

共生社会特論

第4回 機械翻訳における諸問題

2016年12月20日

講義計画

11月29日 自然言語処理基礎

12月 6日 ルールベース機械翻訳

12月13日 統計的機械翻訳

12月18日 レポート締切

12月20日 機械翻訳における諸問題

1月17日 評価方法、レポート採点締切

1月24日 派生文法とウイグル語機械翻訳

1月31日 法令翻訳、レポート返却

レポート課題

書評

- 課題図書から1冊選ぶ
 - 早い者勝ち
- PDF形式1ページ
- 名前ありと名前なしの2ファイルを提出
 - 名前なしは採点用
- メールで提出：詳細は講義用サイト参照
- 書評の内容と、他人への採点・添削を評価

課題の補足説明

- 書評であり、概要でも感想でもない
 - 概要や感想を書いていけない訳ではない
- 講義用サイトにリストを掲載
 - 書評が提出された図書はリストから削除
- 後日、他人の書評を採点する
 - 一人あたり五人分を予定
 - 採点方法は後日説明

レポートの採点方法

- 自分と他5人のレポート、計6本を採点
 - レポート用紙上に **赤色のペン** で記入
- 採点基準
 - 0～10までの整数値の得点(8.5点などは駄目)
 - 複数のレポートに同じ点数を与えてはならない
 - ◇ 例外:0点のみ複数のレポートに与えてよい
 - 重大な不正(盗作等)がある場合は0点
 - 採点の理由を明記する
 - 不備な点を添削する
- 採点理由と添削の内容を、採点

配点

- 学生による採点結果 $10 \times 5 = 50$
- 教員によるレポート採点 20~30
- 採点理由の採点 10
 - 盗作の発見 50
- 添削内容の採点 上限なし

採点後のレポートは返却します

機械翻訳の問題

- 曖昧性
- 言語間の差異
- 状況依存
- 機械翻訳の利用の現状
- 意味理解

曖昧性 (Ambiguity)

- 単語の多義性 → 訳語選択

crane

take

- 品詞の曖昧性

Time flies like an arrow.

- 構文的曖昧性

John saw a girl with a telescope.

意味を考慮した規則や言語モデルで対応

辞書による曖昧性解消

A crane flies.

crane:

- 鶴
(class, 鳥)
- クレーン
(class, 機械)

fly:

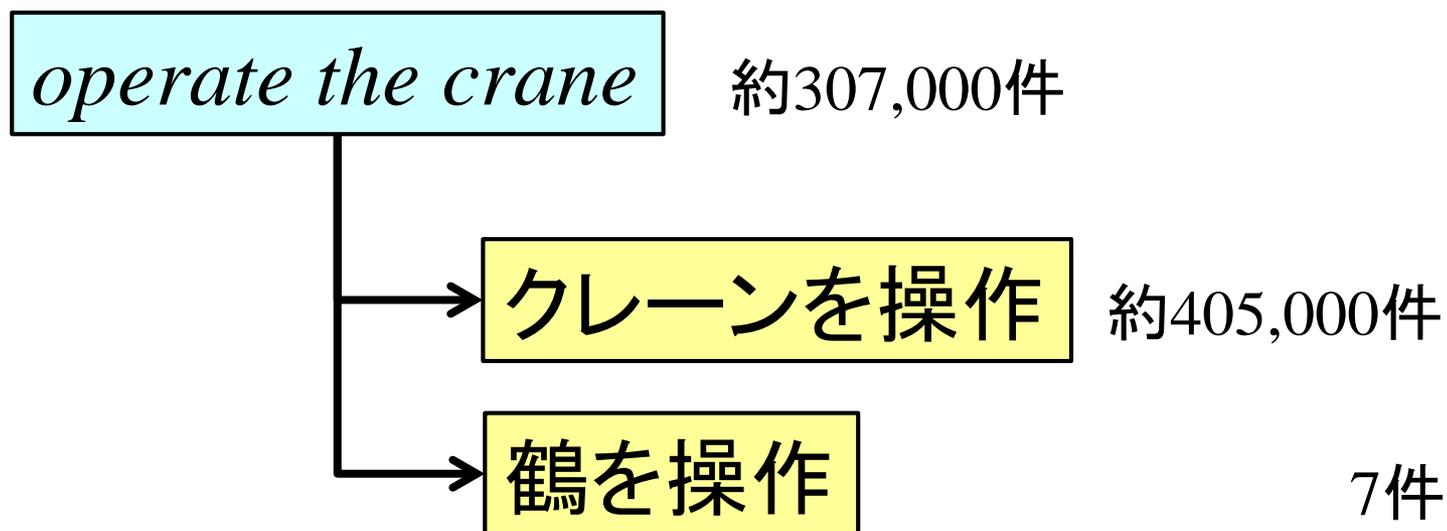
- 飛ぶ
(subj, {鳥, 航空機}, 動作主)

