コンピュータで言葉を理解する
言葉の意味を処理するとは？

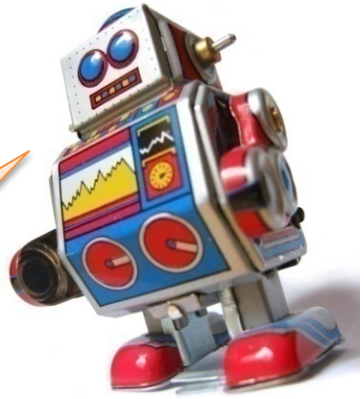
宮尾祐介
国立情報学研究所

自然言語処理

- 人間の言葉(自然言語)を理解するコンピュータを作ることを目指す学問
- 言葉による情報交換、コミュニケーションを助ける
  - かな漢字変換
  - 検索
  - 自動翻訳
  - 対話システム

おいしいイタリアン
知らない？

いいところ知ってますよ。
高級店とカジュアル店の
どっちがいいですか？



自然言語処理の応用

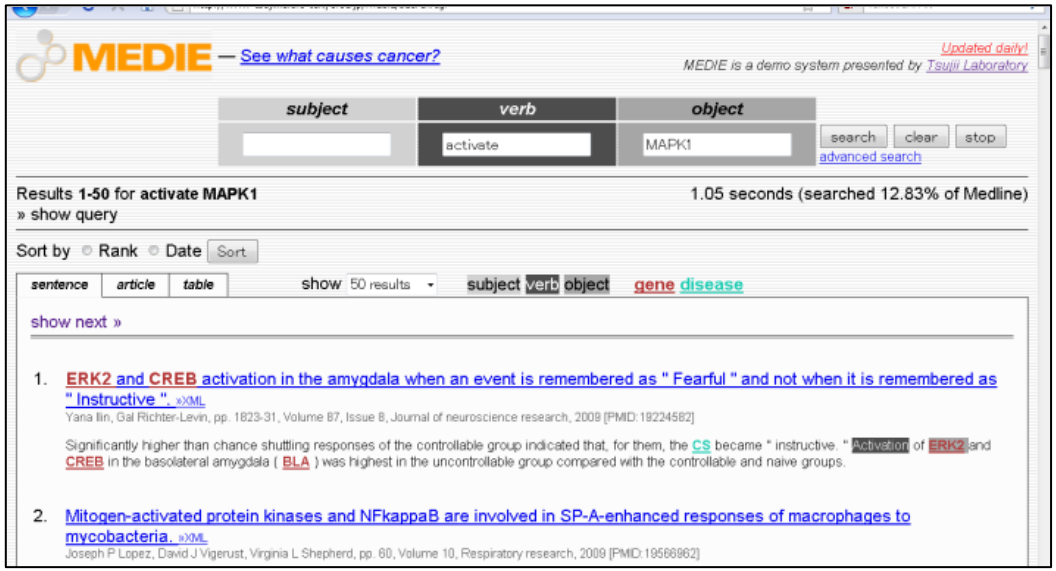
かな漢字変換

自動翻訳

対話システム

自然言語処理の応用

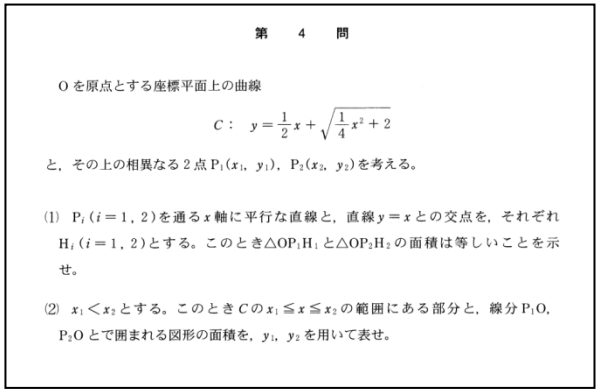
- 生命科学論文の検索システム
  - 膨大な数の論文から、自分が欲しい情報を探す
- 意味に基づく検索ができる
  - 「MAPK1が活性化される」ことについて書かれている論文が欲しい



ロボットは東大に入れるか

- 人工知能で東大入試を突破することを目指すプロジェクト
- 自然言語処理が重要な役割を果たす

問題文



第　4　問

O を原点とする座標平面上の曲線

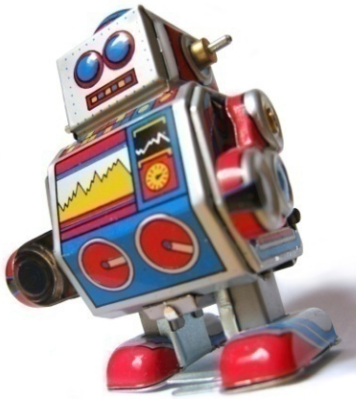
$$C:\quad y=\frac{1}{2}x+\sqrt{\frac{1}{4}x^2+2}$$

と，その上の相異なる 2 点 $P_1(x_1,\ y_1)$，$P_2(x_2,\ y_2)$ を考える。

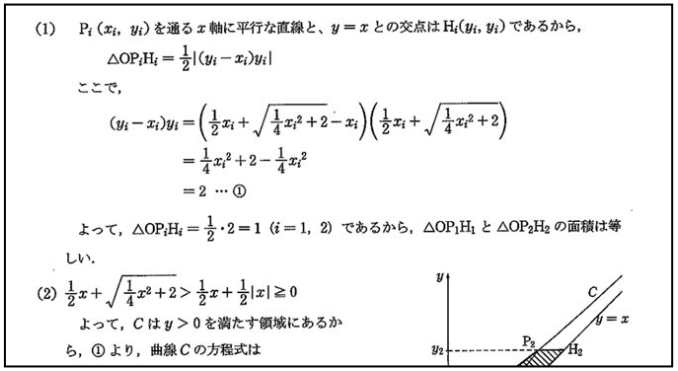
(1)　$P_i\ (i=1,\ 2)$ を通る x 軸に平行な直線と，直線 $y=x$ との交点を，それぞれ $H_i\ (i=1,\ 2)$ とする。このとき $\triangle OP_1H_1$ と $\triangle OP_2H_2$ の面積は等しいことを示せ。

(2)　$x_1<x_2$ とする。このとき C の $x_1 \leqq x \leqq x_2$ の範囲にある部分と，線分 P_1O，P_2O とで囲まれる図形の面積を，y_1，y_2 を用いて表せ。

2010年度東京大学前期試験　数学



回答



(1)　$P_i\ (x_i,\ y_i)$ を通る x 軸に平行な直線と，$y=x$ との交点は $H_i(y_i,\ y_i)$ であるから，

$$\triangle OP_iH_i=\frac{1}{2}|(y_i-x_i)y_i|$$

ここで，

$$(y_i-x_i)y_i=\left(\frac{1}{2}x_i+\sqrt{\frac{1}{4}x_i^2+2}-x_i\right)\left(\frac{1}{2}x_i+\sqrt{\frac{1}{4}x_i^2+2}\right)$$
$$=\frac{1}{4}x_i^2+2-\frac{1}{4}x_i^2$$
$$=2\ \cdots①$$

よって，$\triangle OP_iH_i=\frac{1}{2}\cdot 2=1\ (i=1,\ 2)$ であるから，$\triangle OP_1H_1$ と $\triangle OP_2H_2$ の面積は等しい．

