

アラジン・フォーラムが提供する 言語処理ツールと言語データ



2011/03/02

独立行政法人情報通信研究機構

言語翻訳G グループリーダー

隅田 英一郎

アラジン (=高度言語情報融合) フォーラム

2009年3月設立



辻井潤一 会長

<http://alagin.jp/>

フォーラムのねらい

- コミュニケーションには**3つの壁**がある

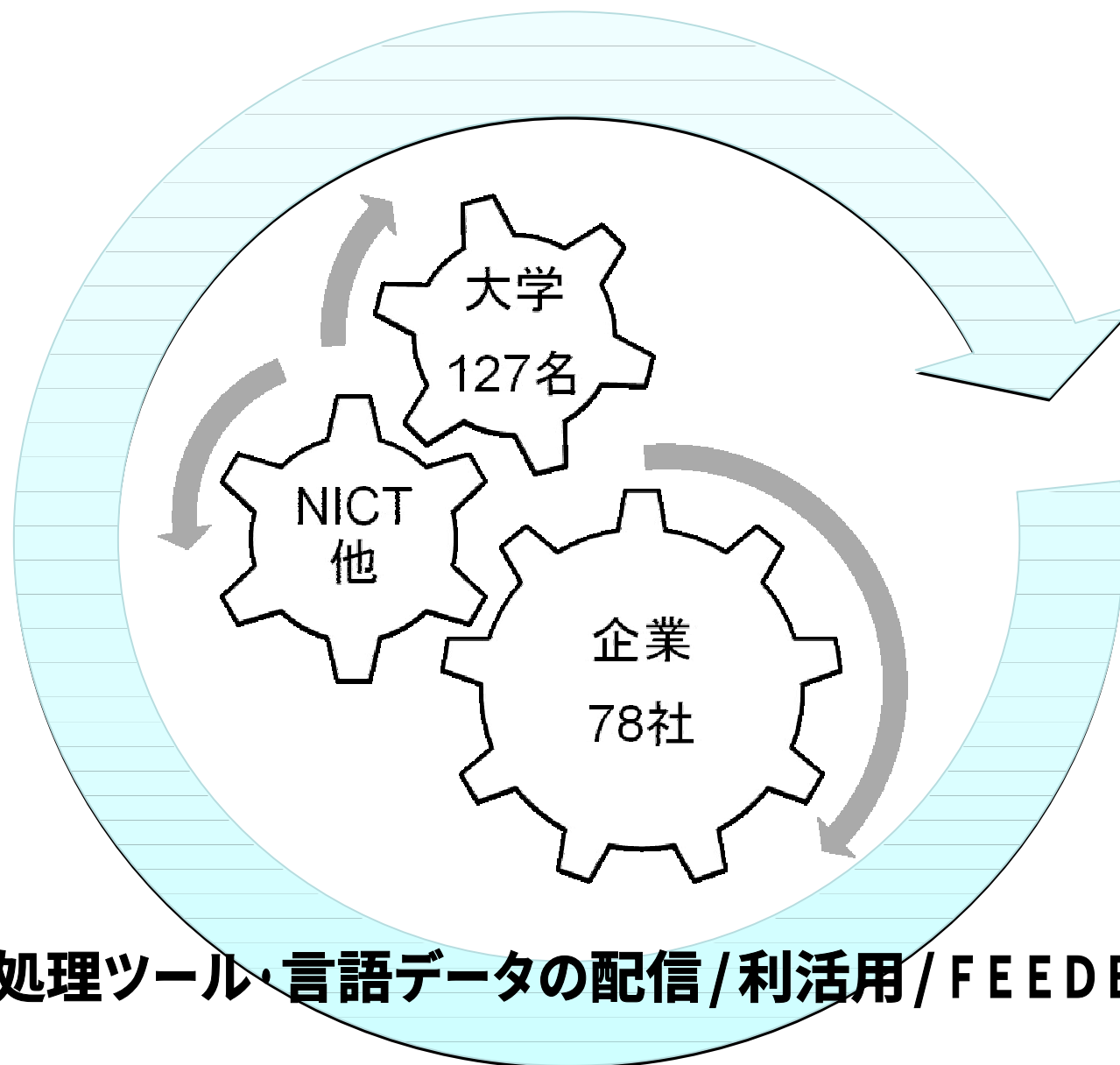
言語の壁：外国語で話されたり、書かれた情報は理解不能！機械と話が通じない！

情報の量の壁：Web上の大量のマルチメディア情報を消化できず本当に有用な情報が見つけられない

情報の質の壁：Web上にはウソも一杯！

- 産学官の協力体制としてフォーラムを組織し、この**3つの壁を乗り越える**。

産学官の連携



言語処理ツール・言語データの配信 / 利活用 / FEEDBACK

ツールやデータの契約総数=413

言語資源サービス配
信サイトを運用

言語データ

概念辞書データ/ツール

文脈類似語データベース

意味的関係抽出器

単語クラス作成器

動詞含意関係データ

上位下位関係抽出ツール

負担・トラブル表現リスト

日本語WordNet

上位語階層データ

日本語異表記判定ツール

日本語係り受けDB

日本語パターン言い換えDB

日本語異表記DB

基本的意味関係の事例ベース

言語処理ツール

世界最高精度の中国語、タイ語等の形態素解析器

カスタム単語集合作成サポートサービス

世界最高精度の中国語構文解析器

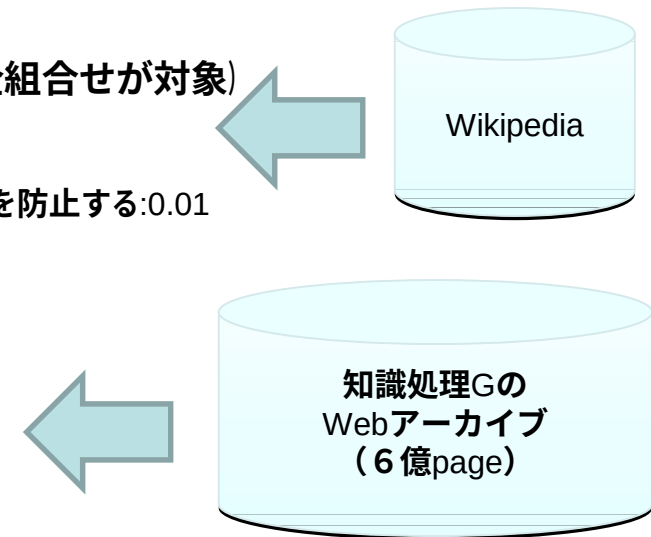
意味的関係抽出サービス

上位下位関係抽出ツール

異表記判定ツール

配信済みの言語資源: Web数億ページやWikipediaからの自動獲得+人手検証

