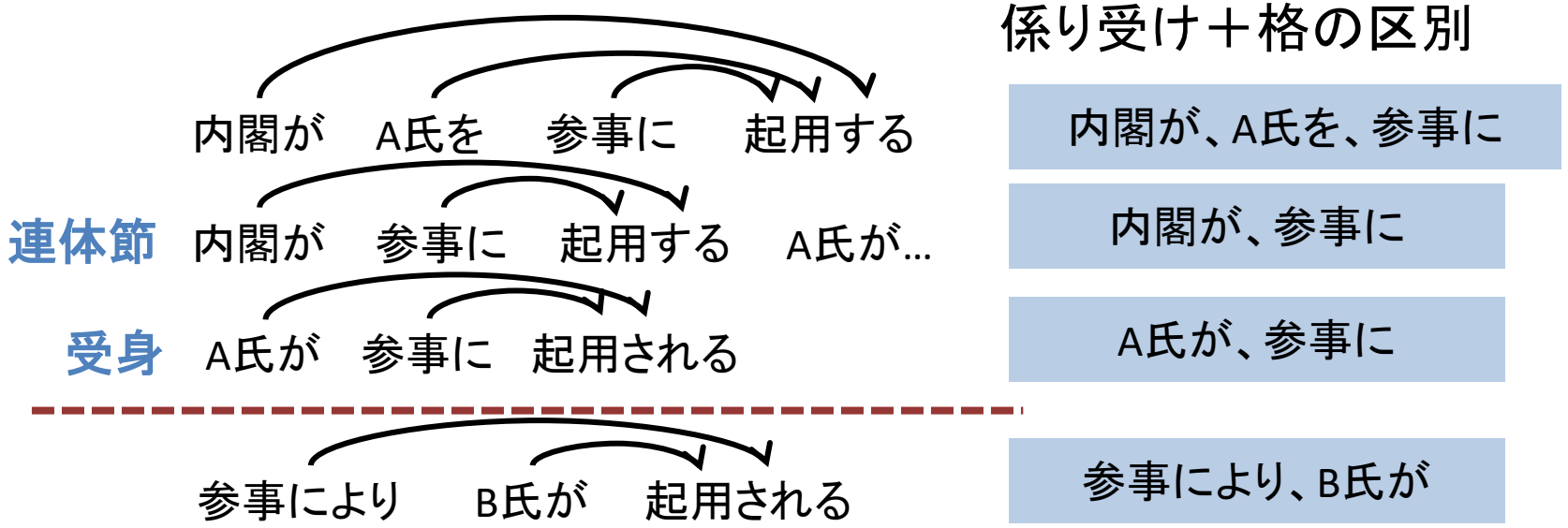
単語間係り受けによる区別

内閣、A氏、参事
内閣、参事
A氏、参事
参事、B氏

格の解析

- 例えば受身表現・連体節 の「正規化」

係り受け＋格の区別



起用する
ガ格（例：内閣が）
ニ格（例：参事に）
ヲ格（例：A氏を）

起用される（受身）
ガ格（A氏が）
ニ格（参事に）
~~ヲ格~~
ニヨリ格（内閣により）

格の解析

- 例えば受身表現・連体節　の「正規化」

内閣が　A氏を　参事に　起用する

連体節　内閣が　参事に　起用する　A氏が…

受身　A氏が　参事に　起用される

委員により　B氏が　起用される

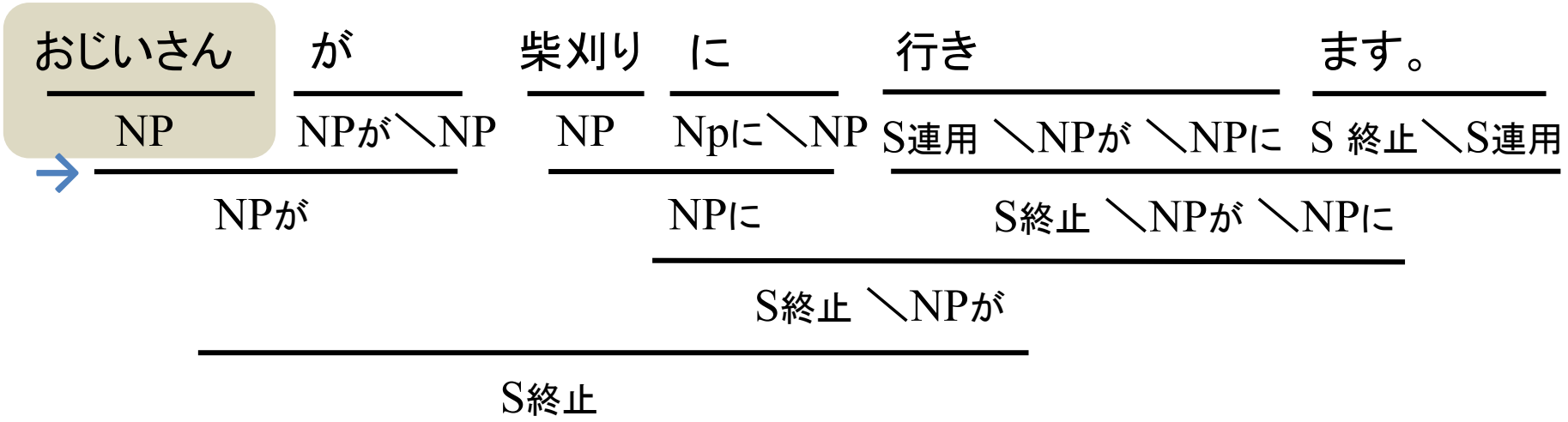
係り受けとあわせて
よりきめ細かい区別が可能

格の「正規化」

格	
ガ格	内閣／？
ヲ格	A氏
ニ格	委員

格	
ガ格	委員
ヲ格	B氏
ニ格	？

CCG文法による構文解析



- 日本語CCG文法の構成要素
  - 辞書：語と範疇（語の振る舞いを示す記号）の組
  - 組み合わせ規則：句と句を組み合わせて句を作る少数の規則

CCG文法による構文解析に必要な要素

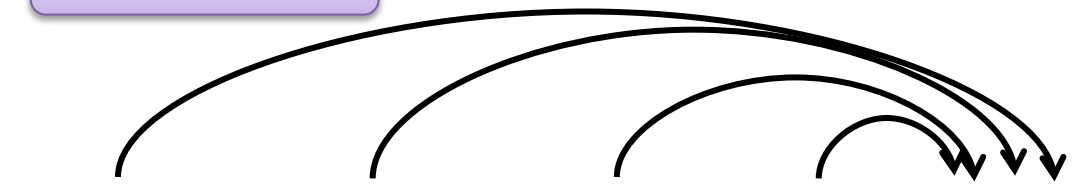
- 日本語CCG文法
  - 辞書：語と範疇（語の振る舞いを示す記号）の組
  - ✔ 組み合わせ規則：句と句を組み合わせて句を作る少数の規則