take:

- 撮る
(subj, 人間, 動作主)
(obj, 写真, 対象)
- 乗る
(subj, 人間, 動作主)
(obj, 乗り物, 対象)
- 飲む
(subj, 人間, 動作主)
(obj, 薬, 対象)

言語モデルによる曖昧性解消

- コーパス中の出現回数から訳語を選択

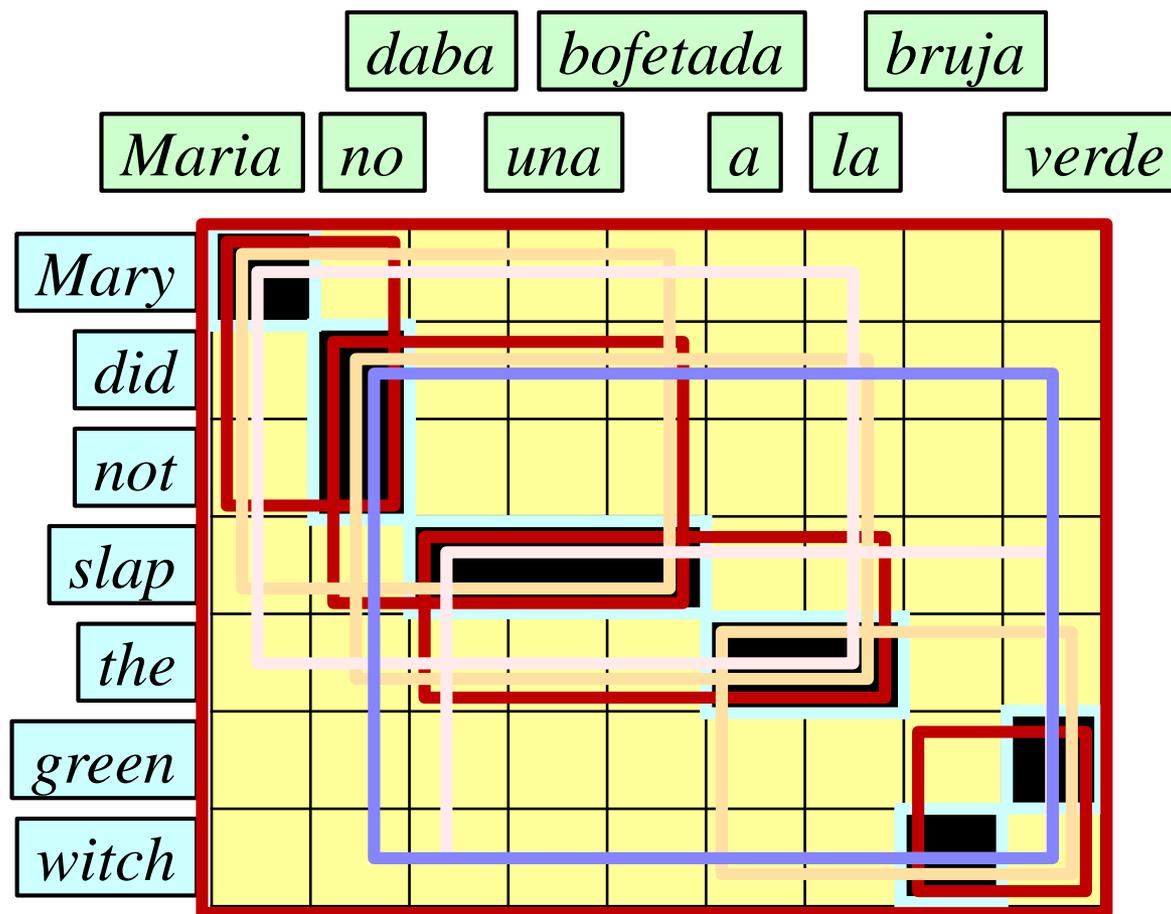


言語間の差異 語順

語順が大きく異なる言語間の翻訳

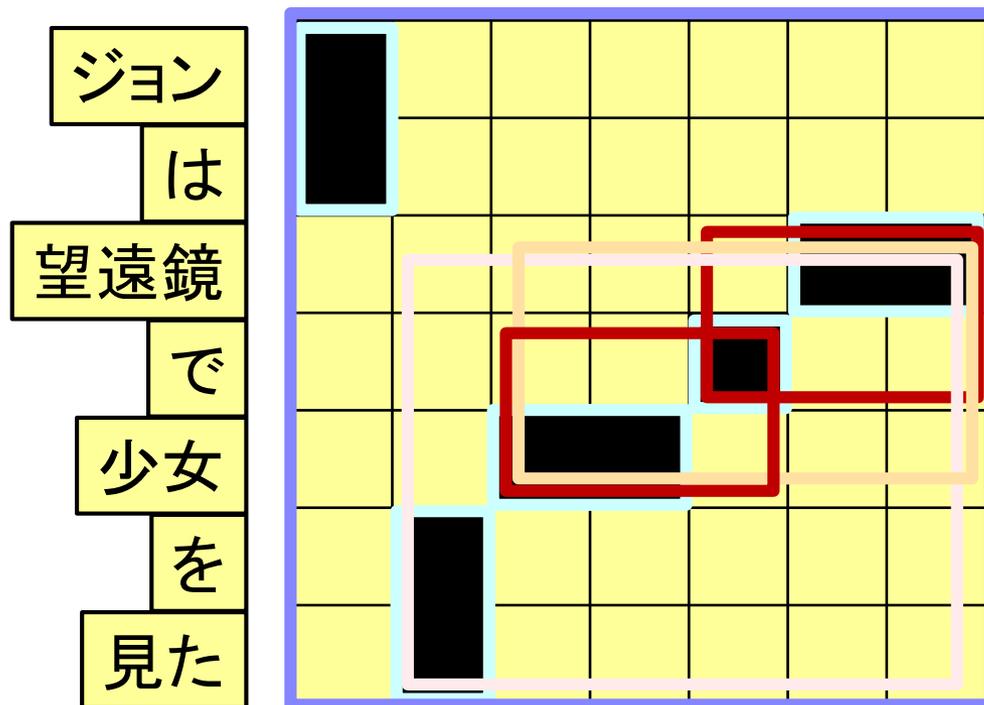
- 可能性のある語順が増える
 - 探索空間が膨大になる
- 句の対応が限られる

句の抽出 英語—西語の例



句の抽出 日英の例

John saw a girl with a telescope



句の対訳ペアの数が少ない

言語間の差異 単語の概念の差

- **lip** 唇だけでなく、その周辺、鼻の下も指す
 - *mustache*(口髭): hair on upper lip
- **皇帝** と *emperor*、**国王** と *king*
 - 元々は別物。「大王」「帝王」
- **弁護士**
 - 法制度が異なるため、対応する語が存在しない
 - 以下はいずれも厳密には異なる
 - ◇ lawyer, attorney, barrister, counselor, consultant

