自然言語処理の基本

阿部 泰之

自己紹介

- 阿部 泰之　/ Hiroyuki Abe

- 業務エンジニア
  （生命保険　主に保険金支払）

- twitter　　/ @taki_tflare
- https://tflare.com

そもそも自然言語処理とは

自然言語とは日本語や英語のこと

マークアップ言語やプログラミング言語と区別するために存在する言葉

自然言語とコンピュータ言語の違いは曖昧性の有無

自然言語には曖昧性が存在する。

自然言語処理の応用例

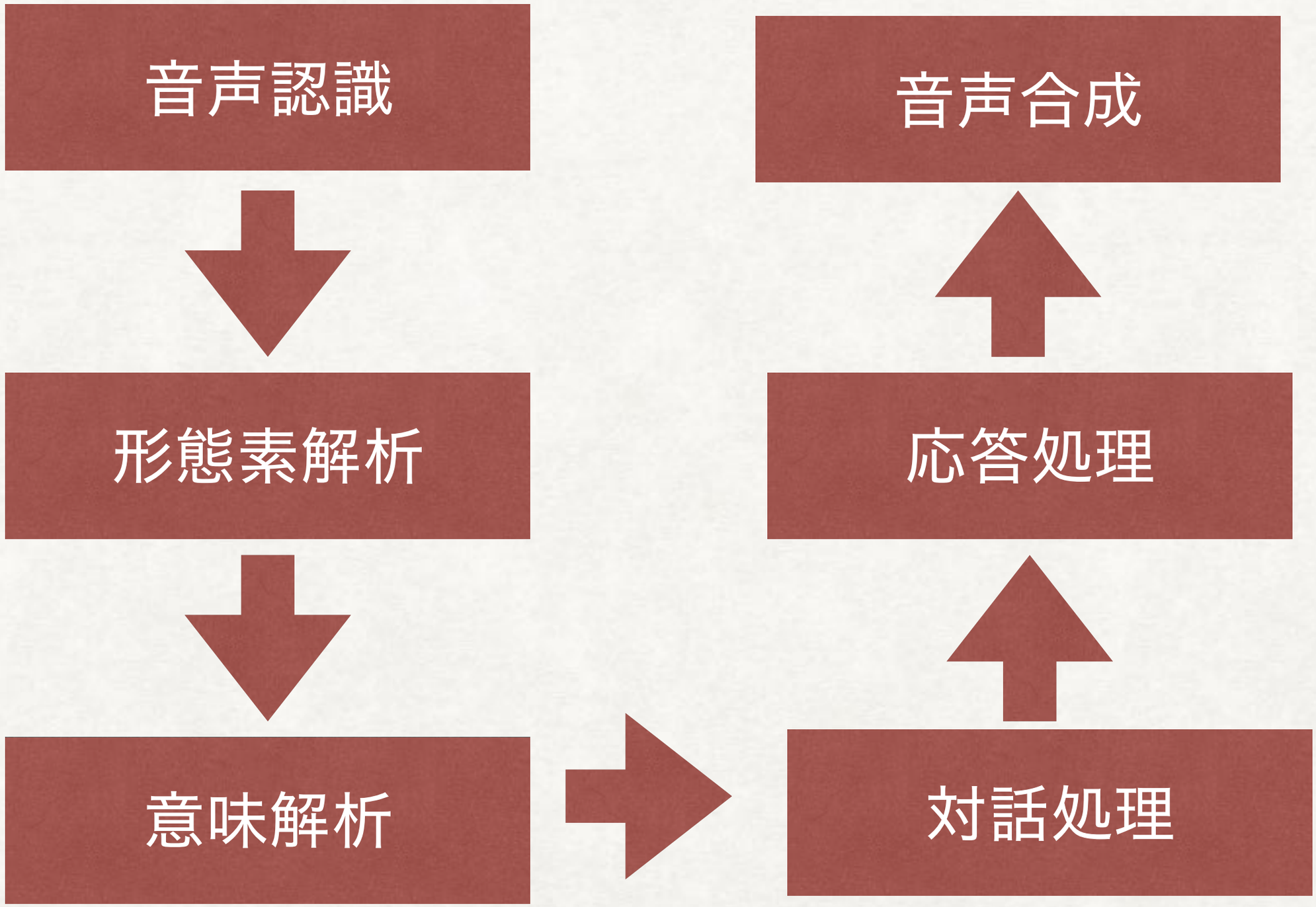
- 検索エンジン
- 機械翻訳

  ある言語を別の言語に翻訳する

- 対話システム

  Siriなど

対話システムの難しさ



対話システムの難しさ

Mac版 Siriでの具体例 (Mac OS X 10.13.5)

「17時頃の天気予報」とSiriに言うと

→表示されない。「すみません。どこにあるかわかりません。」と応答される。

「17時の天気予報」とSiriに言うと

→「今日の午後の天気予報です。」と応答し、

17時の天気（周辺のもの）が表示される。

形態素解析とは

コーパスとは

コーパス（英: corpus）は、言語学において、自然言語処理の研究に用いるため、自然言語の文章を構造化し大規模に集積したもの。構造化し、言語的な情報（品詞、統語構造など）を付与している。

Wikipediaより

均衡コーパスとは

BCCWJ は現代日本語の均衡コーパス（balanced corpus）である。現代日本語書き言葉のできるだけ多くの変種をとりあげ、日本語の全体像を明らかにするための偏りのないサンプルを提供することを目標とした設計が施されている

均衡コーパスの説明に変更

http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_01.pdf　より

現代日本語書き言葉均衡コーパスとは

『現代日本語書き言葉均衡コーパス』(BCCWJ)は、現代日本語の書き言葉の全体像を把握するために構築したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスです。書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などのジャンルにまたがって１億430万語のデータを格納しており、各ジャンルについて無作為にサンプルを抽出しています。

http://pj.ninjal.ac.jp/corpus_center/bccwj/　より

現代日本語書き言葉均衡コーパスとは

全文検索専用のインターフェースは『少納言』（http://www.kotonoha.gr.jp/shonagon/）、

形態素解析済データ検索用のインターフェースは 『中納言』（https://chunagon.ninjal.ac.jp/）と呼ばれている。

http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_01.pdf　より

辞書とは