- 構文解析モジュール
  - ✔ 改変CKYアルゴリズム：文の構造を列挙
  - 曖昧性解消モジュール：文の構造から尤もらしいものを選択

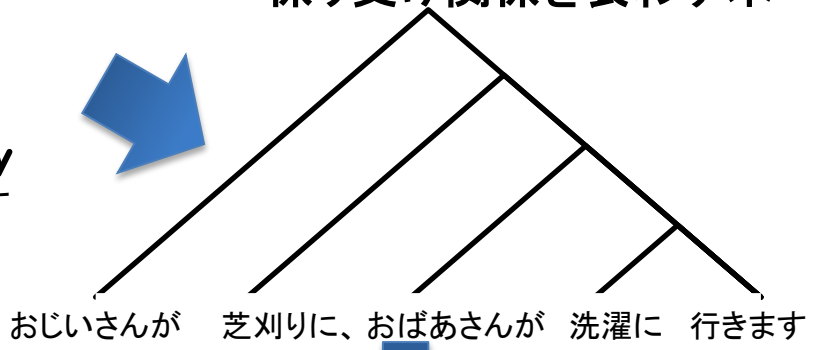
辞書と曖昧性解消については
「コーパス指向の文法開発」手法を用いる

日本語CCG構文解析器の構築：図解

京都コーパス

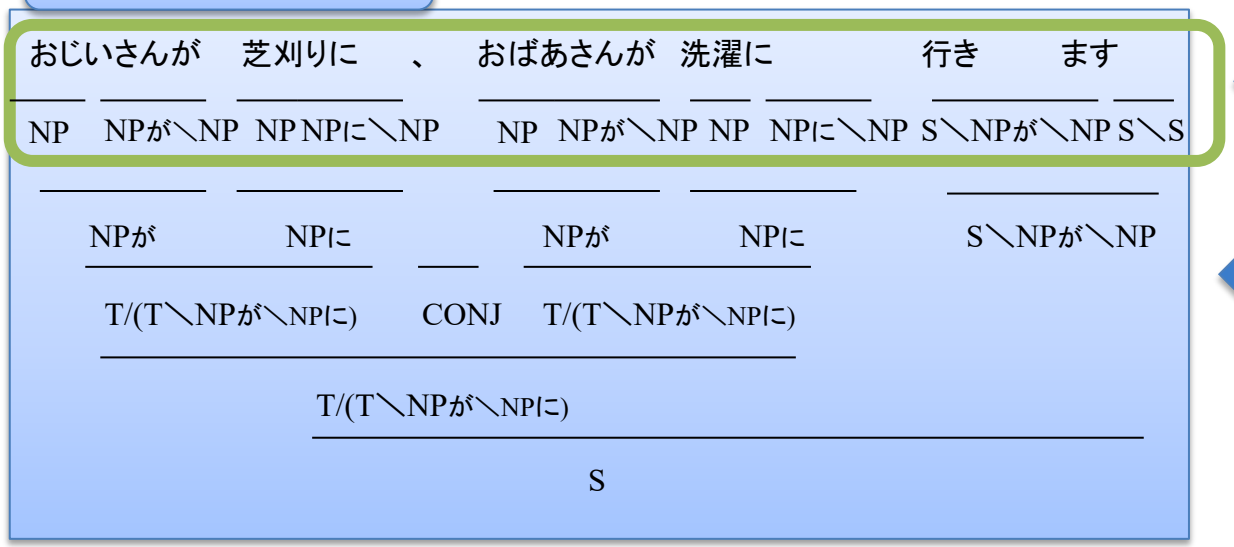


係り受け関係を表わす木



CCGコーパス

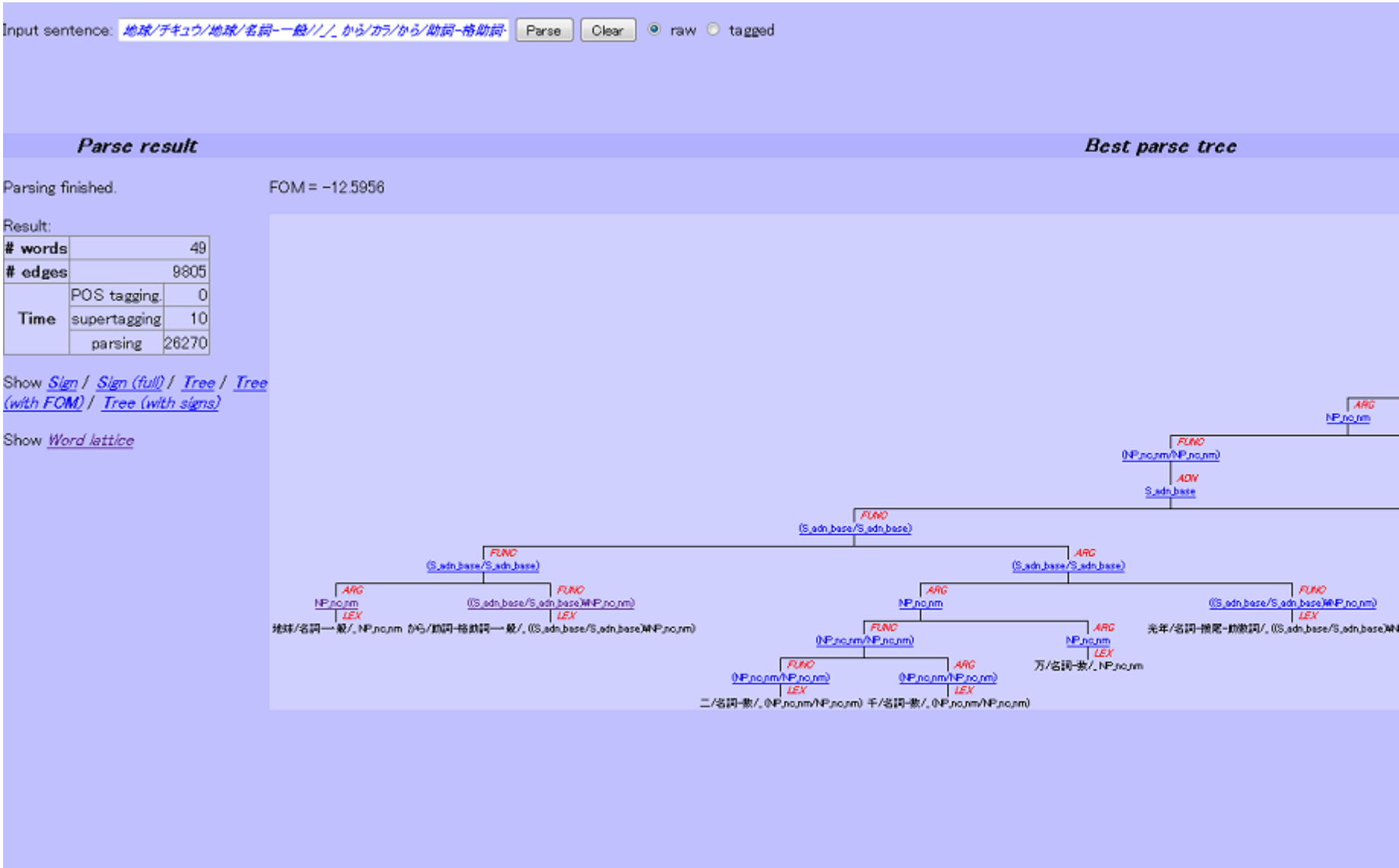
パターンルール



NAISTコーパス

「と」コーパス

CCG構文解析システム



自然言語処理の基礎技術ツール

- 形態素解析
  - Chasen http://chasen-legacy.osdn.jp/
  - JUMAN http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
  - MeCab http://taku910.github.io/mecab/
  - Kuromoji http://www.atilika.org/
- 係り受け解析
  - KNP http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
  - Cabocha http://taku910.github.io/cabocha/

構文解析ミニ演習

- 格構造解析器（CCGパーザ）
  - http://nlp.he.u-tokyo.ac.jp:8000/jparser/demo.html
- 係り受け解析器（KNP）
  - http://lotus.kuee.kyoto-u.ac.jp/nl-resource/cgi-bin/knp.cgi