(2)　$\frac{1}{2}x+\sqrt{\frac{1}{4}x^2+2}>\frac{1}{2}x+\frac{1}{2}|x|\geqq 0$

よって，C は $y>0$ を満たす領域にあるから，①より，曲線 C の方程式は

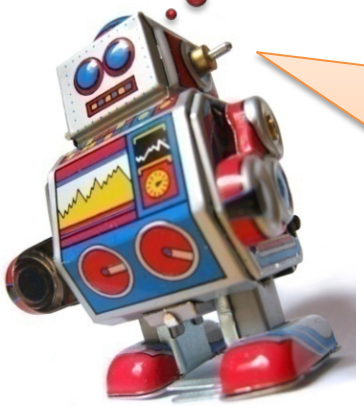
2010年度駿台予備学校大学入試解答速報

自然言語処理の考え方

- 入力･出力＝自然言語
- 入力と出力の間になんらかの「計算」が行われている
  ＝言葉の理解

理解 ＝計算

おいしいイタリアン知らない？



いいところ知ってますよ。
高級店とカジュアル店のどっちがいいですか？

入力

出力

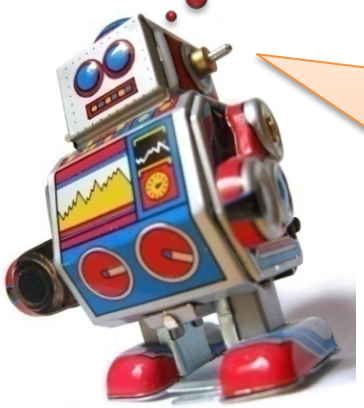
自然言語処理の難しさ

- 人間は無意識のうちに言葉を理解している
- 言葉を理解する仕組みを直接観察することはできない

→ 入力・出力を観察し、その間で起きていることを推測する

理解 ＝計算

おいしいイタリアン知らない？



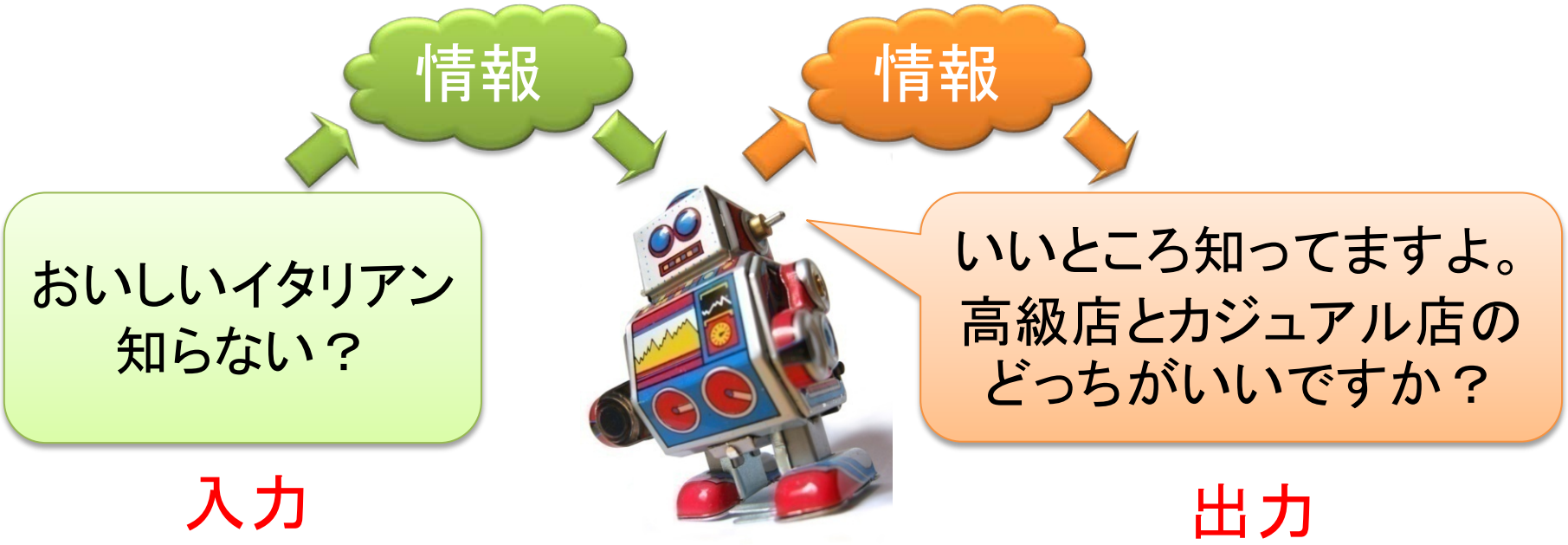
いいところ知ってますよ。高級店とカジュアル店のどっちがいいですか？

入力

出力

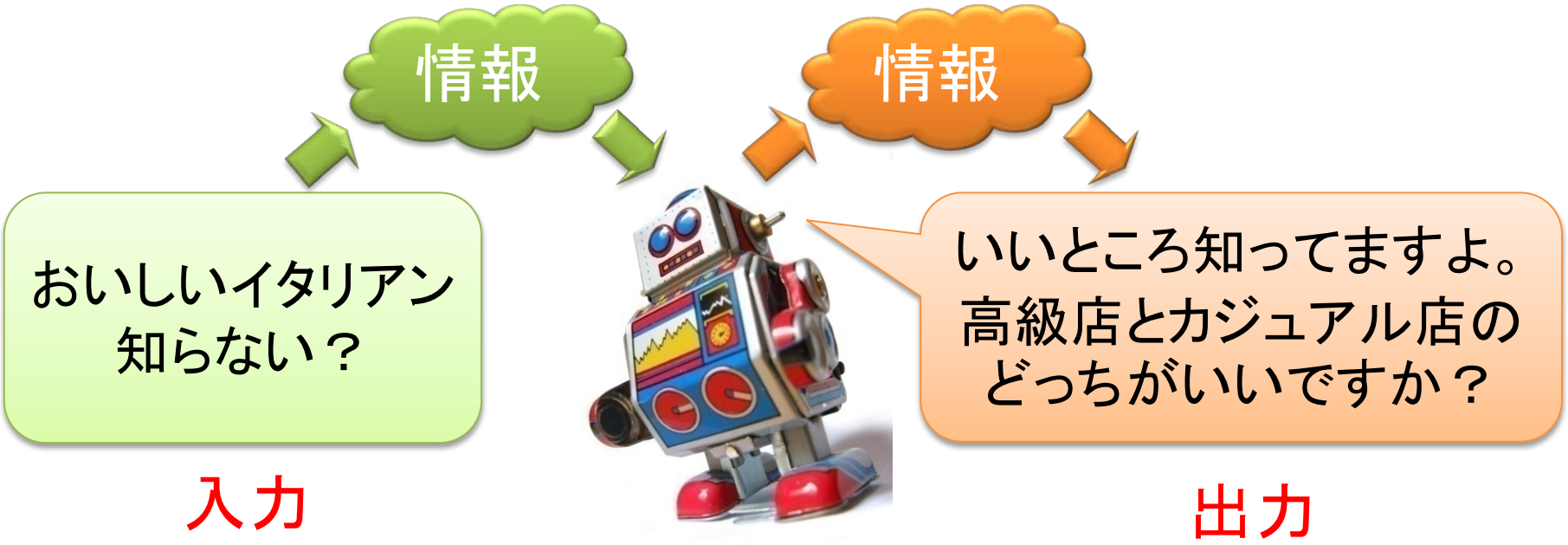
言葉の理解とは？

- 人間は言葉で「情報」をやりとりしている
  - 情報を提供する
  - 情報を獲得する
- 言葉が表す情報＝「意味」



「意味」とは何か？

- なぜ言葉を文字列そのものとして処理するのではなく、「意味」を考える必要があるのか？



文字列と意味の不一致

- 文字列の近さ≠意味の近さ
  - 友達からケーキをもらった。
  - 友達からケヤキをもらった。
  - 友達からモンブランをもらった。
  - 友達がケーキをもらった。
  - 友達からケーキをもらえなかった。
  - 朝早くからケーキをもらった。
- 文字列と意味との対応関係が複雑