1. 文脈類似語データベース (名詞100万語が対象)
 - ・ チャイコフスキー → ブラームス (-0.152) シューマン (-0.163) メンデルスゾーン (-0.166)
2. 動詞含意関係データベース (含意ペア約3万を収録)
 - ・ チンする → 加熱する, デトックスする → 解毒する, 先発復帰する → スタメン出場する
3. 負担・トラブル表現リスト (約2万表現を収録)
 - ・ 風邪 → 病, カルテ改ざん → 不正行為・違反, ハム音 → 異音・雑音・騒音, ハモグリバエ → 有害生物-虫
4. 上位語階層データ (約7万の名詞句を階層化)
 - ・ <type=D>台</>, <type=G>天文台</>, <type=G>公開天文台</>
5. 単語共起頻度データベース (Web一億ページを対象、約100万語の全組合せが対象)
 - ・ 海外旅行 → 国内旅行:0.37, 格安航空券:0.16, ツアー:0.16, 航空券:0.15,
6. 日本語パターン言い換えデータベース (約170万パターン)
 - ・ AはBを防ぐ → AがBを防ぐ:0.02, AでBを防ぐ:0.01, Bを防ぐA:0.01, AはBを防止する:0.01
7. 日本語異表記対データベース (正例約5万ペアを収録)
 - ・ ゴミ置き場 - ゴミ置場, ギタープレー - ギタープレイ, お替わり - おかわり,
8. 日本語係り受けデータベース (係り受け約8億種類を収録)
 - ・ (関サバを食べる 20回) (関サバのお造り 7回)
9. 基本的意味関係の事例ベース (約1.7万ペアを収録)
 - ・ 同義: 狂牛病 - BSE, 対義: 夜 - 朝, 免税事業者 - 課税事業者
10. 日本語WordNet (9.4万語、5.7万概念を収録)
 - ・ 0206897-n → 海豚, ドルフィン, イルカ
11. Wikipedia対訳コーパス (約50万文ペアを収録)