- 上記に同じ文を入力して解析結果を確認すると共に、解析の違いを確認する
  - 例
    - 「気象庁は強風や高波に注意するよう呼びかけています。」

本日の出席キーワード

「syntax」

文脈解析

さらに高度な自然言語処理
– 文脈(コンテクスト)を理解すること –

- 「僕はキツネ！」

イラスト:いらすとや

文脈の理解

A「明日は？」

文脈の理解

A「今日遊びに行かない？」
B「今日は用事があってダメだね。」
A「明日は？」
B「明日なら大丈夫。」

文脈の理解

A「明日は？」

文脈の理解

A「今日雨降るんだって？」
B「天気予報でそう言ってたよ。」
A「明日は？」
B「明日もまだ降るらしいよ。」

文脈の理解

A「明日は？」

→「明日の予定は？」
「明日の天気は？」

対話システム

- ことばの理解＋現実世界の理解＋関連付け

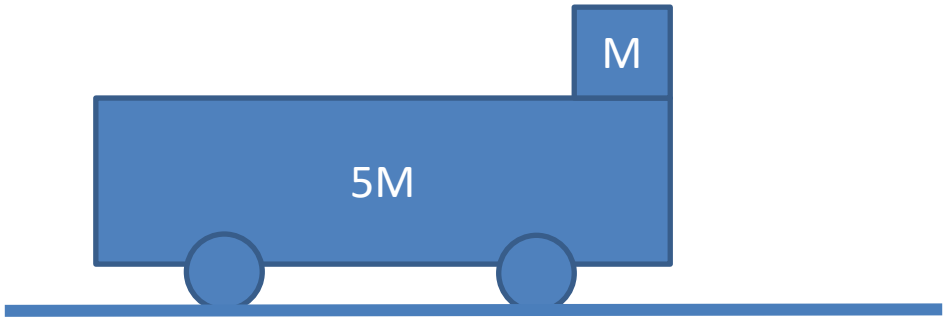
©いらすとや

そこ？
黒い？
カップ？
持ってくる？

どこに？
(誰に？)
どのように？

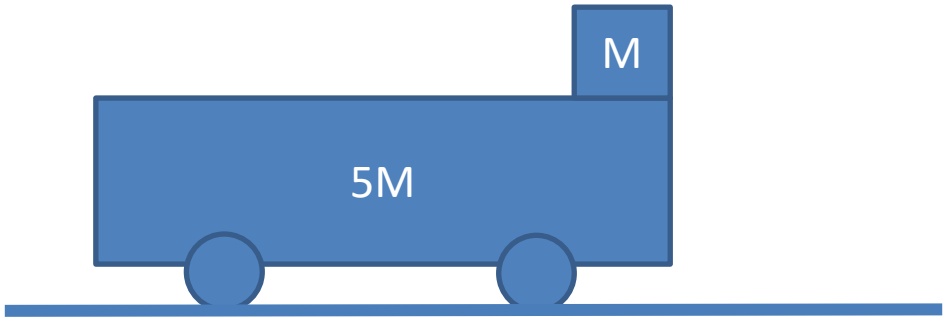
ことばと実世界の関連付け

- センター試験 物理問題
  - 水平でなめらかな床面上に、長さLの水平であらい上面をもつ質量5Mの台車を置き、台車の上面の右端に質量Mの小物体を乗せる。



ことばと実世界の関連付け

- センター試験 物理問題
  - 水平でなめらかな床面上に、長さLの水平であらい上面をもつ質量5Mの台車を置き、台車の上面の右端に質量Mの小物体を乗せる。



ことばの処理(理解)とあいまい性

- ことばを正確に理解するにはあいまい性を正しく扱うことが必要
  - **アップル**
    - 果物？　コンピュータ会社？
  - 黒い目の大きな猫
    - 大きいのは？黒いのは？
  - **そのカップ取ってきて**
    - 「その」の指すものは？
  - **明日は？**
    - 状況により様々な意味

あいまい性：一つの表現 ⇔ 複数の意味

ことばの処理(理解)とことばのゆれ

- ことばを正確に理解するにはことばのゆれを正しく扱うことが必要
  - 表記のゆれ

    コンピュータ　⇔ コンピューター
  - 構造的なゆれ

    school of science ⇔ science school
  - 意味的なゆれ

    コンピュータ ⇔ 計算機

ゆれ：複数の表現 ⇔ 一つの意味

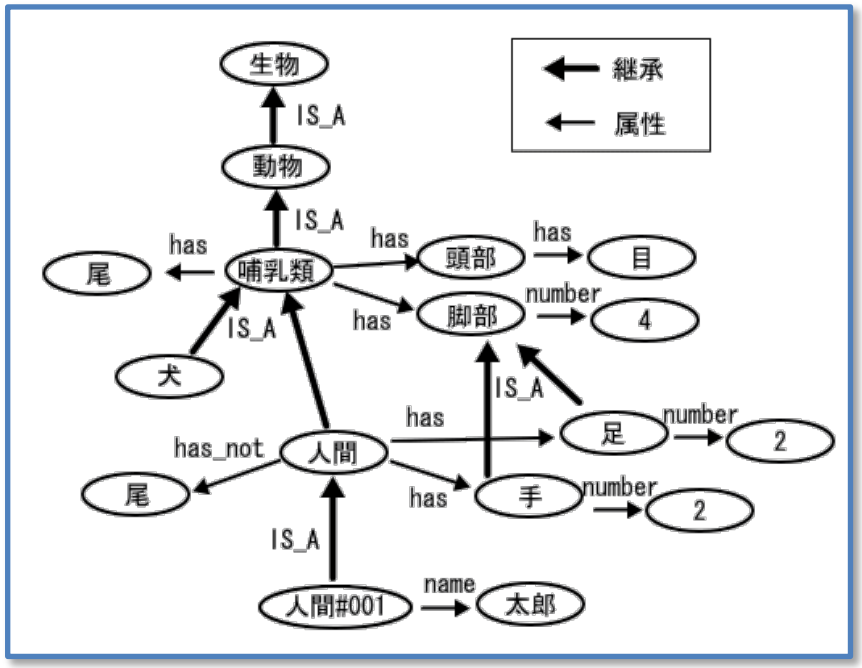
自然言語処理とデータ

- 辞書
  - 形態素／固有語
  - 専門用語
  - センチメント（感情表現）
  - 翻訳（日英、英日、etc.）
- 文法
  - 係り受け
  - 格構造／格フレーム
- 知識
  - 共起データ
  - 意味ネットワーク
  - シソーラス
  - オントロジー

言語情報処理ポータル
http://www.jaist.ac.jp/project/NLP_Portal/

知識表現

- 意味ネットワーク(1960-)は人間の記憶の一種である意味記憶の構造を表すためのモデルである。コリンズとキリアンによって考えられた。
- Cyc(1984-)は、一般常識をデータベース（知識ベース）化し、人間と同等の推論システムを構築することを目的とするプロジェクト。20年たっても書き終わらない。



意味ネットワーク

※ has関係は、part-of関係の逆

(#$isa #$BillClinton #$UnitedStatesPresident)
"Bill Clinton belongs to the collection of U.S. presidents"

(#$genls #$Tree-ThePlant #$Plant)
"All trees are plants".