「意味」とは何か？

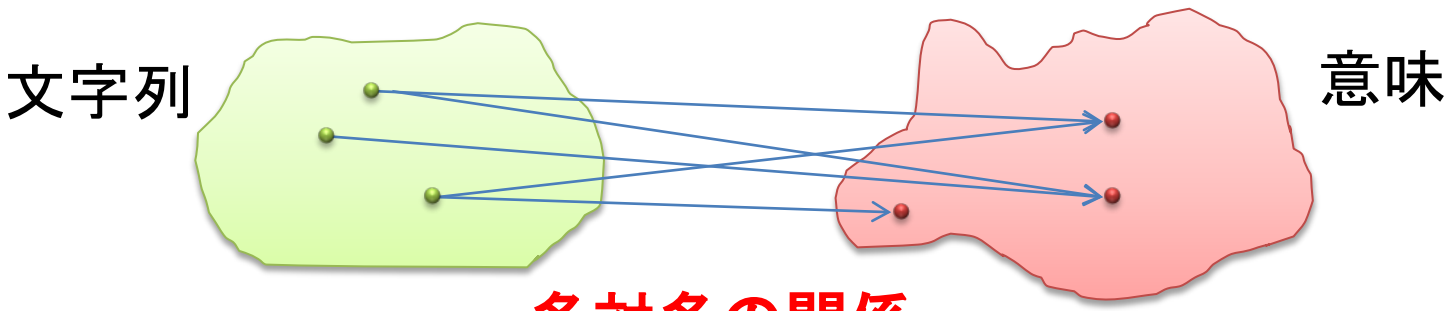
- 「意味」(大辞林)
  1. 言葉・記号などで表現され、また理解される一定の内容。
  2. ある表現・作品・行為にこめられた内容・意図・理由・目的・気持ちなど。
  3. 物事がある脈絡の中でもつ価値。重要性。意義。
- どうやってコンピュータで実現すればよいのか？
- たぶん人間の頭の中にあるが、どういうものか観察できない

→ どういうものか分からないので、その性質から迫る

自然言語処理における「意味」とは何か？

- 異なる文字列が同じ「意味」を表す
  - 私がネコにエサをあげた。
  - ネコが私にエサをもらった。
- 同じ文字列が異なる「意味」を表す
  - かわいい瞳の大きな女の子を見た。
- つまり、意味が「同じ（同値性）」「異なる（差異）」という直感を再現したい

→「同じ」「異なる」をコンピュータで計算できるように、意味の表現方法・計算方法を設計する



多対多の関係

文字列から意味へ

犬に風邪薬を飲ませると貧血状態に陥ります。

どういうときに「意味が同じ」と言えるのか？

ビーグルが風邪薬を食べたら病気になる！

意味の計算

犬に風邪薬を飲ませると
貧血状態に陥ります。

ビーグルが風邪薬を食べたら病気になる

意味の2つの側面

- **構成的意味**：文の中で単語が組み合わされて表現される意味
  - ビーグルが風邪薬を飲んだ。
  - ビーグルに風邪薬を飲ませた。
  - それがビーグルに飲ませた風邪薬だ。
- **語彙的意味**：文とは独立に、単語がもともと持っている意味
  - 犬、ビーグル、柴犬
  - 顔、目、鼻
  - あげた、もらった

つまり…

- 言葉が表す意味とは、
  - 単語を並べることによって表される意味と
  - 単語がもともと持っている意味

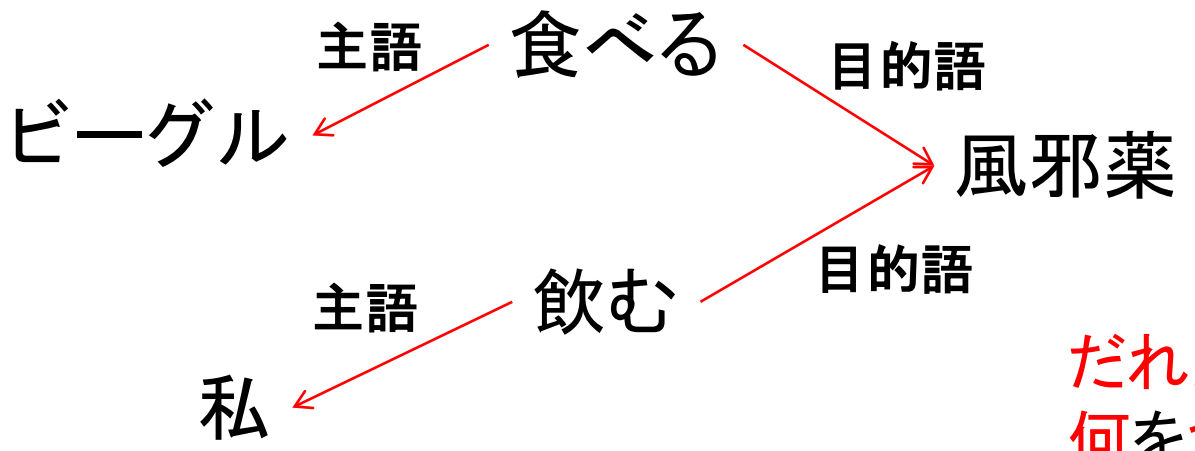
  の組み合わせ
- この二つを表現し、意味が「同じ」「違う」を計算する仕組みを考えればよい
  - 意味のデータ構造とアルゴリズム

構成的意味

- 単語を並べることによって表される意味
- ビーグルが食べた風邪薬を私も飲んだ。
  - だれが食べた？
  - 何を食べた？
  - だれが飲んだ？
  - 何を飲んだ？
- 私が風邪薬を飲んだ。
- 単語と単語のつながり
  - 主語、目的語、…

構成的意味の表現方法

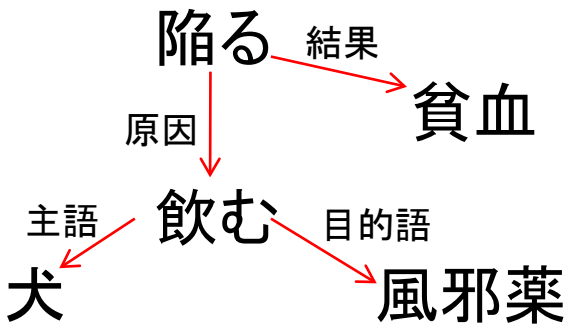
- グラフ構造で表す
  - 点＝単語、辺＝単語間のつながり
  - ビーグルが食べた風邪薬を私も飲んだ。



