人文科学分野における多言語コーパス検索のためのアライメント支援『星の王子さま』と『源氏物語』におけるケーススタディ

大前寛子[†] フレデリック アンドレス[††]
今井倫太[†††] 安西祐一郎[†††]

本論文では、多言語人文科学コーパス検索のためのアライメント支援ツールを提案する。多言語ドキュメントのアライメントを行うことで、多言語コンテンツを扱うことが可能となる。本システムでは、Linguistic DS でタグ付けされたドキュメントデータを用い、このタグを分析することでツリー構造を生成し、キーワードのマッチングを行うことで単語同士の対応付けを行う。また、得られた対応付けから、翻訳時に用いるためのグロッサリーを自動生成する。本論文では、システムを実装し、紫式部の『源氏物語』の現代語訳、及びサン・テクジュペリの『星の王子さま』を実行例として用い、日仏・日英ドキュメントのアライメントを行うことで、システムを評価した。

Alignment Support for Multi-lingual humanity Corpus Retrieval Case Study: The Small Prince and the Tale of Genji

Hiroko Omae,[†] Frederic Andres,[††] Michita Imai[†††]
and Yuichiro Anzai[†††]

In this paper, an alignment support tool for multi-lingual humanity corpus retrieval is proposed. Alignment of multi-lingual documents help to access to multi-lingual contents.We use the documents data tagged. The System creates a tree-structure by analyzing these tags and defines the alignment by matching the keyword. It visualizes the tree-structure and alignment of each segment. It also makes the glossary automatically, which we can use when translating. We implemented the system and made an experiment using "The Small Prince" of Antoine de Saint-Exupery and the "Tale of Genji" (contemporary style version) of Murasaki-Shikibu as an example. We evaluated the system by making the alignment of Japanese-French/Japanese-English documents.

1. はじめに

近年、文化遺産などをデジタル化して保存しようという活動が盛んになり、そのようなデジタルデータの配信の必要性も高まっている。特に古典文学は、自国文化の、他国への紹介のために用いられることが多く、他国へ配信する際のデジタル化及び配信の対象ともなっており、デジタル文書の翻訳が重要な課題となっている。

翻訳分野においては、過去20年以上に渡って、機械翻訳に関する研究が行われてきたが、近年、機械のみで行う翻訳には精度に限界があるとの見方が広まっており、人間と機械が協調して翻訳を行う、翻訳支援システムに注目が集まっている。特にその機能の1つである翻訳メモリは、例文ベースの翻訳支援システムを実現するための機能で、特定のドメインの翻訳に有効であると言われている。

古典文学の翻訳においても、翻訳メモリを用いることは非常に有効であると考えられるが、古い言い回しや、現在では使用されていない用語などは通常辞書には掲載されておらず、語の対応付けを行うことも困難である。本論文では、タグ付けされた文書をもとに、文及び語同士の対応付け(アライメント)を半自動的に行い、グロッサリーを自動的に生成することを支援する、アライメント支援ツールを提案する。このツールを用いることで未知語同士の対応付けを容易に行うことが可能となり、ここで生成したグロッサリーを翻訳メモリによる翻訳に適用することができるようになる。また、他国の言語を、古典文学を通して学ぶ際にも有効である。

本ツールでは、ユーザがシステムの提案した対応付けに間違いがある場合、それを修正することを可能とすることで、間違った対応付けから、間違ったグロッサリーを生成し、誤訳をすることを防ぐ。また、今回は日仏及び日英における文書を扱うこととし、他国でそれぞれ親しまれている、サン・テクジュペリの『星の王子さま』と、紫式部の『源氏物語』の現代語訳を用いた。

本論文の構成は以下の通りである。まず、2章で関連研究について述べ、3章で本ツールについ

† 慶應義塾大学大学院 理工学研究科
Graduate School of Science and Technology, Keio University
†† 国立情報学研究所
National Institute of Informatics
††† 慶應義塾大学大学院 理工学部
Faculty of Science and Technology, Keio University