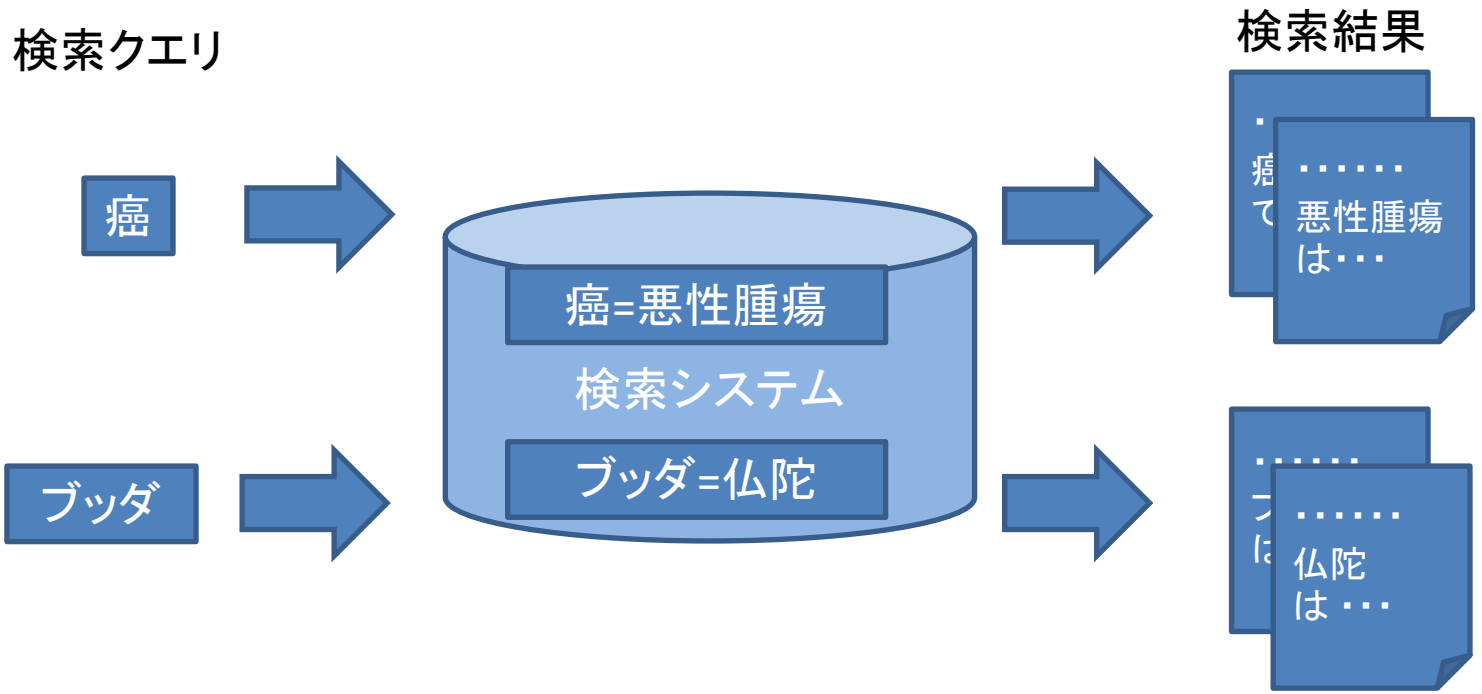
(#$capitalCity #$France #$Paris)
"Paris is the capital of France."

Cycプロジェクトで記述された知識の例

2016年度学術俯瞰講義 第9回「人工知能の未解決問題とディープラーニング」
松尾 豊先生 スライドp.18より引用

知識による情報検索の改良

- シソーラス(同義語辞書)の導入
- ことばのゆれを同一視して検索を行う



導入するシソーラスデータ

- **JST科学技術用語シソーラス**
  - 科学技術用語の同義語・異表記・英語表記等の辞書
  - 12万同義語対、54万語

- **人文社会37万語英和対訳大辞典**
  - 人文社会系専門用語の日英対訳辞書
  - 同一英単語に対する日本語訳語を同義語とみなす
  - 4万同義語対、13万語

JST科学技術用語シソーラス

- 同義語データ例:
  - 脊柱腫瘤, 脊柱腫りゅう
  - 血圧降下剤, 抗高圧症薬, 抗高血圧薬, 血圧降下薬, 降圧薬, antihypertensive, 抗高血圧剤, 抗高血圧症薬, 高血圧治療薬, depressor, 高血圧治療剤, 降圧剤, 高血圧薬
  - 絃, chord, string, 弦
  - アトム, atom, atomic, 原子
  - 原子エネルギー準位, 原子エネルギ準位
  - 核, nuclear, nucleus, nuclei, ニュークリアス, 原子核
  - immunoreactivity, 免疫活性
  - サービス科学, サービスサイエンス

人文社会37万語英和対訳辞典

- 人文社会科学系分野の専門用語の日英対訳辞書
- データ例:

＼E01＼Buddha
＼J01＼ブッダ
＼J02＼ブッダ
＼T01＼[文化〈世界史〉]

＼E01＼Buddha
＼J01＼仏陀
＼J02＼ブッダ
＼T01＼[文化〈宗教〉];[文化〈日本美術〉]

