1. 文章要約
2. 文章分類
3. 固有表現抽出

固有表現抽出（Named Entity Recognition, NER）とは、テキストから特定のカテゴリに属する単語やフレーズ（固有名詞や専門用語など）を識別し、抽出する技術のことです。例えば人名、地名、組織名、日付、金額などの情報を抽出することができます。

「昨日、ジョンは東京で開催されたコンサートに行った。」という文から、ジョン（人名）、東京（地名）、昨日（日付）といった情報を抽出します。情報抽出：ニュース記事やウェブページから重要な情報（人物名、地名、日付、イベントなど）を抽出し、データベースに保存したり、要約を作成するために使用します。

顧客サポート：顧客からのメールやチャットで言及された商品名、問題の詳細、連絡先情報などを抽出し、効率的なサポート対応を行うために使用します。

ソーシャルメディア分析：ツイートやブログ記事からブランド名、製品名、競合他社、感情などを抽出し、マーケティングや市場調査に活用します。

バイオインフォマティクス：研究論文から遺伝子名、タンパク質名、疾患名、薬物名などを抽出し、データベースの構築や研究の助けとなる情報を提供します。

自動タグ付け：ウェブページや文書からキーワードやカテゴリを抽出し、自動的にタグ付けを行うことで、コンテンツの管理や検索性を向上させます。

暗号資産取引：取引所やニュースサイトから暗号資産名、価格変動、関連ニュースなどを抽出し、投資判断の助けとなる情報を提供します。

法的文書解析：契約書や裁判文書

約定情報の例

1. 文の類似度判定

文の類似度判定とは、二つの文がどれだけ似ているかを数値化することです。例えば、「私は犬が好きです」と「私は猫が好きです」は、主語と述語が同じで、目的語だけが異なるだけなので、ある程度類似していると判断できます。一方、「私は犬が好きです」と「今日は雨が降っています」は、主語も述語も目的語も異なるので、あまり類似していないと判断できます。

文の類似度判定にはさまざまな手法がありますが、代表的なものに以下のようなものがあります。単語分散表現を用いた手法文中に含まれる単語をベクトル化し、そのベクトル間の距離や角度で類似度を測定する。

例：コサイン類似度、Word Mover’s Distance

文分散表現を用いた手法

文全体をベクトル化し、そのベクトル間の距離や角度で類似度を測定する。

例：Doc2Vec、BERT

統計的手法

文中に含まれる単語や文字列の出現頻度や重み付けなどを利用して類似度を測定する。

例：TF-IDF、Jaccard係数

文の類似度判定は、情報検索やテキストマイニングなどの応用に役立ちます。例えば、

情報検索では、ユーザーが入力したクエリとドキュメントとの類似度を測定してランキング付けする。

大量のテキストから重複や偽造を検出したり、要約や分類を行ったりする。

1. 文章要約

文章要約には、主に抽出的要約と生成的要約の2つのアプローチがあります。

抽出的要約（Extractive Summarization）: このアプローチでは、元のテキストから最も重要だと思われる単語やフレーズ、文を抽出し、それらを組み合わせて要約を作成します。抽出的要約は、元のテキストの一部をそのまま用いるため、文法や意味の整合性が保たれることが多いです。ただし、抽出的要約では、新しい表現や言い回しを生成することは難しいです。

生成的要約（Abstractive Summarization）: 生成的要約では、元のテキストの内容を理解し、新しい表現や言い回しで要約を生成します。このアプローチでは、要約が元のテキストにない表現を含むことがありますが、より自然で簡潔な要約が作成できることがあります。生成的要約は、ディープラーニングやニューラルネットワークを用いたモデルで実現されることが多く、自然言語生成（NLG）の一形態とも見なされます。

文章要約技術は、ニュース記事の要約、レポートの概要作成、教材の短縮版作成など、さまざまな分野で役立ちます。自然言語処理技術の進化により、より効果的で高品質な要約が生成されるようになっています。

1. 質問応答

質疑応答（Question Answering）タスクは、特定の質問に対する正確な回答をテキストやデータベースから見つけ出す技術です。質疑応答システムは、ウェブ検索、情報検索、知識ベース、機械学習などの技術を組み合わせて機能します。質疑応答タスクは、主に以下の3つのカテゴリに分類されます。

知識ベースの質疑応答（Knowledge-based QA）:

このタスクでは、システムは既存の知識ベースやデータベースを利用して質問に回答します。例えば、質問「東京タワーの高さは？」に対して、システムは知識ベースから東京タワーの高さ（333メートル）を検索し、回答します。

文書検索ベースの質疑応答（Document-based QA）:

このタスクでは、システムは与えられた質問に関連する文書やテキストを検索し、そこから回答を抽出します。例えば、質問「ハリー・ポッターの作者は誰ですか？」に対して、システムは関連する文書やウェブページを検索し、J.K.ローリングが作者であることを回答します。

機械学習ベースの質疑応答（Machine Learning-based QA）:

このタスクでは、システムは機械学習アルゴリズムやディープラーニングモデルを用いて質問に回答します。この場合、システムは大量の質問と回答のペアから学習し、新しい質問に対しても適切な回答を生成する能力を獲得します。例えば、質問「エベレスト山の標高は？」に対して、訓練データをもとに学習したモデルが8,848メートルと回答します。

質疑応答タスクは、検索エンジン、仮想アシスタント、カスタマーサポートボット、教育アプリケーションなど、さまざまな分野で活用されています。自然言語処理技術の進化により、質疑応

1. 含意判定

含意判定タスクとは、二つの文がどれだけ論理的に関連しているかを判定するタスクです。例えば、「太郎は人間だ」と「太郎は動物だ」は、前者が正しいならば後者も正しいという関係になります。これを含意関係（entailment）と呼びます。一方、「太郎は人間だ」と「太郎は男性だ」は、前者が正しいからといって後者が正しいとは限りません。これを矛盾関係（contradiction）と呼びます。また、「太郎は人間だ」と「太郎は学生だ」は、前者と後者の間に論理的な関係がないことになります。これを中立関係（neutral）と呼びます。

含意判定タスクでは、このように二つの文の間の関係を含意・矛盾・中立の三種類に分類することが多く行われています。しかし、実際にはもっと細かく分類できる場合もあります。例えば、前提：「太郎は人間だ」と「太郎は動物だ」は、前提が真である場合仮説も真であるので、含意関係と判定します。前提：「太郎は人間だ」と仮説：「太郎は犬だ」は前提が真であった場合、仮説は偽であるので、矛盾関係となります。また前提：「太郎は人間だ」と仮説：「太郎は東京に住んでいる」は前提が真でも偽でも仮説の真偽はわからないので中立関係となります。

含意判定タスクは、自然言語処理の中でも難易度の高いタスクです。文同士の関係を判断するためには、単語や文法だけでなく、常識や背景知識や推論能力なども必要です。そのため、最近では深層学習や事前学習モデル（BERTなど）を用いた手法が注目されています。

含意判定タスクは、質問応答や情報検索やテキスト要約などの応用に役立ちます。例えば、

質問応答では、ユーザーが入力した質問文と候補となる回答文との含意関係を測定して最適な回答を選択する。

情報検索では、ユーザーが入力した検索クエリとドキュメント内の文との含意関係を測定してランキング付けする。

テキスト要約では、元のテキストから重要な文を抽出したり生成したりする際に含意関係を考慮する。

1. 文章分類

文書分類タスクとは、与えられた文書をあらかじめ定められたカテゴリに分類するタスクです。例えば、ニュース記事を政治・経済・スポーツなどのジャンルに分類したり、レビュー文を肯定的・否定的・中立的な感情に分類したりすることができます。

文書分類タスクにおいて、学習済みモデルとは、大規模なコーパス（テキスト集合）で事前に学習された自然言語処理モデルのことです。学習済みモデルは、文書の特徴や意味を表現する能力が高く、様々なタスクに適応できることが期待されています。学習済みモデルを用いることで、手元のラベル付きデータ（カテゴリーと文書のペア）で微調整（ファインチューニング）するだけで高い精度の文書分類が可能になります。

1. 文章生成

文書生成タスクとは、与えられた入力に基づいて文書を生成するタスクです。例えば、要約や翻訳や対話などが文書生成タスクに含まれます。

文書生成タスクにおいて、学習済みモデルとは、大規模なコーパス（テキスト集合）で事前に学習された自然言語処理モデルのことです。学習済みモデルは、文書の特徴や意味を表現する能力が高く、様々なタスクに適応できることが期待されています。学習済みモデルを用いることで、手元のラベル付きデータ（入力と出力のペア）で微調整（ファインチューニング）するだけで高品質な文書生成が可能になります。