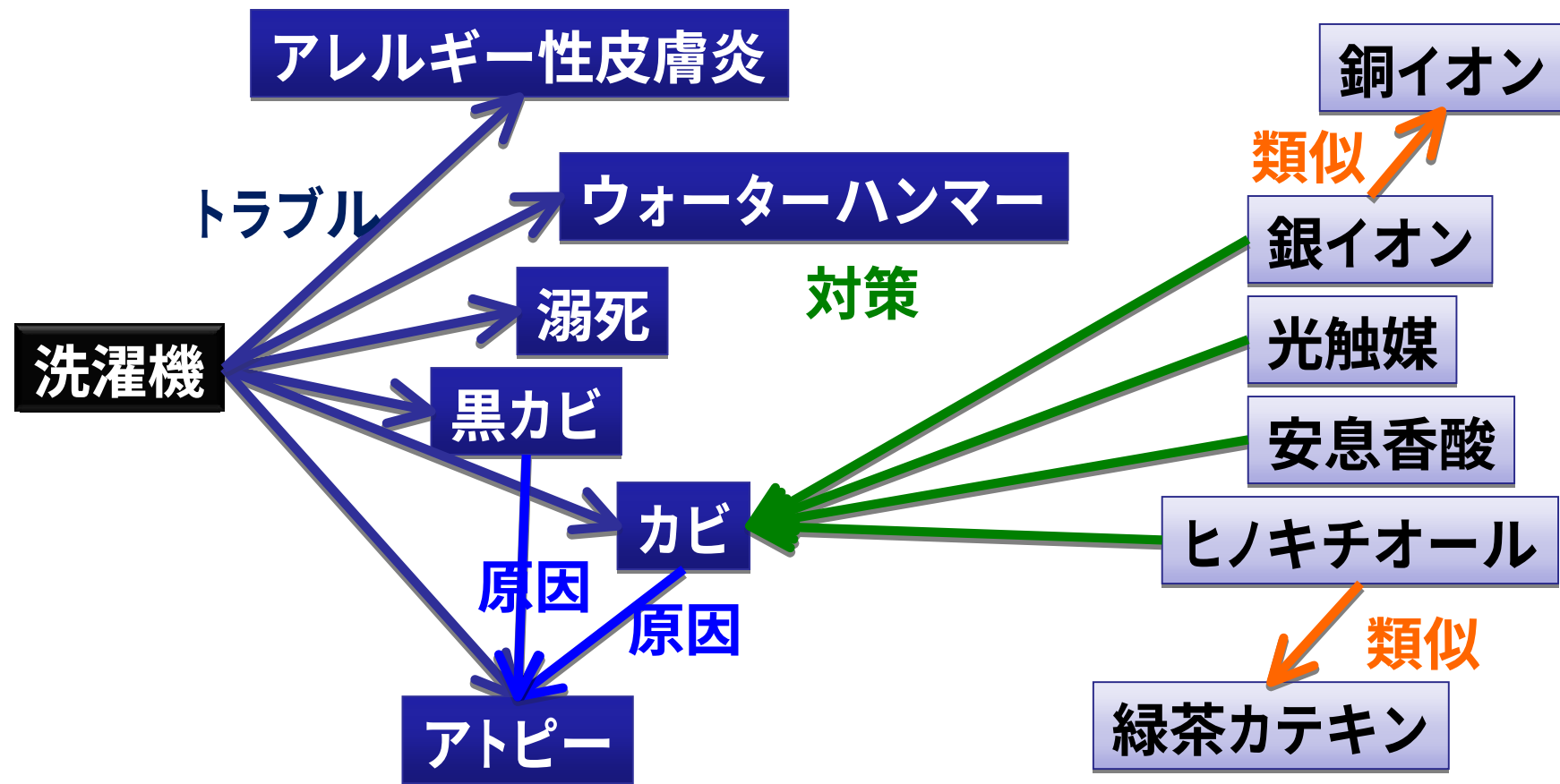


Webサービス、ツール

12. カスタム単語集作成サポートサービス	14. 中国語形態素解析システム *世界最高性能! (近日公開予定)	16. 上位下位関係抽出ツール
13. 意味的關係抽出サービス	15. 中国語係り受け解析システム *世界最高性能!	17. 異表記判定ツール (近日公開予定)