E01: 英単語
J01: 日本語単語
J02: 読み
T01: 分野

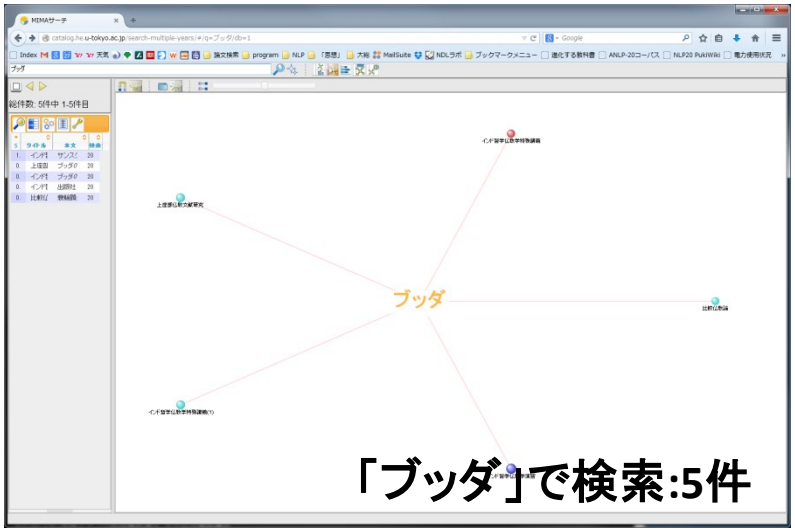
人文社会37万語英和対訳辞典

- 同一の英単語に対する訳語を同義語とみなす
- 同義語例
  - Buddha: ブッダ,仏陀
  - Buddhist temple: 寺院,寺,仏閣
  - yearly income: 年間所得,年収
  - library: 図書館,図書室,文庫,蔵書
  - lifelong education: 生涯学習,生涯教育

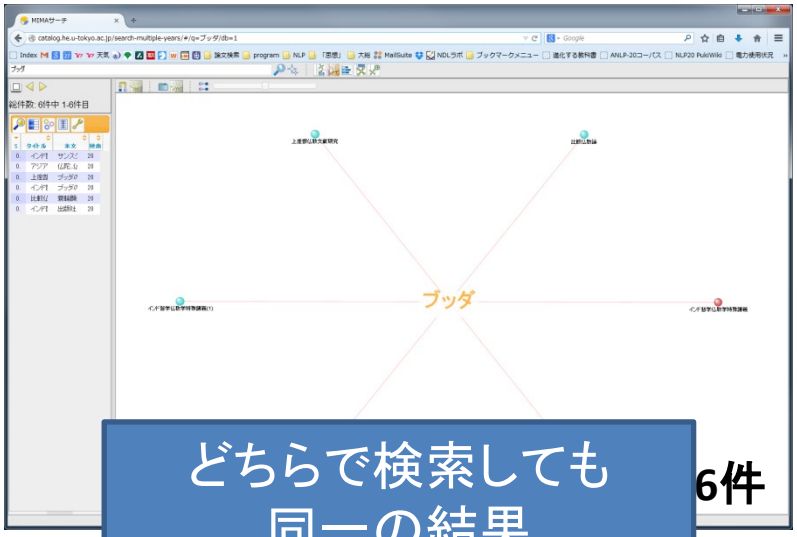
活用例(人文社会用語辞書)

「ブッダ」「仏陀」

表記のゆれの解消

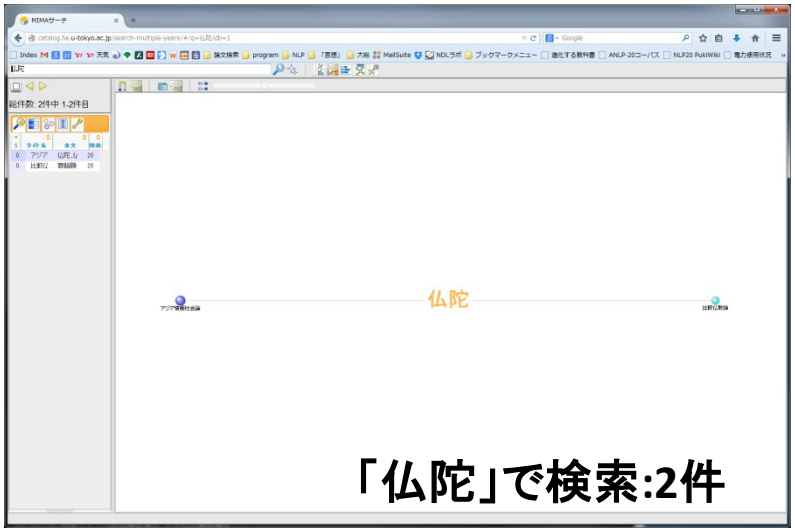


「ブッダ」で検索:5件



6件

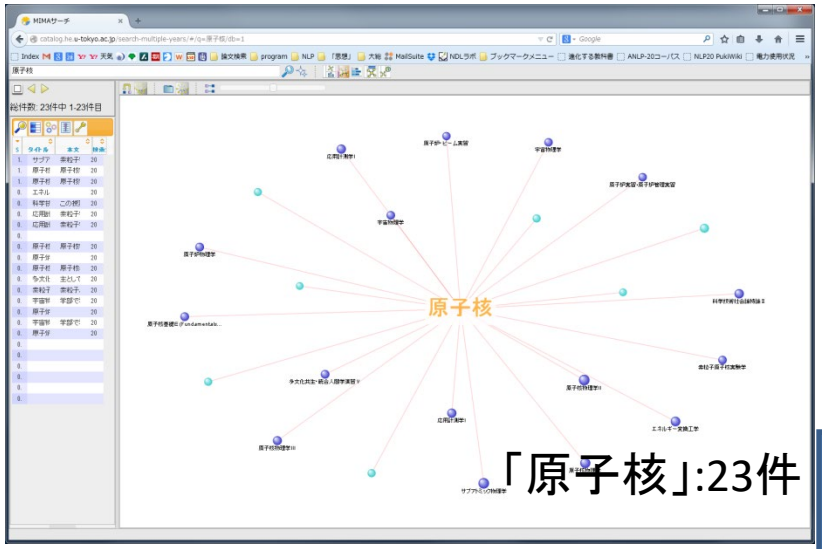
どちらで検索しても
同一の結果
(どちらかを含む文書)