だれが食べた？
何を食べた？
だれが飲んだ？
何を飲んだ？

構文解析

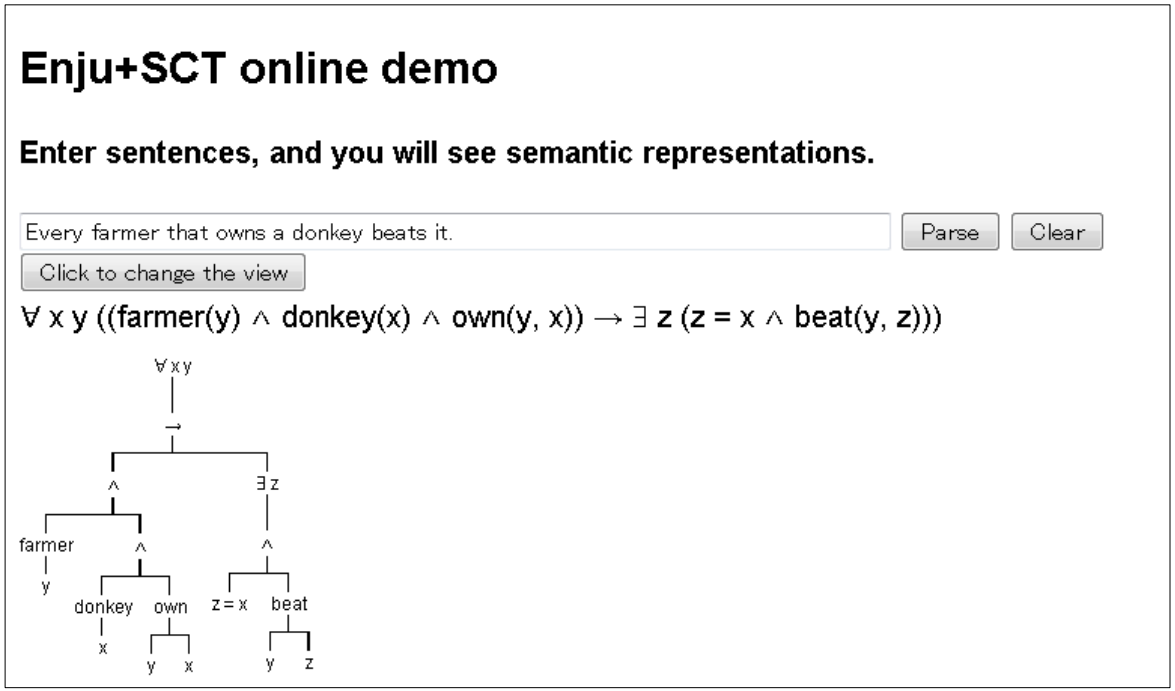
- 構成的意味は文の構造が分かれば計算できる
- 文の構造の計算＝構文解析



犬に 風邪薬を 飲ま せる と 貧血状態に 陥ります。

構文解析の研究

- Enju: 英語の構文解析器
- 90%の精度で意味表現（構成的意味）を計算

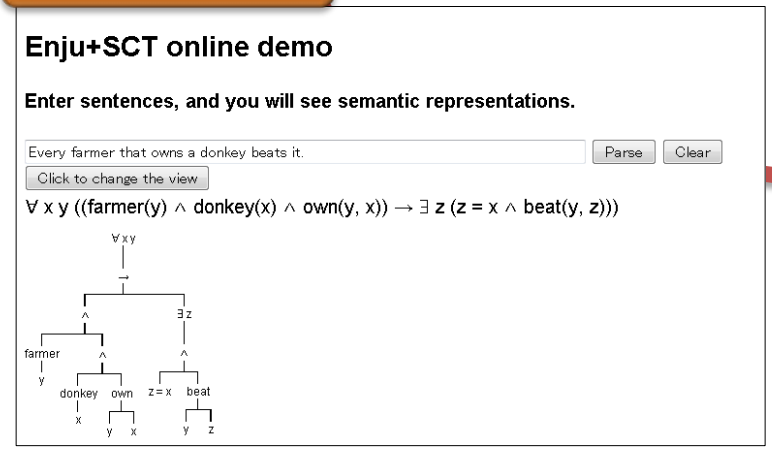


意味の計算

言葉

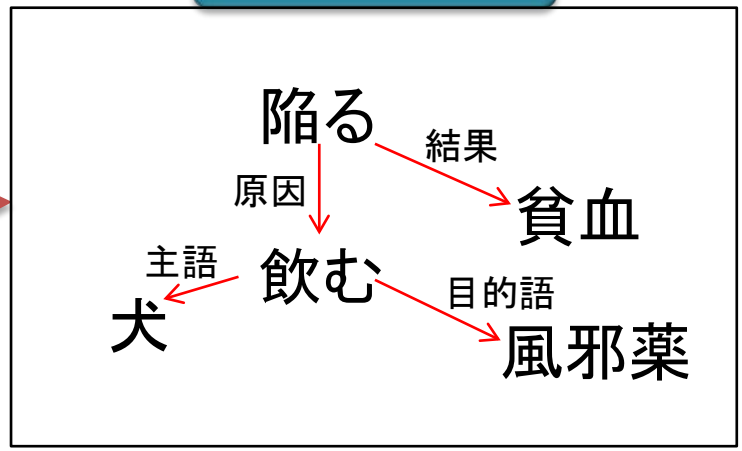
犬に風邪薬を飲ませると
貧血状態に陥ります。

構文解析



構成的意味を計算

意味表現



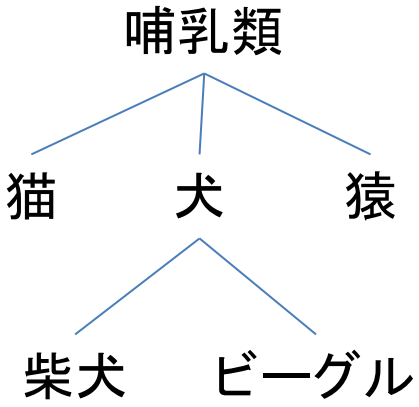
ビーグルが風邪薬を食べたら病気になる

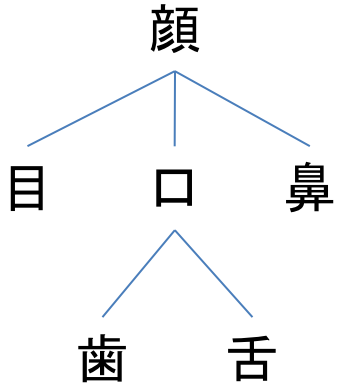
語彙的意味

- 単語がもともと持っている意味
- いろいろな種類の意味的関係
  - 同義・反義関係
    - 犬、イヌ、わんちゃん
  - 上位・下位関係
    - 犬、ビーグル、柴犬
  - 全体・部分関係
    - 顔、目、鼻

語彙的意味を表す方法

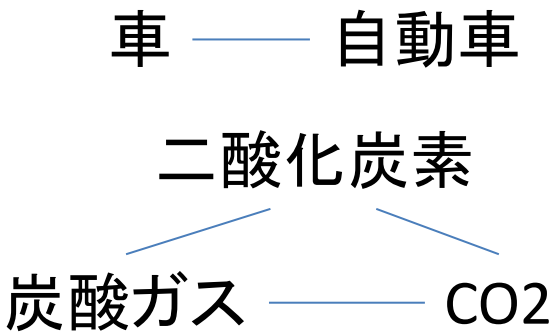
- 意味ネットワーク：単語の意味的関係をグラフ構造で表す
  - 点：単語
  - 辺：意味的関係
- いろいろな種類の関係
  - 同義・反義関係
  - 上位・下位関係
  - 全体・部分関係





同義(類義)関係

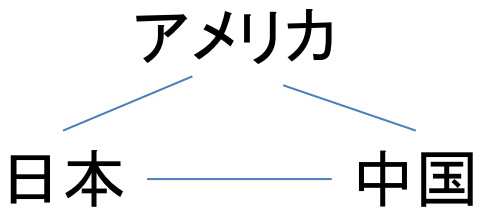
- 同義(類義):(ほとんど)全ての文脈で置き換え可能な単語
  - 車、自動車
  - 二酸化炭素、炭酸ガス、CO2



