用例に基づく大意翻訳のための類似文検索

下畑 光夫, 隅田 英一郎
ATR 音声言語コミュニケーション研究所
{mitsuo.shimohata,eiichiro.sumita}@atr.co.jp

松本 裕治
奈良先端科学技術大学院大学
matsu@is.aist-nara.ac.jp

1 はじめに

用例翻訳は、原言語と目的言語の文の対から構成される用例コーパスから入力文と類似した文を検索し、その対訳文を修正することで翻訳文を生成する翻訳方式である [1]。用例翻訳方式を音声対話翻訳に適用した場合 [2] に、2つの問題が生じる。一つは、長い入力文の翻訳精度が低いことである。入力文が長くなると、類似文が用例コーパスに存在する確率が低くなるために翻訳が出力されなくなる場合が多く生じる。もう一つは、入力される発話文と用例コーパスの間でスタイル (文体) が異なることによる翻訳精度の低下である。発話文を収録したコーパスは構築にコストがかかるため、スタイルは異なるが収集が容易なコーパスを用例コーパスに導入せざるを得ない。その結果、入力される発話文と用例コーパスの間のスタイルの差異のために類似文の検索が困難となる。

本論文では、上述の2つの問題にロバストな用例翻訳のための類似文検索方法を提案する。まず、翻訳の目的を主要な部分を捕らえればよいと緩和する。このアプローチを "大意翻訳" と呼ぶ。これにより、長い入力文が与えられた場合に主要な部分を共有する短い文が類似文と判定できる。また、これらの類似文の検索は内容語に基づいて行う。その出現順も基本的には利用しない。これにより、機能語や語順の違いなどによるスタイルの違いを吸収した類似文の検索ができる。

以下、2節では大意翻訳の概念について述べる。3節では大意を共有する類似文を用例コーパスから検索する方法について述べ、4節では類似文検索性能を入力文のスタイルの差異、入力文の長さで比較した実験を報告する。

((隣の部屋が) うるさいので) 部屋を代えて (下さい)
(ちょっと高いですが) シングルで (予約を) お願いします
(あなたの) パスポートを見せてください

図 1: 大意を共有する類似文の例

2 大意翻訳

一般的に、機械翻訳では与えられた入力文全体を正確に翻訳することが目的である。しかし、長い入力文の全体を正確に翻訳しようとするために、訳質が悪い翻訳文を出力する場合が多く見られる。ここで、我々が対象としている対話翻訳という用途では、入力文全体でなく主要な部分のみの翻訳であっても充分有用であると考えられる。それは、重要でない要素は欠落しても対話の進行において大きな影響は与えないか、もしくは必要であれば後の対話で問い返すことも可能だからである。つまり、重要でない部分も正確に翻訳しようとして翻訳性能が低下することより、重要な部分だけでも正確な訳を出力する方が有用であると考えられる。

この考えをふまえ、我々は対話翻訳用の目的を "入力文全体の正確な翻訳" ではなく "入力文の大意をとらえた翻訳" と緩和した。大意をとらえた翻訳では入力文の重要部分を翻訳できればよく、重要でない語、句、節などの欠落は許容する。この翻訳方針を "大意翻訳" と呼ぶ。

大意翻訳という翻訳方針を採用することで、用例翻訳では入力文との類似文を検索する条件を大きく緩和することができる。具体的には、主要部分を捕らえていれば短い文でも類似文と認定する。これにより、長い入力文であっても類似文を検索できる確率が高くなる。図1に入力文とその中の重要でない要素の例を示す。カッコで囲まれた部分は重要でない情報であり、大意翻訳ではその部分の欠落は許容する。

3 大意翻訳のための類似文検索

用例翻訳における大意翻訳とは、用例コーパスから入力文と大意を共有する類似文(原言語)を検索することに相当する。本論文では、類似文の検索は内容語に基づいて行い、機能語の情報は基本的に捨象している。これにより、文末表現や助詞の有無など機能語に起因するスタイルの違いに対しロバストな検索を行うことができる。また、内容語の出現順も基本的に使用していないため[1]、語順の異なりに対してもロバストな検索ができる。3.1節で、内容語に基づく検索の詳細について述べる。

一方、機能語は内容語の格、修飾関係やモダリティ、テンスなどの情報などの重要な情報を表している。3.2節では格、修飾関係に対する考察、3.3節ではモダリティ、テンス情報の導入について述べる。

3.1 内容語に基づく検索

類似文の検索は、内容語に基づいて行う。入力文と用例コーパス各文に対して形態素解析を施し、各文に対する内容語[2]リストを作成する。用例コーパス中の各文が、入力文と類似と判定されるには下の2つの条件を満たす必要がある。