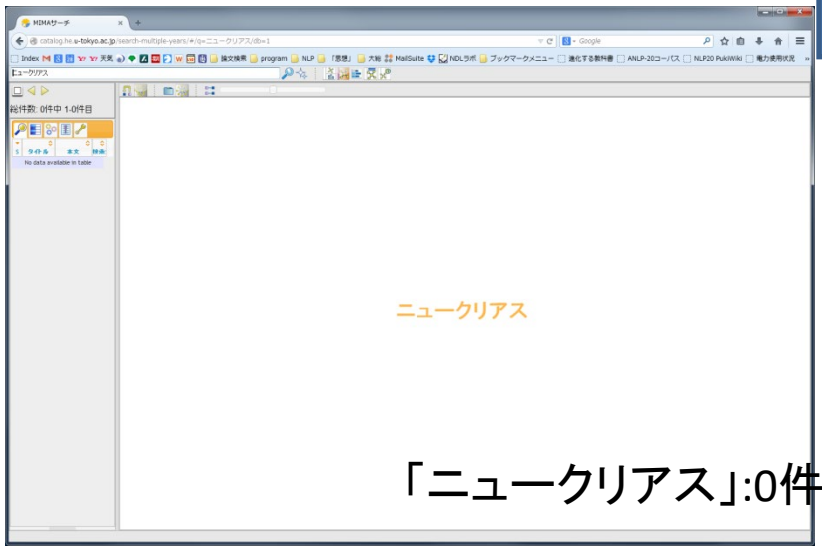
「仏陀」で検索:2件

「仏陀」で検索:6件

活用例(科学技術用語辞書)
「原子核」「ニュークリアス」「nuclear」



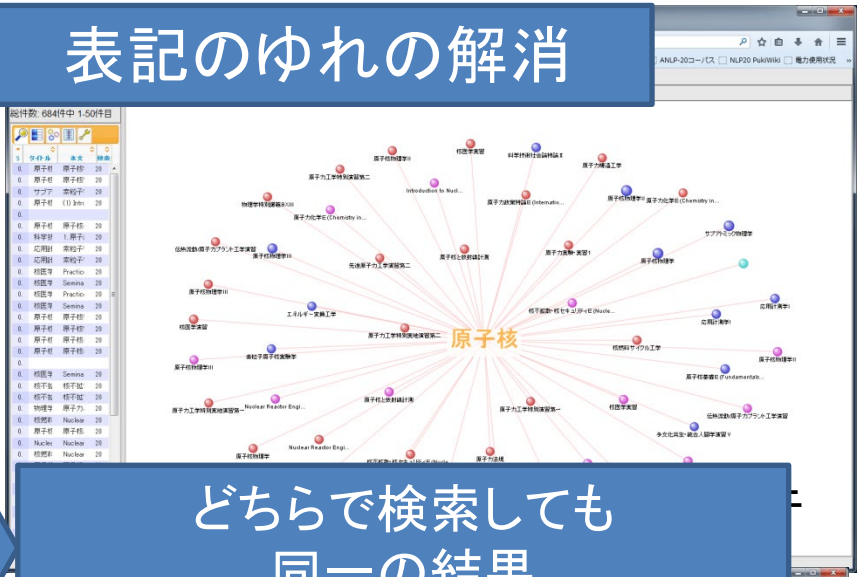
「原子核」:23件

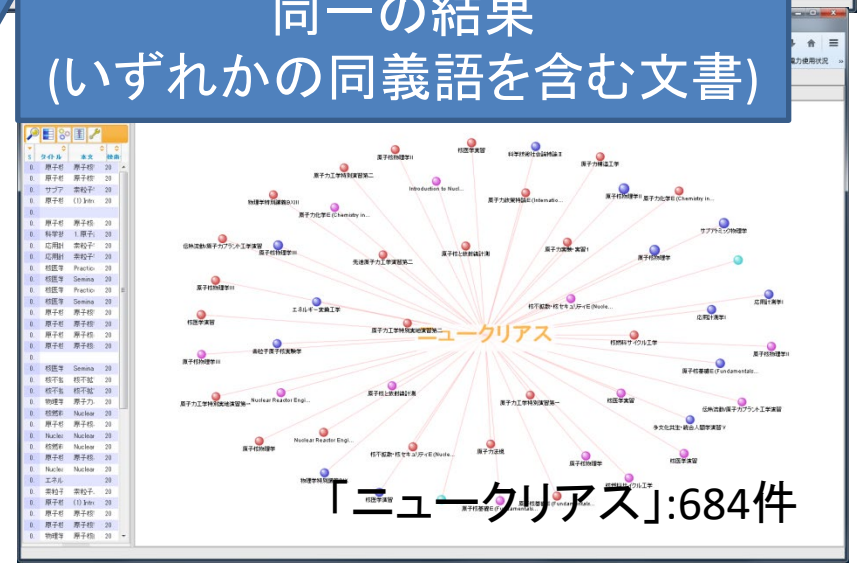


「ニュークリアス」:0件

表記のゆれの解消

どちらで検索しても
同一の結果
(いずれかの同義語を含む文書)