言語間の単語の粒度の差

- cattle 牛
- bull 雄牛
- cow 雌牛
- calf 子牛
- ox 雄の去勢牛 (使役用)
- steer 雄の去勢牛 (肉用)
- heifer 未経産牛
- rib
- fillet
- sirloin
- tenderloin
- plate
- chunk
- round
- shank

言語間の単語の粒度の差(2)

ウイグル語

- serke 去勢した雄ヤギ
- serkesh 去勢した二歳のヤギ
- seynek 一歳の雄ヤギ

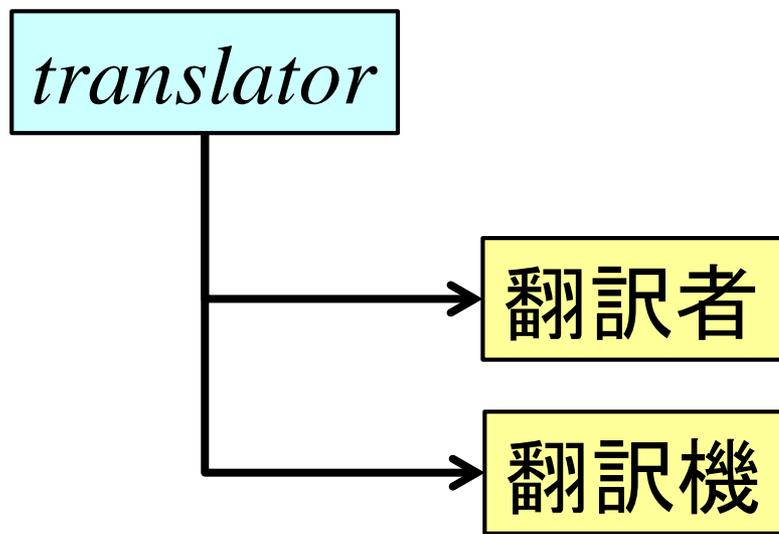
言語間の単語の粒度の差(3)

出世魚

	10cm	20cm	30cm	40cm	60cm	70cm	80cm	それ以上
関東		ワカシ	イナダ		ワラサ	メジロ	ブリ	ブリ
関西		ワカナ	ワカナ・ツバス	ツバス	ハマチ	メジロ	メジロ	ブリ
北陸	ツバス、ツバイン	ツバス	コズクラ	ハマチ	フクラギ	ガンド	ブリ	ブリ

英語では yellowtail というが、該当する魚は多数ある

言語間の単語の粒度の差(4)



相対的な日の呼び方

- 一昨々日 three days ago
two days before yesterday
- 一昨日 the day before yesterday
- 昨日 yesterday
- 今日 today
- 明日 tomorrow
- 明後日 the day after tomorrow
- 明々後日 two days after tomorrow

子孫の呼び方

- 子 child
- 孫 grandchild
- 曾孫 great-grandchild
- 玄孫 great-great-grandchild
- 来孫
- 昆孫
- 仍孫(じょうそん)
- 雲孫

親族を表す言葉

- 日本語

- 「伯父」「叔父」「伯母」「叔母」

- 中国語

- 「伯父」「叔父」「舅父」「姑母」「姨母」

- 英語

母方

父方

母方

- grandparent

- sibling

- ウイグル語

- 「親」に相当する単語がない

- ✧ 「父母」(ata-ana)はある

代表する語

- onion 玉ねぎ
- green onion ねぎ

情報量の差

私は世界中に友達がいる

I have a friend all over the world

誤訳

- 数の情報がない

U goshni yêdi.

- ウイグル語では三人称の代名詞が一種類

主語や目的語の省略

次章で詳細を説明します。

We explain the details in the next section.

スイッチを入れることができます。

You can turn on the switch

補う語が文脈や状況に依存

主語の役割の違い

The switch allows to run on lamp



スイッチでランプを点けることができます。

「スイッチが」とは言わない

概念の違い

ハンドルの前に座ってください



Sit down in front of the driving wheel.