1. 類似文は、すべての内容語が入力文の内容語に一致または類義の関係で含まれていなければならない。

2. 類似文は、少なくとも一つの内容語が入力文中の内容語と一致しなければならない。

図2に内容語リストによる類似文判定の例を示す。用例文1,2は入力文のサブセットとなっているため類似文となる。用例文3は、“ツイン”という入力文にはない語があるが、これは“シングル”と類義関係にあるため、類似文となる。用例文4は、同一語、類義語のいずれにも該当しない語“公演”があるため、類似文とはならない。

上記の必用条件を満たした用例文が類似文と認定されるが、複数の類似文が存在する場合はさらに以下の条件(優先順に記述)によりソートされる。

[1]複文の主文部を捕らえるために、後半部に出現した内容語に重みをつける処理が加わっている。

[2]名詞、動詞、形容詞、形容動詞、副詞が該当する。

入力文	(明日, あさって, シングル, 予約)	
用例文1	(明日, あさって, 予約)	○
用例文2	(シングル, 予約)	○
用例文3	(明日, ツイン, 予約)	○
用例文4	(明日, 公演, 予約)	×

図 2: 内容語リストによる類似文判定

1 同一の内容語の数
2 類義の内容語の数
3 共通する機能語の数
4 異なる機能語の数(少ない文ほど優先度高)

3.2 意味世界の狭いドメインにおける格関係、修飾関係

音声翻訳は、旅行対話などの狭い意味世界のドメインに適用されることが多い。ここで、意味世界が狭いとは構成する内容語が与えられた時に内容語が形成する格関係、修飾関係の種類が少ないことを指す。

例えば、ある文を構成する内容語リスト(泥棒, 私, 財布, 盗む)が与えられたとする。この内容語リストから以下のように様々な格、修飾関係を持つ文を考えることができる。

(1) 泥棒が私の財布を盗んだ
(2) 私は泥棒の財布を盗んだ
(3) 泥棒は私と財布を盗んだ

しかし、一般的な旅行対話を対象として考えると、ほとんどの場合は内容語リストは(1)の文を表していると考えてよい。(2)や(3)といった意味を表す場合はほとんど起らないと考えてよい。

格、修飾関係は厳密には格助詞などを解析しないと決定できないが、意味世界の狭いドメインではそれらの情報がなくても高い確率で限定されていることが期待できる。したがって、意味世界の狭いドメインで、ある2文が共通の内容語リストを共有している場合、それらの格関係、修飾関係も高い確率で一致していると考えられる。

3.3 モダリティ・テンスによる区別

文の大意は、モダリティ、テンスにより異なるため、区別しなければならない。内容語“予約”をもつ文が、モダリティ、テンスにより大意が異なる例を図3に示す。

モダリティ	テンス	文
依頼	その他	予約をしてください
質問	過去	予約をしてたでしょうか
その他	その他	予約をしています
その他	過去	予約をしていました

図 3: モダリティとテンスによる意味の異なり

モダリティ	表現
依頼	たい (助動詞), ほしい (形容詞) 願う (動詞), てください (助動詞) ていただける (助動詞)
疑問	か (終助詞), ね (終助詞)
否定	ない (助動詞), ません (助動詞)

表 1: モダリティ識別のための表現

そこで、用例文が入力文と類似であると判定されるには、入力文と同じモダリティ、テンスを有するという条件を付加している。モダリティとして (依頼、質問、その他)、(否定、その他)、テンスとして (過去、その他) を用いている。入力文のモダリティ、テンスは、表層的に識別している。識別に利用している語の例を表 1に示す。

4 実験

本手法と類似文の検索性能を比較する用例翻訳として [2] を用いた。この用例翻訳方式では、入力文と用例文の間の類似度を、構成単語列に DP マッチングを適用することで得られる編集距離で算出する。したがって、機能語、語順が類似度に反映されている。用例翻訳では [2] のように機能語も類似文判定の要素として導入することが一般的である。本手法を “提案方式”、比較に用いた用例翻訳を “従来方式” と呼ぶ。

4.1 データ

旅行会話において頻出する表現を収録したコーパス [3] を、用例翻訳用の用例コーパスとして用いた。コーパスから内容語を 1 つ以上含む文を取り出し、用例コーパスとして 101,786 文、テストデータとして 1,578 文を抽出した。このコーパスのスタイルを “基本表現” と呼ぶ。このコーパスに収録されている文は、話し言葉によく見られる文末表現が多く含まれているが、話し言葉特有の助詞の欠落、言い直し、語順の入れ替えなどはほとんど見られず、文法的には話し言葉とは性質が異なっている。