「ニュークリアス」:684件

シソーラス利用の問題点

- 同義語でない語を同義語と扱うことがある
  - 「原子核」「ニュークリアス」
    - ⇔「nuclear」
    - ⇔「原子力」
  - 「証券」「有価証券」
    - ⇔「instrument」
    - ⇔「機器」「計器」

同じ英単語の異なる意味(あいまい性)を正しく扱う必要がある

word2vec

- ニューラルネットワークを用いて単語をベクトル表現化する手法
- テキスト中の各単語についてその周辺に出現する単語の情報を基に計算
- word2vecで作られたベクトル表現はある種の単語の意味・概念を表現する
  - ベクトル間の演算や意味の演算が可能
    - 例: v(king) - v(man) + v(woman) = v(queen)
- 今回は Wikipedia のテキストから各単語のベクトルを学習

対象文書のベクトル化

「数学ＩＡ」シラバス
工学全分野で必要不可欠な道具である、常微分方程式、ベクトル解析、変分法について学ぶ。実践的な理解を目指す。…

単語抽出

工学、全、分野、で、必要、不可欠、だ、道具、だ、ある、・・・

ベクトル化

工学: (-0.28987, 2.20560, -0.13070　0.67409, …)
全:　(0.72628, 0.84896, 1.94840, 0.66509, …)
分野: (1.17059, 1.94050, 1.00932, 1.04591, …)
で:　(1.89374, 2.01249, -0.65686, -2.03772, …)
必要: (-0.76447, 1.06354, 2.38880, -0.42196, …)
・・・

平均

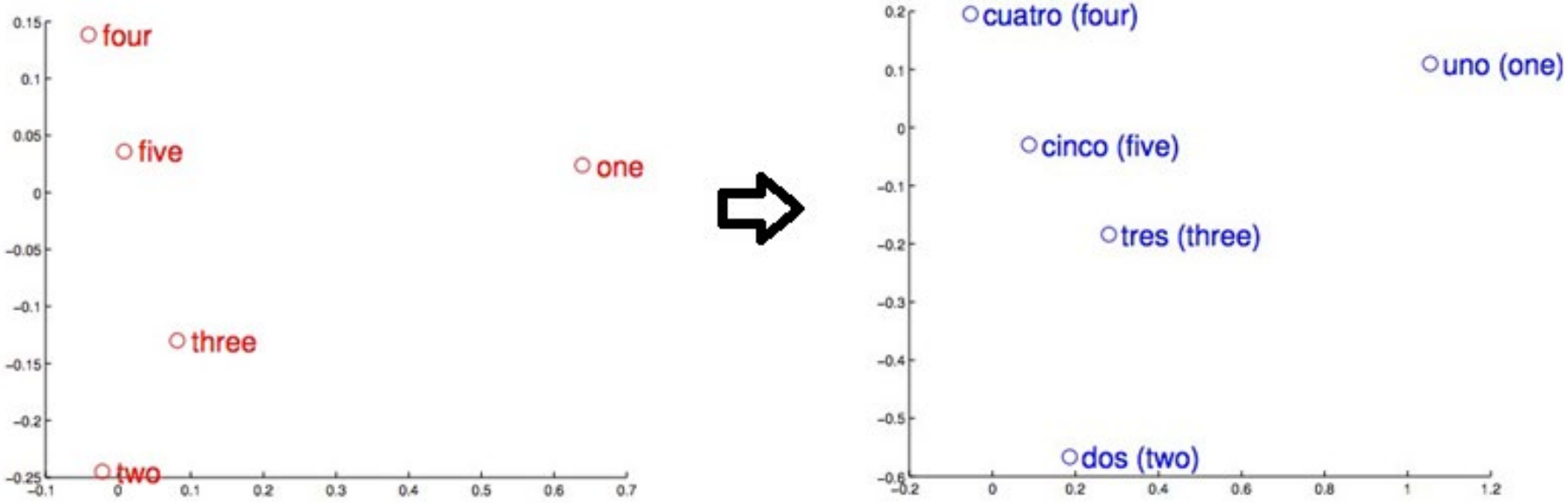
(0.77248 1.13985 -0.11331 -1.13872, …)

対象テキストを200次元の特徴ベクトルに変換

特徴ベクトル(word2vec)の「翻訳」

[Milkov et. al. 2013]

- word2vecは言語間で一定の対応を取ることができる
- 関係を表す写像を行うことで「翻訳」ができる



英語　　スペイン語

Deep Age「機械翻訳」より
https://deepage.net/bigdata/machine_learning/2016/09/02/word2vec_power_of_word_vector.html

まとめ

- 自然言語処理
  - 人間のことばをコンピュータで処理する技術
    (人間のことばをコンピュータで**理解**させる技術)
  - 様々な基礎技術・応用技術
    - 形態素解析、構文解析、情報検索・・・
  - 今やいたるところで既に活用されている
  - ことばを正しく処理するにはあいまい性・ことばのゆれ等に対する適切な処理が必要

参照：UTokyo OCW

- UTokyo OCW
  - 東大の正規講義の資料・映像を一般に公開

    http://ocw.u-tokyo.ac.jp

  - 学術俯瞰講義「ビッグデータ時代の人工知能学と情報社会のあり方」(2016Aセメスター)

    http://ocw.u-tokyo.ac.jp/course_11381/

推薦図書

『いちばんやさしい機械学習プロジェクトの教本 : 人気講師が教える仕事にAIを導入する方法』
韮原 祐介(著)、インプレス、2018年