誤訳

日本語の「前」の意味が曖昧

ハンドルの前に物を置かないでください

迷惑の受身

- 迷惑や被害を受身形で表す
- 自動詞の受身形もありえる
- 英語にはない表現

子供に泣かれた

雨に降られた

I was fallen by rain.

誤訳

状況依存

僕はウナギだ



I am the eel.

- 注文の時には間違い
- 料理が運ばれて来たときにはありえる

状況依存(2)



- 発話の日時に依存

文学作品

古池や蛙飛びこむ水の音

➤ 蛙は何匹？

The ancient pond
A frog leaps in
The sound of the water.
(Donald Keene)

The old mere!
A frog jumping in
The sound of water
(正岡子規)

Old pond — frogs jumped in — sound of water.
(Lafcadio Hearn)

その他の問題

- 分野依存
 - 分野ごとの対訳辞書の作成
- そもそも、原文が間違っている！

機械翻訳の利用の現状

- 情報受信型
 - 意味が分かればよい
 - 現状の機械翻訳でも役に立つ
- 情報発信型
 - 流暢な出力が要求される
 - 現在の機械翻訳ではまだまだ
 - 後編集が必要
 - ◇ 最初から人間が翻訳した方が早いという意見も

機械支援翻訳 (Computer Aided Translation)

- 前編集
 - 制限言語
- 後編集
- 翻訳メモリ
 - Trados
- 集合知
 - 「みんなの翻訳」

前編集

- 原文の間違いを修正（情報発信型）
- 原文を簡単化
 - 分割
 - 係り受けの修正
 - 曖昧な単語の言い換え
 - 制限言語
 - ◇ 規格化日本語

文章作成の基準（一部）

1. 一つの文には一つの事柄を書く
2. 1文を50文字以内に収める
3. 箇条書きを利用して簡潔に書く
4. 助詞を省略しない
5. 必要に応じて主語を明示する
6. 主語を述部を正しく対応させる
7. 主語以外に安易に「は」を付けない
8. 目的格の助詞には「が」より「を」を使う

