医療情報アクセスシステム

背景・課題：最新の医療情報を知ることは人々の切実な願望です．しかし，その多くは英語で記述されており，難解な長文も数多くあります．こうした文を日本語で読めるようにすることは大きな技術課題です．また，内容理解のためには，単なる翻訳だけではなく，文書の意味構造を明らかにすることも必要です．

アプローチ：英日統計翻訳の高精度化のため，文を節に分割し，それらを個別に翻訳した後，統合する手法，英文を日本語の語順に並び替える手法を導入しました．内容理解に関しては，文書の意味構造解析，つまり修辞構造解析を利用し，半教師あり学習と効率的な特徴抽出手法を導入しました．

到達点：高精度な英日翻訳と修辞構造解析器を実現し，医療情報アクセスシステムを作成しました．その第一ステップとして医学生物学分野の英語論文アブストラクトのデータベースである PubMed に対し，日本語で検索し，その結果も日本語で提示，さらに情報の取捨選択も容易にできるシステムを作成しました．

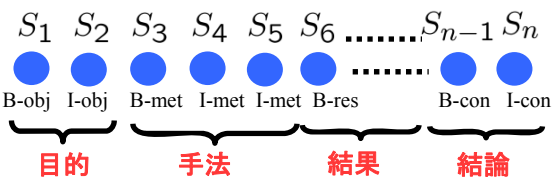
概要：

※米国国立医学図書館（NLM）が提供するPubMedを利用しています．
京都大学 金子周司氏らが開発したライフサイエンス辞書を利用しています．

修辞構造解析

論文アブストラクト(文書)をセクション(意味段落)へと分割:文系列のセグメンテーション問題

⇒半教師あり系列ラベリングを利用



各セクションに偏って出現する単語列を特徴として学習

＋

英日統計翻訳

長文は節に分割して翻訳:Divide and Translate

John lost the book that borrowed last week from Mary

→ John lost the book _s
that borrowed last week from Mary.

英文を日本語の語順へ:Head Final English

John saw a beautiful girl yesterday.

→ John yesterday a beautiful girl saw.

ジョンは 昨日 美しい 少女を 見た → 逐語訳

日本語クエリ

膵癌の化学療法？

言語横断検索

PubMed:
英語論文アブストラクト
データベース

英語文書

S_1 There is no established second line chemotherapy with.....
S_2 .....
S_3 .....
S_n The combination of capecitabine and oxaliplatin.....
.....

情報抽出型翻訳

修辞構造付き日本語訳

目的:ゲムシタビンによる化学療法行った患者に対する2次治療法は未だ確立されてない.
.....
手法:.....
.....
結果:.....
結論:カペシタビンとオキサリプラチンの併用療法が有効.

関連文献

磯崎秀樹，“英日翻訳における語順について”，言語処理学会第16回年次大会，pp.884-887，2010.

平尾努，鈴木潤，磯崎秀樹，永田昌明，“半教師あり系列ラベリングによるアブストラクトのセクション分割”，言語処理学会第16回年次大会，pp.98-100，2010

J. Suzuki, H. Isozaki: “Semi-Supervised Sequential Labeling and Segmentation using Giga-word Scale Unlabeled Data”, Proc. ACL-08:HLT, 2008

連絡先

平尾　努(Tsutomu Hirao)

協創情報研究部 言語知能研究グループ