冬休みに *x* で草津まで行く。
駅前の道はたくさん *x* が走っている。
今度 *x* を修理に出すつもりだ。
昨晩家の前に赤い *x* がとまっていた。

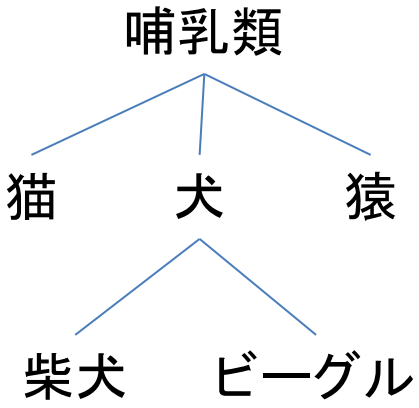
反義関係

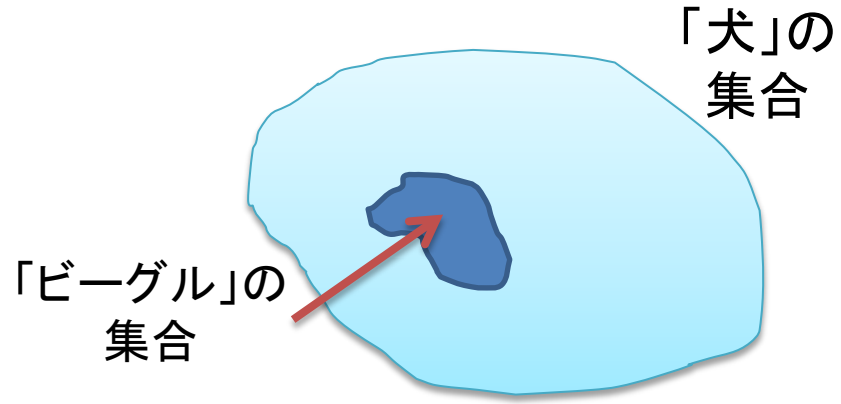
- 「反義」を定義することは実は難しい
  - 関係が無い単語対は反義ではない
  - 否定は必ずしも反義では無い
  - 反義ではなく、排他性（同時に成り立たないこと）を考えることが多い
    - おとな、こども
    - アメリカ、日本、中国
    - 犬、猫、猿
    - 大きい、小さい



上位・下位関係

- 単語 A の全ての性質を単語 B が持っているなら、A は B の上位語
  - 哺乳類　＞　犬、猫
  - 家具　＞　テーブル、ソファ
- 「B は A の一種である」「B is a A」
- A が指す集合が B が指す集合より大きい
- IS-A 関係とも言う





動詞の意味関係

- 同義・反義、上位・下位、全体・部分関係は同様に定義できる
  - 勉強する ≒ 学ぶ
- ただし、項の対応関係を考える必要がある

X が Y に Z をあげる

X が Y に Z をもらう

あげる(X, Y, Z) ＝ もらう(Y, X, Z)

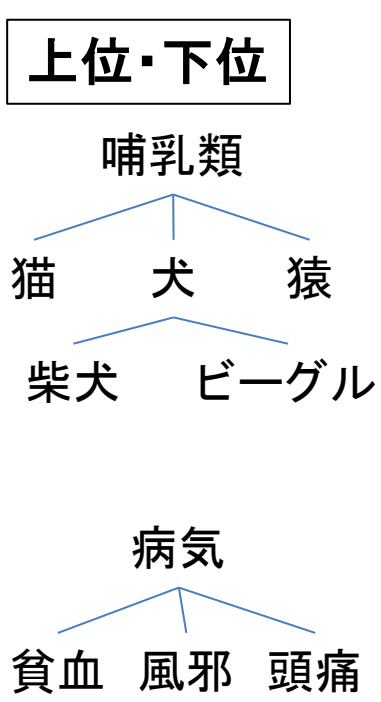
- 動詞特有の関係もある
  - 含意関係：A が成り立つなら、必ず B も成り立っている
    - X が Y に陥る → X が Y になる
    - X を後悔する → X が起きた

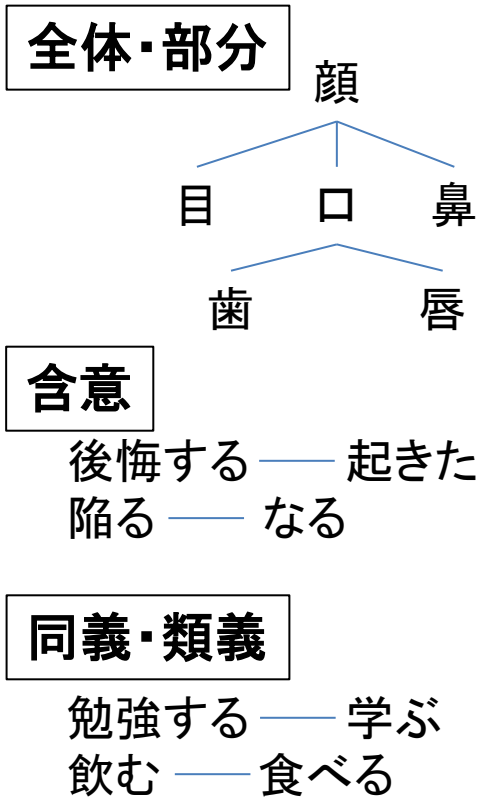
フレーズの意味的関係

- 単語だけでなく、フレーズでも意味的関係が考えられる
  - 強い雨が降る ＝ 大雨になる （同義関係）
  - いびきをかく → 寝ている （含意関係）
  - ノーベル文学賞を受賞する
    → 作家である （含意関係）