文章作成の基準（続き）

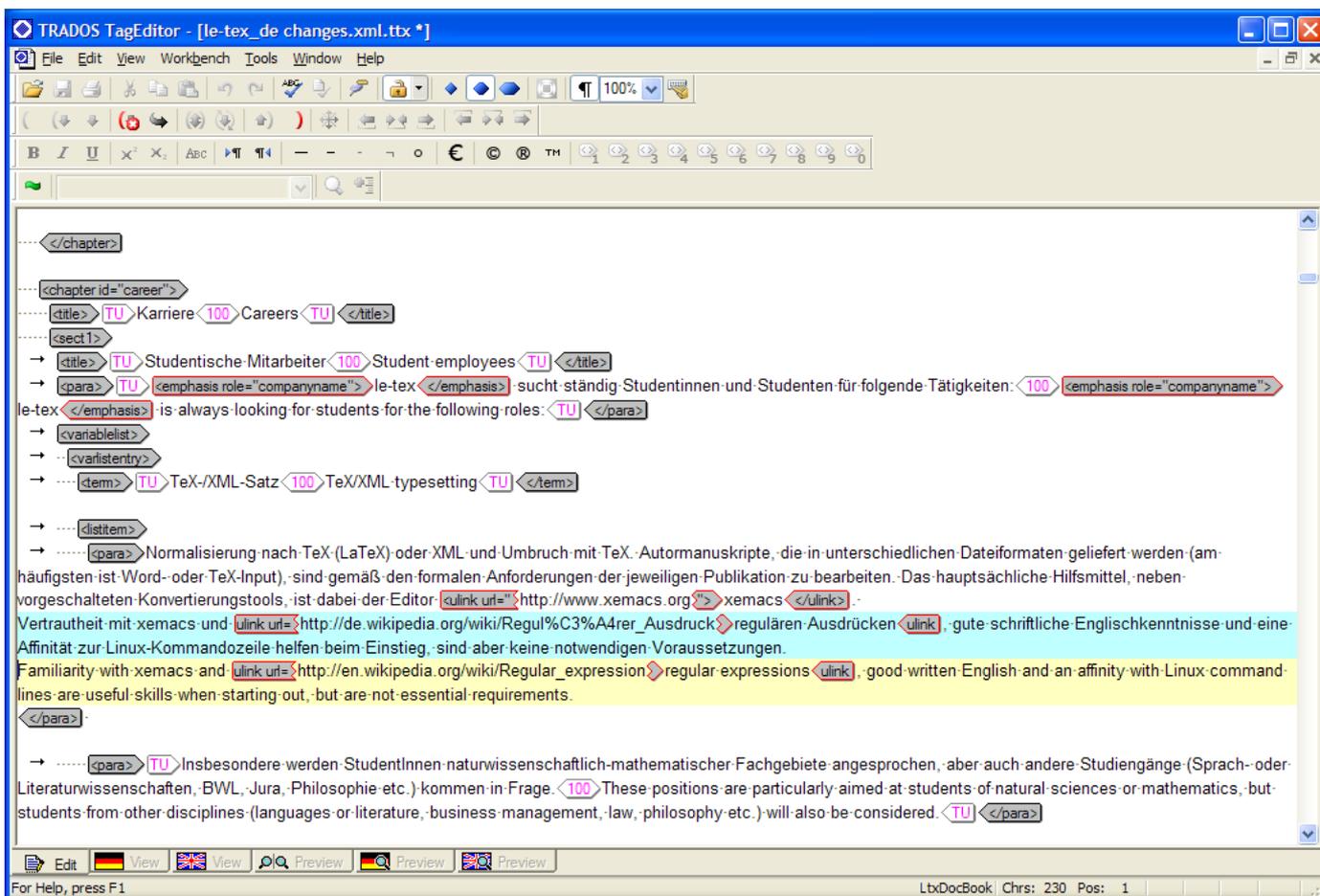
9. リスト内の項目のスタイルを統一する
10. 具体的な表現や直接的な表現を使う
11. 重複表現を省いて簡潔に書く
12. 正しい文法に沿って書く
13. 漢字で書くことが標準になっている言葉は漢字で書く
14. 誤字をなくす
15. 文中で記号を多用しない

後編集

- 機械翻訳の結果を修正
- 全部人手で翻訳するのとどちらが良いか？

翻訳メモリ

- 翻訳の再利用



みんなの翻訳

- ウェブ上での翻訳
- みんなの過去の翻訳を共有・再利用

<http://trans-aid.jp/>

翻訳と言語理解

Can machines think?

チューリング・テスト (Turing Test)

(Turing '50)

- 判定者から見えない場所に人かコンピュータを配置
- 判定者は対話を通じて、相手が人かコンピュータかを判定
- 判定者が区別できない場合、このコンピュータはテストに合格
- 2000年までに5分間の会話で30%の審査員を騙せるようになると予測

チューリングテスト 続き

- 機械に知能があるか否かを判定
- 行動主義的
 - 機械の内部に立ち入らない

人工無能 (chatterbot)

- 人間と対話するプログラム
- チューリング・テストに挑戦
- 「人間らしい」モデルを構築
 - 相手の発話中のキーワードと内部ルールとのマッチングでそれらしい返答を作成
 - ◇ 「〇〇とは何ですか？」
 - 学習するものもある
- Twitter上のBOTなど

ELIZA (Weizenbaum '66)

- 最初の人工無能
- セラピストの対応を模倣
- デモ

<http://www.manifestation.com/neurotoys/eliza.php3>

<http://shower.human.waseda.ac.jp/~m-kouki/cgi-bin/eliza/>

人工痴能うずら

• IRC上の人工無能（2000年の会話例）

その1(2000/11/11)

05:09 <#うにつくす:Hisasa> こんにちは、うずらさん

05:09 >#うにつくす:uzura< てすと勉強おしまーい

05:09 <#うにつくす:Hisasa> 難しいテストですか？

05:10 <#うにつくす:Hisasa> うずら、うずら

05:11 <#うにつくす:Hisasa> うずらさん、お返事して

05:11 >#うにつくす:uzura< ひゆるりら～

05:11 <#うにつくす:Ken02work> にんきものね

05:12 >#うにつくす:uzura< 片手でもてるパソコンほしいのよね

05:12 <#うにつくす:Hisasa> うずら、あまり賢くない...