スタイル	方式	正類似文出力率
基本表現	提案方式	79.2%
	従来方式	77.5%
対話	提案方式	58.4%
	従来翻訳	50.3%

表 2: スタイルに対する正類似文出力率

また、ホテルのフロントでの模擬対話を収録し、テキストに書き起こしたコーパスもテストデータとして用いた [4]。このデータからは、内容語を 1 つ以上含む文 769 文を用いた。このコーパスのスタイルを “対話” と呼ぶ。

これら 2 つのテストデータはスタイルが異なっている。例えば、“基本表現” は平均文長が 6.4 語であるが、“対話” は 9.3 語である。また、それぞれのパープレキシティは 16 程度であるが、クロスパープレキシティは 60 程度になる。

また、内容語の類義関係を与えるためのシソーラスとして角川類語辞典を用いた。角川類語辞典では、意味素性は 3 階層からなっている。実験では、最下位層の素性まで一致した場合にその 2 語が類義関係にあるとした。このシソーラスは提案方式、従来方式の両方式で用いている。

4.2 評価

提案方式では、評価は検索された類似文と入力文を比較することで行った。つまり、比較する両文はどちらも原言語である。入力文と検索文を比較し、入力文において重要でない情報が検索文で欠落している場合は正しいとした。しかし、重要な情報が欠落したり、モダリティなどの基本的意味が異なるものは誤りとした。また、各入力文について検索された類似文の中から最も類似度の高い 1 文を抽出し、評価した。

4.3 入力スタイルの違いに対する精度

表 2に実験結果を示す。正類似文出力率とは、与えた入力文に対し正しい類似文を出力した比率を表す。基本表現スタイルでは、提案方式と従来方式の差は 1.7% とほぼ同等の性能を示している。対話スタイルでは、基本表現スタイルと比較して 2 方式とも性能が

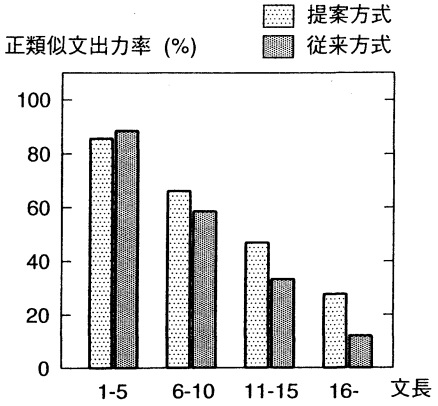

図 4: 文長で分類した正類似文出力率

低下しているが、大意翻訳の方が性能の低下が少なくなっており、両者の差は 8.1% となっている。これより、大意翻訳の方がスタイルの異なりにロバストであるといえる。

また、基本表現において提案方式と従来方式の双方が同程度の性能であることは、3.2節で述べたように、狭い意味世界のドメインでは機能語を参照しなくも格、修飾関係が高い確率で一致するという前提が成り立つことを示している。

4.4 入力文の長さに対する精度

表 2で示した対話スタイルの評価結果を、さらに入力文の長さで区分した結果を図 4に示す。入力文は、構成単語が 1 ～ 5、6 ～ 10、11 ～ 15、16 以上、という 4 つに分類した。入力文が長くなると、提案方式と従来方式は共に正訳出力率が減少しているが、提案方式の方が低下率が緩やかである。これより、提案方式は従来方式と比較すると長い文に対してロバストであるといえる。

5 まとめと今後の課題

本論文では、発話を対象とした用例翻訳のための類似文検索方法を提案した。まず、発話の中で用いられるという事を考慮し、主要な部分を捕らえた翻訳を目的とした大意翻訳を提案した。この目的の下では、主要部分を共有していれば短い用例文でも類似文と認定することができるため、長い入力文に対してロバストな類似文検索ができる。また、類似文の検索は内容語に基づいて行い、機能語の情報は捨象している。これにより、文末表現や語順など機能語に関連するスタイルの異なりにロバストな検索が可能となった。ただし、モダリティ、テンスの情報は文の大意に関わるため、表層的に識別を行い、類似文の検索において区別している。

現段階では本手法は文を構成する内容語を平等に扱っており、検索された類似文の中に主要な情報が欠落した文が多く存在している。今後は、内容語の重要度を算出し、それに基づいて類似文の検索、ソートを行うことで、正類似文出力率の向上を図っていきたい。

謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである.

参考文献

[1] M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, pp. 173–180, 1981.

[2] E. Sumita. Example-based machine translation using DP-matching between work sequences. In *Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 1–8, 2001.

[3] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd international conference on language resources and evaluation (LREC)*, pp. 147–152, 2002.

[4] T. Takezawa. Building a bilingual travel conversation database for speech translation research. In *Proc. of the 2nd international workshop on East-Asian resources and evaluation conference on language resources and evaluation*, pp. 17–20, 1999.