て詳述する。4章で実験・評価、及び結果・考察について述べ、5章で今後の展望、結論について述べる。

2. 関連研究

2.1 翻訳メモリ

翻訳メモリ[1]とは、過去に翻訳した原文-訳文の組み合わせを翻訳パターンとしてデータベースに格納しておく機能である。例文ベースの翻訳システムに新たな文が翻訳対象として入力されると、その文と同一又は類似したパターンをデータベース中から検索し、必要ならば変更を加えて出力する。

翻訳パターンを用いることで翻訳にかかる労力を軽減することができ、ソフトウェアのアップグレードに伴うマニュアルの改訂など、変更の少ない文書の翻訳に有効であるとされている。また、入力文と同一の文が翻訳パターンにない場合には、その文を翻訳パターンとして新たに格納することで、精度の向上を図ることができるため、同一ドメインの文書を複数翻訳する場合には大変便利である。

図1に翻訳メモリを用いたシステムの概要を示す。

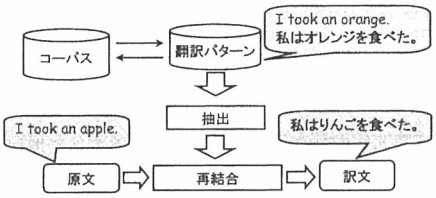

図1 翻訳メモリを用いたシステム

2.2 翻訳支援ツール

本項では、本研究に関係があると思われる既存の翻訳支援ツールを2つ挙げ、その問題点について触れる。

- Waligner
  Waligner[3]は、タイ語と英語のドキュメントのアライメントを行うツールである。まず、共通の既知語と、ドキュメント内の距離をもとに対応する文を抽出した後、その文に含まれる未知語の対応付けを行う。対応する語として複数の候補を挙げ、ユーザに選択させることが可能だが、候補がたくさん挙がる場合や、たくさん挙がっても正解が含まれていない場合があり、必ずしも精度が高いとは言えない。しかし、タグ付けされていないドキュメントも扱うことができる、という点では非常に有効である。
- JGloss
  JGloss[4]は、日本語ドキュメントのアノテーション支援ツールである。システムに日本語ドキュメントを入力すると、形態素解析を行い、各単語の読みと英訳を表示する。文節の区切り方・読み・訳が間違っている場合には、ユーザ自身がそれを修正することができ、ユーザ辞書を作成することも可能である。ただし、このツールはアライメント支援ツールではないため、2言語のドキュメント間での対応付けを行うことはできない。また、各単語ごとの英訳は表示されるが、文としての訳を表示することはできない。さらに、このツールはEDICTという日英辞書や茶筌などの形態素解析器といった、特定のツールを用いる必要があるため、精度がそのツールに依存してしまうという問題がある。

2.3 Linguistic DS

Linguistic DS[2]とは、多言語間に共通の統語・意味等に関する注釈をつけることを目的とした、GDA(Global Document Annotation・大域文書修飾)のXMLタグ集合であり、Mpeg-7によって標準化されている。Linguistic DSを用いると、文書の意味構造化を行い、多言語で統一的記述を行うことが可能となる。