05:15 <#うにつくす:Hisasa> うずら、うにつくすって何？

05:15 >#うにつくす:uzura< うにつくすだと、IRCクライアントは何がおすすめですか

05:17 <#うにつくす:Hisasa> ぎょ～、うずらにシカトされてる

05:17 >#うにつくす:uzura< うひん。

05:35 <#うにつくす:Excel> うずらって基本的に何＝何で覚えてるわけじゃないから、
そゆ反応無理だと思うんですが...(^^;

05:35 >#うにつくす:uzura< む 私は何に反応してるのだ

05:37 <#うにつくす:Ken02work> すごい返した

05:37 <#うにつくす:Ken02work> 人間でもかなわんかも..

人工痴能うずら

• 会話例

- 06:12 <#うにつくす:uzura> ext2 から脱出の旅とか(ムリ
- 06:13 <#うにつくす:lumin> ドキュメント読めばね(笑
- 06:13 <#うにつくす:ririn> uzura のは昔の私の発言だ... ^^;
- 06:13 <#うにつくす:ririn> 盗聴記録自体はだれでも触れちゃうですきゃ。
- 06:16 <#うにつくす:lumin> 会社の中からはならね。
- 06:17 <#うにつくす:ririn> なるほど。
- 06:17 <#うにつくす:uzura> 私って基本的に何＝何で覚えてるわけじゃないから、
そゆ反応無理だと思うんですが...(^^;
- 06:17 <#うにつくす:HMX-12> うひひ
- 06:17 <#うにつくす:ririn> 自分のシステムを説明できるとは。
- 06:20 <#うにつくす:Mjnk> ぼっとですか? >うずら
- 06:21 <#うにつくす:lumin> トリ人間だよ
- 06:21 <#うにつくす:uzura> まあ～ポットとかの仲間かな
- 06:21 <#うにつくす:HMX-12> うひひ
- 06:21 <#うにつくす:kuroyagi> ポットなのか...(^^;

ローブナー賞

- チューリングテストのコンテスト
 - 1991年開始
 - 人間とチャットボットと5分間ずつ対話
 - 審査員は、どちらが人間か判定する
 - 30%の審査員を騙せたら合格

 - 「最も人間らしいコンピュータ賞」
(Most Human Computer)
 - 「最も人間らしい人間賞」(Most Human Human)

Eugene Goostman

- 13歳のウクライナの少年という設定
- 2014年のローブナー賞にて
33%の審査員が人間と判定

中国語の部屋 (Chinese Room)

(Searle '90)

- 部屋の中に中国語が分からない人を配置
- 部屋には英語から中国語への変換規則が書かれた本がある
- 部屋の中の人、この本に従い、
入力された文を中国語に翻訳して返却

チューリング・テストに合格するが
知性を持たない

チューリング・テストに対する反論

強いAIと弱いAI

- 強いAI

- 考えることができ、心を持つAI
- 知性と推論能力を持つ

- 弱いAI

- 考えたり心を持っているように振る舞うAI
- 実用的な問題を解決する

コンピュータは「頭が良い」か？

- 「頭の良さ」とは何か？
 - 計算が速いこと？
 - 知識が豊富なこと？
 - 質問に答えられること？
 - チェスや将棋に強いこと？
 - 試験問題が解けること？
- コンピュータは人間より「頭が良い」か？
 - AIの分野では、コンピュータが人間を超えると「それは頭の良さを示さない」と言われる