シソーラス、オントロジー

- 同義・反義関係、上位・下位関係などを収録した辞書
- 自然言語処理では欠かせないデータ
  - WordNet
  - 分類語彙表
  - 日本語語彙体系

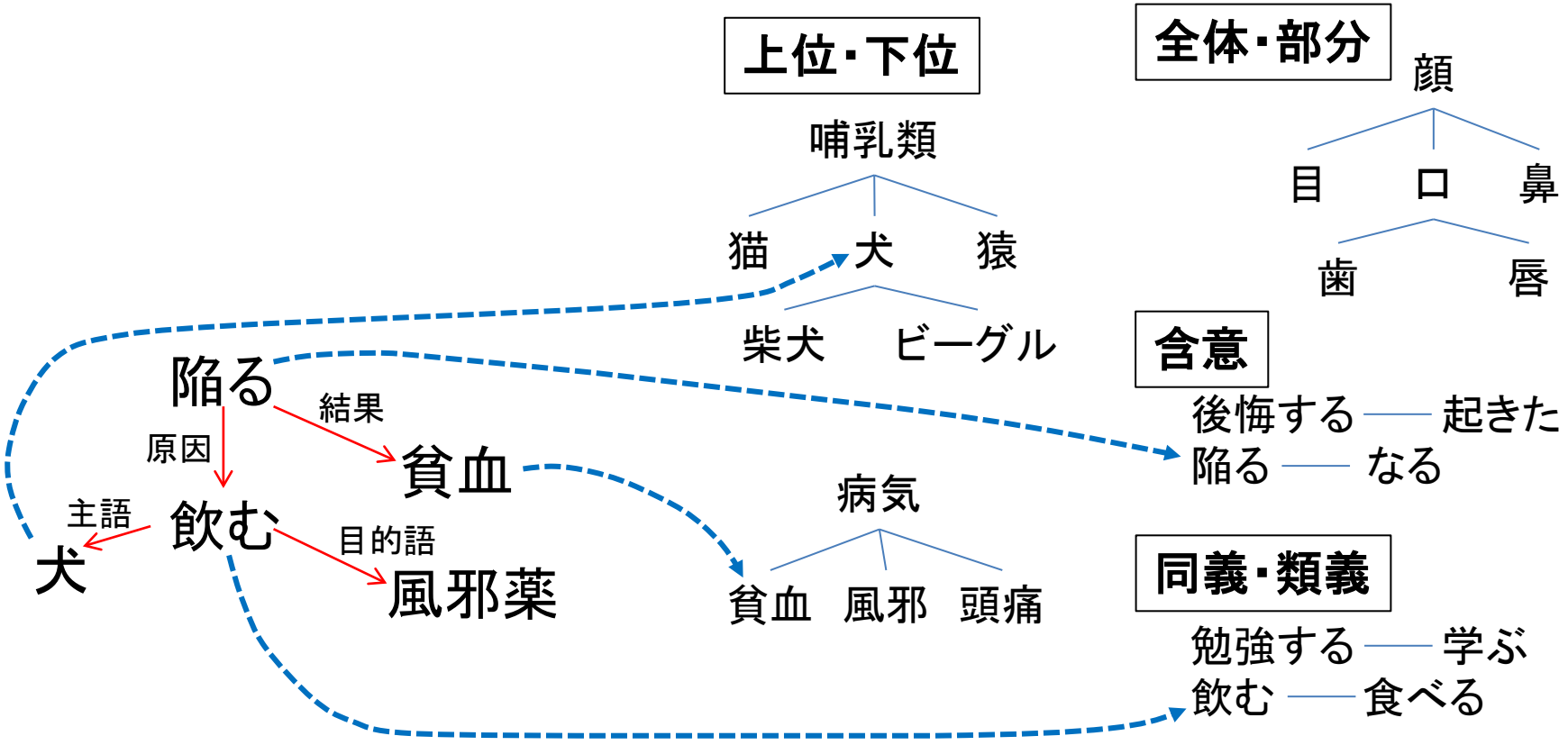




語彙的意味の計算

- 文中の単語をシソーラス・オントロジーにひもづける

→ 単語の置き換え可能性を表す



シソーラス・オントロジーをどうやって作る？

- シソーラス・オントロジーを作るのは大変
  - 世の中には無数の単語がある
  - 常に新しい単語が生まれている
    - スマホ、スーパークールビズ，…
- 大量のテキストから自動獲得する
  - ポイント：同じような意味の単語は同じような環境に現れる