表1に、Linguistic DSを用いたタグ付けの例を示す。

表1 Linguistic DSを用いたタグ付けの例

初めに、神が天と地を創造した。

```
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xml:lang="jp">
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="LinguisticType">
      <Linguistic>
        <Sentence id="b.GEN.1.1" type="verse">初めに、
          <Phrase>
            <Phrase id="GOD">神</Phrase>
            が
          </Phrase>
          <Phrase>
            <Phrase>
              <Phrase id="HEAVEN">天</Phrase>
              と
              <Phrase id="EARTH">地</Phrase>
            </Phrase>
            を
          </Phrase>
          <Phrase>創造</Phrase>
          した。
        </Sentence>
      </Linguistic>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

Linguistic DSを用いると、係り受けや動詞の依存関係、代名詞や先行詞の関係も容易に示すことができ、アライメントへ適用した場合の精度向上が予想されるが、今回は、表1程度の単純なタグ付け文書を対象とする。

3. アライメント支援ツール

本章では、提案するツールの設計と実装について述べる。

3.1 システムの概要

本システムの概要を図2に示す。

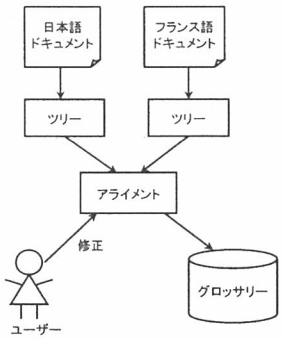

図2 システム概要

システムに入力として与えるドキュメントは、Linguistic DSでタグ付けされているものとする。これらのドキュメントを、タグをもとにツリー構造で表示する。また、タグの属性をもとに、語同士の対応関係を自動的に調べ、表示する。ここで表示された対応関係が間違っている場合には、ユーザが訂正することも可能である。このようにして得られた対応関係を、グロッサリーに格納する。

3.2 文のアライメント及びツリー生成

Linguistic DSでタグ付けされたドキュメントから、タグに従ってツリー構造を生成する。タグ付けされた文書ファイルをシステムに入力すると、まず、タグをとった、プレインなテキストを表示する。ユーザがこのテキスト中の一文を選択すると、その文のツリー構造と、他方のドキュメントにおける対応する文のツリー構造を共に表示する。対応付けはどちらの言語からでも行えるようにする。

3.3 単語のアライメント生成

得られたツリー構造から、単語のアライメントを生成する。単語の場合も、対応付けはどちらの言語からも行えるようにする。Linguistic DSでは、重要な単語にキーワードを属性として付与する。このキーワードをもとに対応付けを得、その関係を視覚化して表示する。キーワードが付与されていない単語に対しては、他の属性を持つタグをもとに対応付けを得る。システムの表示した結果が間違っている場合、ユーザがその結果を訂正することも可能である。

3.4 グロッサリー生成

ユーザが結果を訂正した後、辞書に格納したい単語については、登録することができる。後に双方の言語からの検索ができるように、二言語分生成する。インデクシングは、今回は単純にアスキーコード順とした。

4. 実 行 例

本論文では、紫式部の『源氏物語』現代語訳及びサン・テグジュペリの『星の王子さま』を用いて実行例を示す。尚、本ツールはJ2SDK1.4.1で実装した。スナップショットを図3に示す。

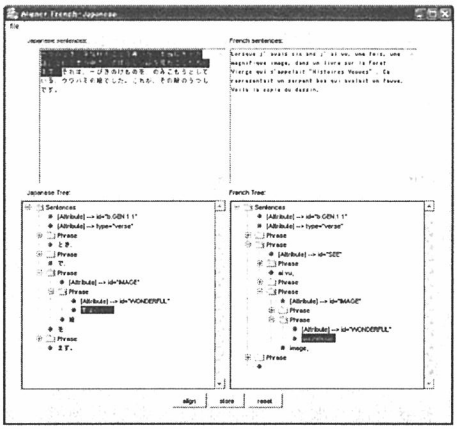

図3 実行例

図3は、『星の王子さま』第1章の第1文を実行した例である。以下に仏日双方の文を示す。

> Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Foret Vierge qui s'appelait "Histoires Vecues".
>
> 六つのとき、原始林のことを書いた「本当にあった話」という本の中で、すばらしい絵を見たことがあります。

この文では、"magnifique" =「すばらしい」、"image" =「絵」、"livre" =「本」、"Foret Vierge" =「原始林」という単語を取得・格納することができた。

ただし、フランス語の特徴として、動詞の活用、名詞・形容詞の性数の一致に伴う活用が行われる、ということが挙げられる。現時点では原形を取り出して格納することはできないが、タグの属性としてそれぞれの品詞の原形を与えておき、解析時に取得することは比較的容易に実現できると考えられる。

次に、『源氏物語』の「桐壺」第1章第1文を示す。

この文では、「帝」= "Emperor"、「御代」= "Court"、「身分」= "rank"という単語を取得・格納することができた。

ここでは、「女御」という単語が"gentlewomen of the Wardrobe"に、「更衣」が"gentlewomen of

どの帝の御代のことであったか、女御や更衣たちが大勢お仕えなさっていたなかに、たいして高貴な身分ではないで、きわだって御寵愛をあつめていらっしゃる方があった。