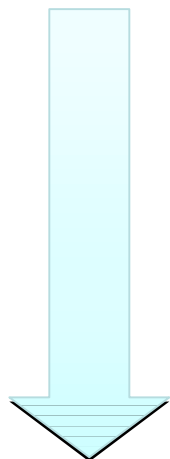
概念辞書＝単語の意味のネットワーク

約250万語をカバー



世界最高レベルの高精度な多言語解析ツール

布朗一行于今晚离沪赴广州。

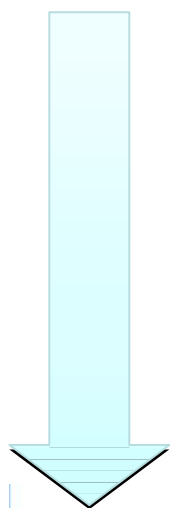


形態素解析

- ハイブリッドモデル
- MIRAアルゴリズムによる学習
- エラー駆動の未知語処理

手法	精度 (F値)
N&U 07	93.32
Jiang 08	93.41
提案手法	93.67

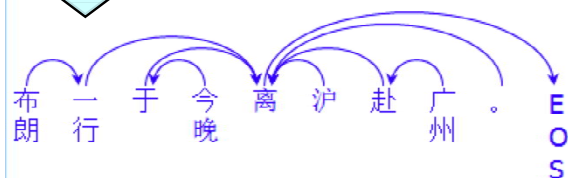
布朗/NR 一行/NN 于/P 今晚/NN ...



係り受け解析

- グラフベース解析
- 大規模コーパスからの教師なし学習
- 対訳コーパスの利用履歴素性

手法	精度(%)	文正解 (%)
MST (ベース)	88.18	47.12
提案手法	89.91	48.56
最新システム	91.93	55.45



ALAGINの活動：講習会・シンポ等（その1）



開催日	イベント名	
H22.7.1 ～ 7.2	技術開発部会： 自然言語処理分科会 ・第1回自然言語処理分科会セミナー	鶴岡慶雅（北陸先端科学技術大学院大学）
H22.8.24 ～ 8.27	技術開発部会：音声処理分科会 ・音声認識・音声対話技術講習会	河原達也（音声処理分科会長、京大）、西村竜一（和歌山大）、 篠崎隆宏（東工大）、李晃伸（名工大）、山下洋一（立命館大）、 荒木雅弘（京都工繊大）、高橋徹（京大）、秋田祐哉（京大）、 酒向慎司（名工大）、駒谷和範（京大）、西本卓也（東大）、翠 輝久（NICT）
H22.9.1	普及促進部会 ・第1回普及促進部会	奥村明俊（普及促進部会長、NEC）、中村哲（NICT）、高木浩 光（産業技術総合研究所）、木村俊也（ミクシィ）、平井聖一 （インターグループ）
H22.9.27 ～ 9.28	技術開発部会： 自然言語処理分科会 ・第2回自然言語処理分科会セミナー	宮尾祐介（国立情報学研究所）

ALAGINの活動：講習会・シンポ等（その2）

開催日	イベント名	講師/講演者	内容
H23.1.7	技術開発部会：自然言語処理分科会 ・第3回自然言語処理分科会セミナー	吉永直樹 (東京大学)	大規模 Webテキストを対象としたアプリケーション (ブログの話題分析など) を実現するために必要となる「機械学習」の分類器の学習・テストの高速化手法の紹介。
H23.1.21	技術開発部会：音声処理分科会 ・音声処理分科会セミナー	河原達也 (音声処理分科会長、京都大学)、堀貴明 (NTT)、Paul Dixon (NICT)	音声認識デコーダの最近の動向についての紹介。 近年最も注目を集めている方式「T3デコーダ (WFST: 有限状態トランスデューサ)」の利用法に関するチュートリアル。
H23.2.7	企画推進委員会 ・国際標準化活動のご紹介 ～講演会とディスカッション～	中村哲 (企画推進委員長、NICT)、堀智織 (NICT)、芦村和幸 (W3C/慶應)	部会横断的な活動として「標準化」をテーマに、音声関連技術を対象にした国際標準化活動への取り組みを紹介。今後のALAGIN標準化への取り組みについてディスカッションを行う。
H23.2.28	普及促進部会： ・第3回普及促進部会	中村哲 (NICT)、関根聡 (普及促進副部会長、楽天/ニューヨーク大学)、山口英 (奈良先端科学技術大学院大学)、山田恒夫 (放送大学)	ALAGIN活動・成果の紹介と講演。講演タイトル：「楽天の自然言語処理技術に関連したサービスの研究開発」、「最近の情報セキュリティの状況と課題」、「ITを利用した言語教育とその教材開発の展望」

アラジンのツールや言語資源で何ができるか？