朝見たらぽげらが真っ赤に熟していた。
おいしそうだったので、またぽげらを食べてしまった。
塩をちょっとかけたぽげらは激ウマだね。

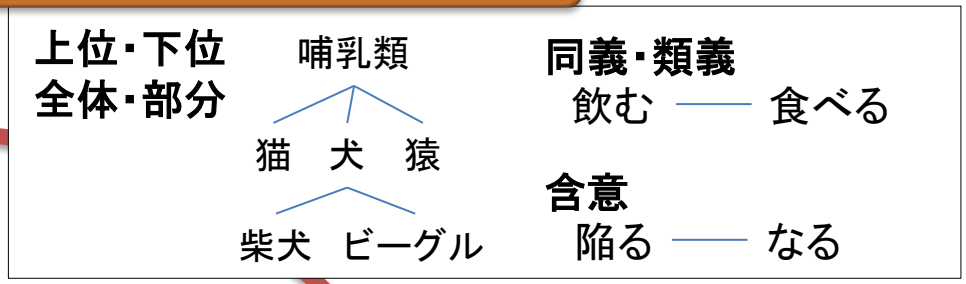
意味の計算

言葉

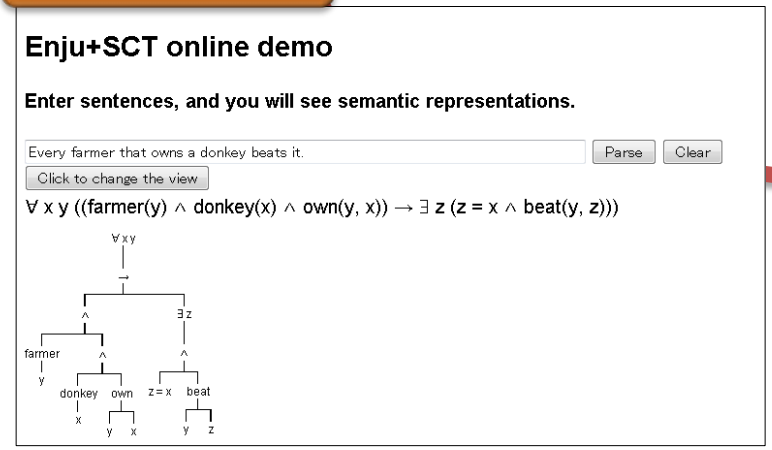
犬に風邪薬を飲ませると
貧血状態に陥ります。

シソーラス・オントロジー

語彙的意味を計算

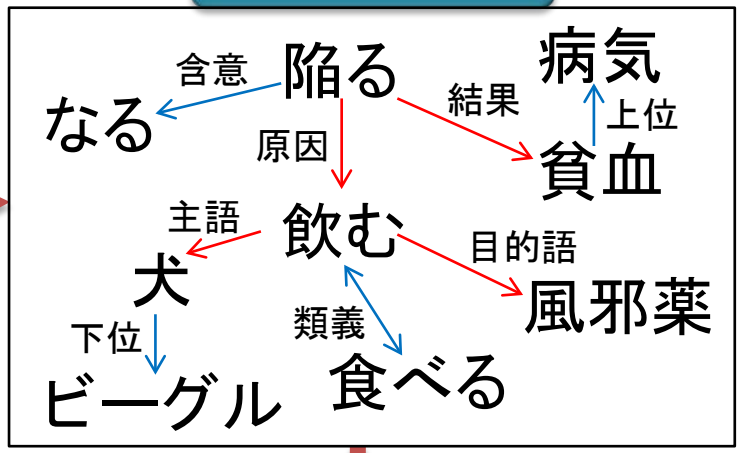


構文解析



構成的意味を計算

意味表現



ビーグルが風邪薬を食べたら病気になる

構成的意味と語彙的意味の相互作用

- 語彙的意味：単語の置き換え可能性を表す
- いつでも置き換えられるわけではない
  - 構成的意味により、置き換え可能性が変わる

犬に風邪薬を飲ませると貧血状態に陥ります。

下位

類義

上位

含意

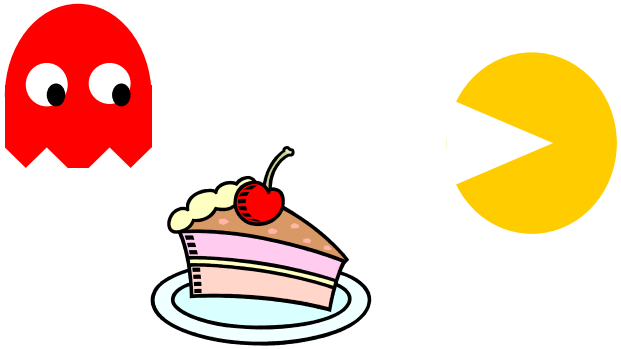
ビーグルが風邪薬を食べたら病気になる！

あいまい性

- あらゆるところで、言葉と意味との関係にはあいまい性がある

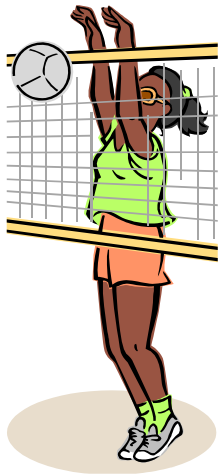
構成的意味のあいまい性

友達とケーキを食べた
せんべいとケーキを食べた



語彙的意味のあいまい性

今日はネットにつなげない
ボールがネットに引っ掛かった



あいまい性の問題

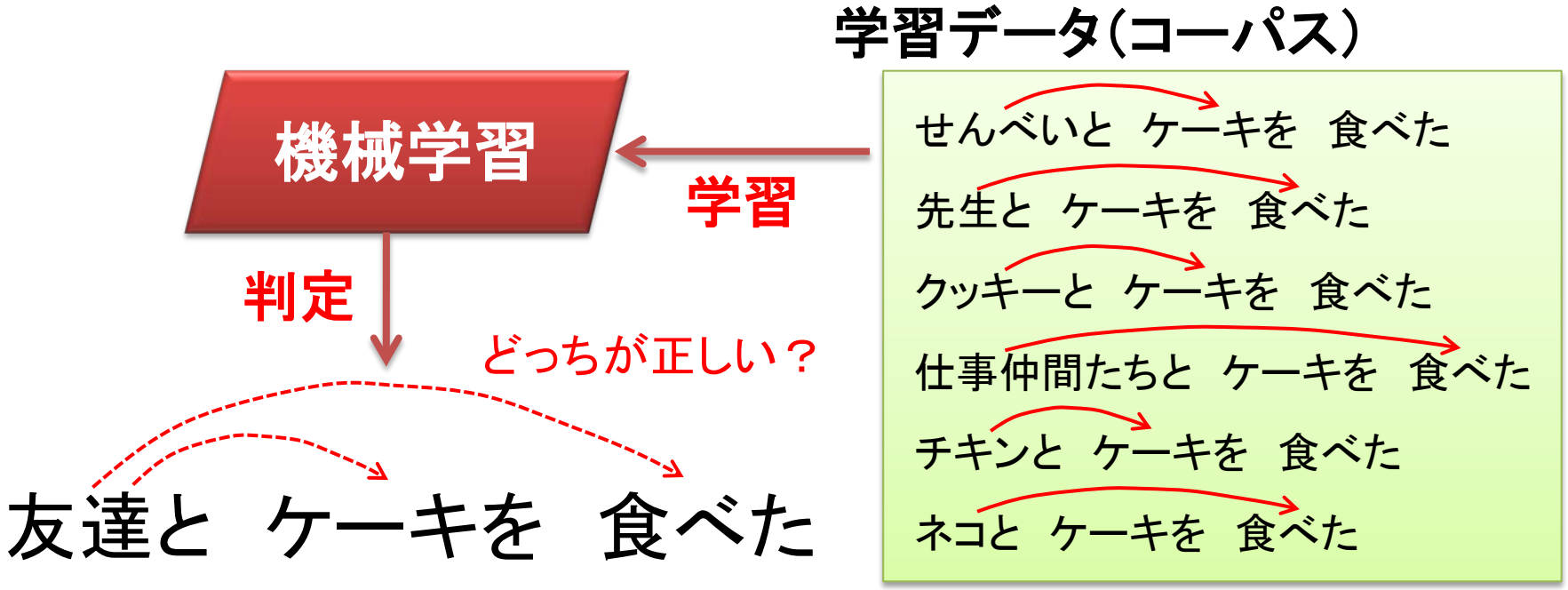
- あいまい性
  - 友達とケーキを食べた
  - せんべいとケーキを食べた
  - ネットで検索しよう
  - ネットで掃除しよう

ほとんど同じ文字列なのに意味的関係が違う

- 文脈がないと解決できない曖昧性もある
  - 大きな黒い瞳の女の子を見かけた

あいまい性解消

- いろいろな意味表現から、人間の解釈に最も近いものを選ぶ
- 機械学習：人間が作った学習データ（コーパス）から、規則性・傾向を自動的に学習する



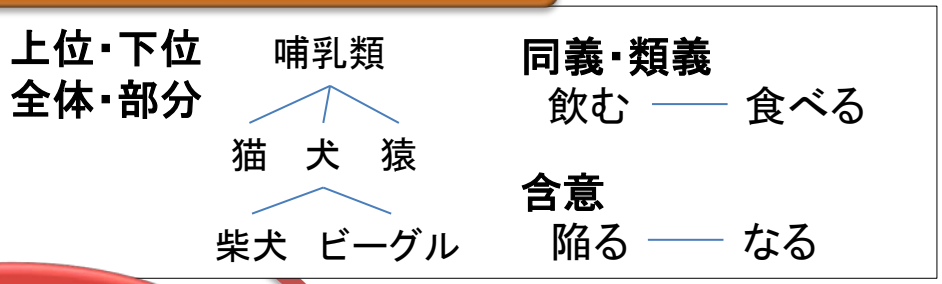
意味の計算

言葉

犬に風邪薬を飲ませると貧血状態に陥ります。

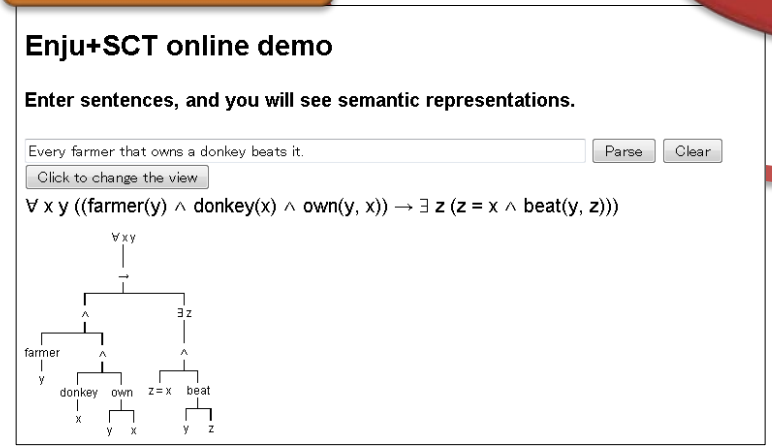
シソーラス・オントロジー

語彙的意味を計算



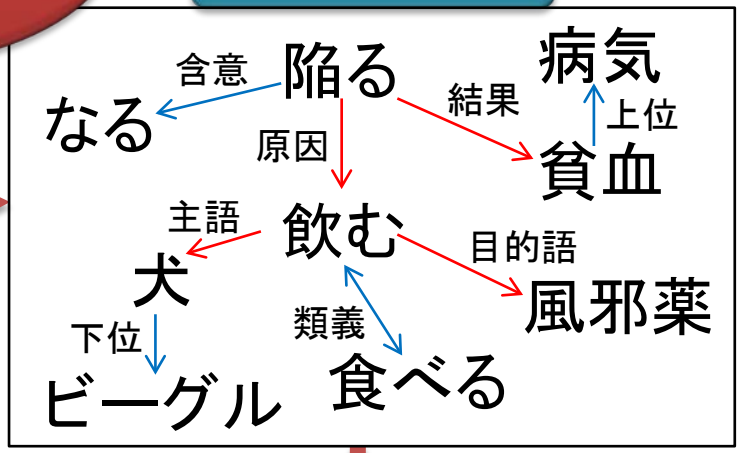
あいまい性解消

構文解析



構成的意味を計算

意味表現



ビーグルが風邪薬を食べたら病気になる

含意関係認識

- 二つの文の間に含意関係が成り立つかどうかを判定する問題
- 含意関係：ある文 t_1 が正しいとした時、もう一方の文 t_2 が正しいと言える

t_1: 犬に風邪薬を飲ませると貧血状態に陥る。
→ t_2: ビーグルが風邪薬を食べたら病気になる。

t_1: 川端康成は「雪国」などの作品でノーベル文学賞を受賞した。
→ t_2: 川端康成は「雪国」の著者である。

含意関係認識で
大学入試問題を解く

- 知識を問う問題＝教科書や参考書を見れば答えられる問題
- 教科書や参考書の内容をどれだけ記憶しているかが問われる
- 記憶している内容と問われている内容が同じ意味を表しているかどうかを認識する
  → 含意関係認識