At the Court of an Emperor there was among the many gentlewomen of the Wardrobe and Chamber one, who though she was not of very high rank was favoured far beyond all the rest.

Chamber"と訳されている。このように1つの単語に対して複数の単語が対応する場合、システムが自動的に判断することは現時点ではできないため、ユーザが改めて指定しなければならない。これに関しても、今後扱うことができるように改良していく必要がある。

5. おわりに

本章では、本ツールにおける今後の課題、本ツールを用いた翻訳メモリの展望について述べ、最後に結論を述べる。

5.1 今後の課題

冒頭にも述べたが、Linguistic DS は意味構造を考慮したものであり、その構造を詳細に表現することができる。今回はシステムの処理が複雑になることを避けるため、利用するタグの種類も少なくし、簡単な表現のみ扱えるようにしたが、詳細にタグ付けしたものを用いることで、システムによる対応付けの精度が上がり、ユーザによる修正の労力が削減されると考えられる。そのため、より複雑なタグ解析を可能にすることも課題である。

また、現時点では、Linguistic DS のタグ付けを自動的に行うことは難しいが、この点が最も労力と時間のかかる作業である。そのため、Linguistic DS のアノテーションエディタを作成し、タグ付けへのコスト削減を図る予定である。また、このエディタを用いて長文のテキストをタグ付けし、本ツールの定量的な評価を行うことも今後の課題である。

5.2 翻訳メモリへの組み込み

本研究の最終的なゴールは、より精度とパフォーマンスの高い翻訳メモリを実現することである。翻訳メモリでは、基本的に文単位での翻訳を行うことを予定している。しかし、例えば図1のように、"I took an apple."という文を入力し、メモリ内に"I took an orange."という文があった場合、"apple"という単語が日本語の「りんご」を指すという情報は得られない。そこで、今回のツールを用いて作成したグロッサリーを適用することが考えられる。また、「オレンジ」と「りんご」はともに果物であり、食べるものである、という意味情報を加えることで、"took"を「取る」ではなく、「食べる」と訳すことができるような、意味を考慮した翻訳システムを今後開発していく予定である。

意味を考慮した翻訳システムが実現できれば、古典文学の翻訳においても、ある文と意味的に類似した文を検索・翻訳することが可能になると考える。

5.3 結　　論

本論文では、多言語人文科学コーパス検索のためのアライメント支援ツールを提案した。本ツールは、翻訳時に必要なグロッサリーを半自動的に生成し、ユーザが修正を施すことを可能とする。また、システムを実装し、紫式部の『源氏物語』の現代語訳、及びサン・テクジュペリの『星の王子さま』を実行例として用いた。今後翻訳メモリへ組み込むために、様々な点において拡張をおこなって

参考文献

1) Emmanuel Planas, Osamu Furuse: "Formalizing Translation Memories," Proceedings of Machine Translation Summit VII, Singapore, 1999.
2) Hasida K., Andres F.,Boitet C., Calzolari N., Declerck T., Farshad Fotouhi, William Grosky, Shun Ishizaki, Asanee Kawtrakul, Mathieu Lafourcade, Katashi Nagao, Hammam Riza, Virach Sornlertlamvanich, Remi Zajac, Zampolli, A.: "Linguistic DS," IO/IEC JTC1/SC29/WG11, MPEG2001/M7818
3) Nithiwat Kampanya, Prachya Boonkwan, Asanee Kawtrakul: "Bilingual Unknown Word Alignment Tool for English-Thai," Joint International Conference of SNLP-Oriental COCOSDA 2002, 9-11 May 2002, Hua Hin, Prachuapkirikhan, Thailand
4) Michael Koch: "JGloss User's Guide," 2002
5) Menezes Arul, Stephen D. Richardson: "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora," Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics, Toulouse,
6) Van Zaanen M.: "ABL: Alignment-Based Learning," Proceeding of COLING'2000, International Conference on Computational Linguistics, Saarbruecken, Germany, 2000.
7) サン＝テグジュペリ作, 内藤 濯訳: "星の王子さま," 2000.
8) Murasaki Shikibu, translated by Arthur Waley: "The Tale Of Genji," Tuttle Publishing, 2002.