問 6　下線部⑥に関連して，兵制や兵士について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。 6

① ポエニ戦争後，重装歩兵として従軍した農民層は経済的に豊かになった。
② 八旗は，順治帝が創設した軍隊である。
③ イェニチェリは，オスマン帝国の常備軍であった。
④ フランク王国では，テマ制（軍管区制）の下で屯田兵制が行われた。

問題

同じことを言っている？

教科書

イェニチェリ

イェニチェリ（Yeniçeri）は、14世紀から19世紀の初頭まで存在したオスマン帝国の常備歩兵軍団で、スプーンをシンボルにしていたことが知られている。常備軍団カプクルの中核をなし、火器で武装した最精鋭であった。トルコ語でイェニは「新しい」、チェリは「兵隊」を意味する。

オスマン帝国が拡大する過程で、従来の騎射を主戦術とするトルコ系軽騎兵の軍事力に頼らない君主の直属兵力として創設された。創始時期については諸説あるが、14世紀後半のムラト1世の治世とするのがもっとも

含意関係認識で
大学入試問題を解く

2009年度センター試験世界史B

兵制や兵士について述べた文として最も適当なものを，次のうちから一つ選べ。

① イェニチェリは，オスマン帝国の常備軍であった。

② フランク王国では，テマ制（軍管区制）の下で屯田兵制が行われた。

イェニチェリ
＝イェニチェリ軍団

オスマン帝国の常備軍
← 皇帝直属の常備軍

東地中海の強国 オスマ

．．．イェニチェリ軍団は，軍楽隊，工兵隊，大砲隊，鉄砲隊などをそなえた皇帝直属の常備軍で，のちにヨーロッパで発展する近代的陸軍の先がけであった。

ビザンツ帝国

．．．7世紀前半のヘラクレイオス1世は，アラブ軍の侵入にそなえて地方の体制を整えるためにテマ（軍管区）制を採用した。兵士には軍役とひきかえに世襲地が与えられ（屯田兵制），自作農の増加によって農村社会は活力にあふれるようになった。

東京書籍　平成19年度　世界史B

NTCIR

- 国立情報学研究所が主催している国際ワークショップ
- 自動翻訳や検索などの技術の評価のため、共有データを提供
- 国内外の研究グループが同じデータを用いて評価を行い、知見を共有する
- NTCIR において、センター試験を題材とした含意関係認識のデータを提供

大学入試にチャレンジ

- センター試験の選択肢とWikipediaを使い、含意関係認識の評価データを作成
- 対象科目
  - 世界史A・B、日本史A・B、政治経済、現代社会
- 6チームが参加
  - IBM東京基礎研究所
  - CMU（カーネギーメロン大学）
  - 京都大学
  - 東北大学
  - 北陸先端大学院大学
  - JUCS (Jadavpur University)

データの作り方

センター試験

問 6　下線部⑥に関連して，兵制や兵士について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。 6

① ボエニ戦争後，重装歩兵として従軍した農民層は経済的に豊かになった。
② 八旗は，順治帝が創設した軍隊である。
③ イェニチェリは，オスマン帝国の常備軍であった。
④ フランク王国では，テマ制（軍管区制）の下で屯田兵制が行われた。

Wikipedia

イェニチェリ

イェニチェリ（Yeniçeri）は、14世紀から19世紀の初頭まで存在したオスマン帝国の常備歩兵軍団で、スプーンをシンボルにしていたことが知られている。常備軍団カプクルの中核をなし、火器で武装した最精鋭であった。トルコ語でイェニは「新しい」、チェリは「兵隊」を意味する。

オスマン帝国が拡大する過程で、従来の騎射を主戦術とするトルコ系軽騎兵の軍事力に頼らない君主の直属兵力として創設された。創始時期については諸説あるが、14世紀後半のムラト1世の治世とするのがもっとも

含意関係：あり

t_1: イェニチェリは，14世紀から19世紀の初頭まで存在したオスマン帝国の常備歩兵軍団である。

t_2: イェニチェリは，オスマン帝国の常備軍であった。

実際のデータ

含意関係：あり
t1: パルテノン神殿は、古代ギリシア時代にアテナイのアクロポリスの上に建設された、アテナイの守護神であるアテーナーを祀る神殿である。
t2: パルテノン神殿の建つ丘は、アクロポリスと呼ばれている。

含意関係：なし
t1: パルテノン神殿は、ドーリア式神殿の最高傑作と言える作品である。
t2: パルテノン神殿は、ヘレニズム文化の影響下で建設された。

含意関係：あり
t1: スレイマン1世率いるオスマン帝国は絶頂期を迎えていた。
t2: オスマン帝国は、スレイマン１世の時代が最盛期であった。

- データ利用の覚書を提出すれば、利用可能

評価結果（試験の正答率）

	世界史A	世界史B	日本史A	日本史B	政治経済	現代社会	合計	正答率
IBM-1	18	13	6	6	10	11	64	57.7
IBM-2	14	11	6	5	10	10	56	50.5
IBM-3	11	10	3	1	6	9	40	36.0
CMU-1	8	8	1	2	2	4	25	22.5
CMU-2	10	12	3	5	6	7	43	38.7
CMU-3	10	12	10	4	7	11	54	48.6
京都-1	9	6	3	2	3	4	27	24.3
京都-2	9	14	5	5	10	12	55	49.5
京都-3	9	14	5	5	10	12	55	49.5
東北-1	12	8	2	4	4	3	33	29.7
北陸-1	8	10	3	2	12	11	46	41.4
北陸-2	8	9	7	4	5	10	43	38.7
北陸-3	7	13	7	7	10	10	54	48.6
JUCS	7	13	3	1	4	8	36	32.4
対象問題数	27	24	16	12	12	20	111	
全問題数	33	36	36	24	38	36	211	

意味に関わるその他の問題

- 時間、アスペクト、様相
  - お風呂が沸きました
  - お風呂が沸いています　} → お風呂に入れる
  - お風呂を沸かしました
  - お風呂を沸かしています → まだお風呂に入れない
- 参照関係
  - イェニチェリ軍団は，軍楽隊，工兵隊，大砲隊，鉄砲隊などをそなえた（オスマン帝国の）皇帝直属の常備軍で…
- メタファー、メトニミー
  - 官邸は首脳会談の日程を発表した。
  - つらい時期を乗り越えた。

これらの問題を解決してはじめて、
人間と同等の意味理解が可能になる

おわりに

- 言葉が表す意味を計算するためには、様々な自然言語処理技術を総動員する必要がある
  - 構文解析、シソーラス・オントロジー、あいまい性解消、…
- 現在最先端の意味処理技術で、センター試験をある程度解くことができる
  - しかし、人間と同等の意味理解にはまだ遠い
- 意味理解ができれば言葉を使いこなせるわけではない
  - 人間は、文の意味を理解してから、それに対する反応を考える
  - 意味処理は、言葉の理解に向けた第一歩

宿題

- 「のび太はお風呂に入っています」
  → のび太は風呂場にいる
- どうして分かるのか？

参考資料

- Speech and Language Processing
  Daniel Jurafsky & James Martin, Prentice Hall
- 言語処理学事典
  言語処理学会編　共立出版
- 言語と情報科学
  松本裕治(編)　朝倉書店