LE-2

用例翻訳のための同義表現同定

Identifying Synonymous Expressions for Example-based Machine Translation

下畑 光夫[1]
Mitsuo Shimohata
mitsuo.shimohata@atr.co.jp

隅田 英一郎[1]
Eiichiro Sumita
eiichiro.sumita@atr.co.jp

1 はじめに

用例翻訳は、パラレルコーパスに基づく翻訳方式の一つである。与えられた入力文に類似した原言語の文をコーパスから収集し、それらと対をなす目的言語の文を基に出力文を生成する [1]。

コーパス中の各原言語文が入力文と類似か否かを判定する 2 文間の類似度計算は、用例翻訳にとって重要である。非類似文を “類似” と判定すると誤訳を生じる可能性が高くなり、逆に類似文を “非類似” と判定すると翻訳可能な入力文 (カバレッジ) が小さくなる。

本論文では、パラレルコーパスから同義表現を獲得し、それらを用例翻訳における類似度計算に利用する方法について述べる。類似度計算において 2 文間に存在する同義表現を同定することで類似文の範囲を拡大し、カバレッジを拡大することができる。獲得する同義表現は、語彙的差異を対象としていることと局所的文脈情報を含んでいるという特徴がある。また、同義表現の同義性は目的言語の性質を反映しており、この性質も翻訳のための類似度計算に有効となっている。

2 同義表現の獲得とその特徴

同義表現は、パラレルコーパスを学習データとして自動獲得される [2]。まず、同一の目的言語文を持つという条件で原言語文の集合を形成する。同一の目的言語文を持つ原言語文は基本的に等しい意味を持つとすると、この原言語文集合は同義文集合となる。次に、同義文集合から同義文対を取り出し、語を単位として DP マッチングを適用する。語彙的な差異のみを持つような同義文対を抽出するため、2 語を越える差異を持つ同義文対は除外する。さらに、DP マッチングにより検出された差異ならびに差異の前後の語を併せて取り出すことで、同義表現が獲得される。

この方法により、語彙的、局所的な同義表現が獲得される。また、同義表現に差異の前後の語を含めることで、同義性が成り立つ文脈の制約が付加される。獲得した同義表現集合の例を、図 1に示す。“です”、“のです”、“でしょう” という語は単独では同義表現とはいえないが、 J1 に示すように前後の語を制約することにより適切な同義表現となっている。

J1	いい	です	か
	いい	のです	か
	いい	でしょう	か
J2	#	切符	は
	#	チケット	は
E1	#	Could	you
	#	Would	you
	#	Will	you
E2	tell	him	to
	tell	her	to

図 1: 獲得した同義表現

加えて、この同義表現は訳文が一致することを同義性の根拠としているため、対訳的観点において同義な表現をその中に含んでいる。用例翻訳に適用する場合には、この性質が有効となる。例えば同義表現集合 J2,E2 は、原言語だけの観点では同義表現とは言いにくいが、目的言語への対訳という観点では同義表現としてもかまわない。例えば、英語では “切符” と “チケット” は共に “ticket” と訳されるため、日英翻訳においては J2 のように両者を同一視してもよい。また、日本語の会話文では目的語は省略されることが多いため、 E2 のように “him” と “her” を同一視してもかまわない。

3 用例翻訳への適用

本手法を組み込んだ用例翻訳の構成を図 2に示す。用例翻訳が備えるバイリンガルコーパスから同義表現を獲得し、それらを文間類似度計算部分に組み込む。入力文が与えられると、コーパス中の原言語文との間で類似度計算を行う。その際に、 2 文間に存在する同義表現は同一視される。類似とされる原言語文と対をなす目的言語文を取り出し、それらから出力文を生成する。

[1]ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Labolatories

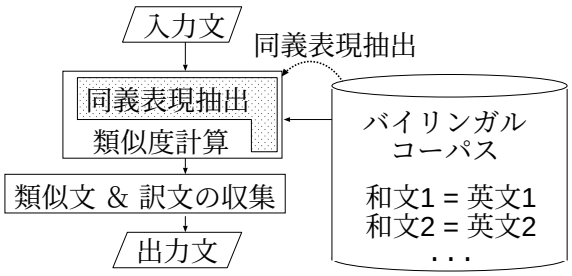

図 2: 同義表現同定処理を導入した用例翻訳

4 評価実験

本手法の効果を見るために、同義表現同定処理を行うシステム (“同定あり”) と行わないシステム (“同定なし”) の 2 つを比較した。“同定なし” では、 exact マッチにより入力文と類似か否かを決定した。“同定あり” では、同義表現を同一視した上で exact マッチを行い、類似か否かを決定した。どちらのシステムも、類似文の対訳文そのものを出力文とした。1 入力文に対して出力文が複数ある場合は、すべての出力文を評価した。

実験には、日本語と英語からなる旅行会話を対象としたバイリンガルコーパスを用いた [3]。学習データ約 10 万文と評価データ約 8 千文を取り出し、学習データから同義表現を獲得した。学習データは用例翻訳のコーパスとしても使用し、評価データの翻訳を行った。日本語から英語 (日英) とその逆方向 (英日) の 2 方向の翻訳について評価した。

評価のポイントは、翻訳可能な入力文数と翻訳精度の 2 つである。同義表現同定により翻訳可能な入力文数がどの程度拡大するか、また拡大した類似文により得られる翻訳の精度を実験で評価した。

4.1 翻訳可能な入力文

翻訳可能な文を、少なくとも 1 文の類似文が検索された文と定義した。評価データを与えて得られた翻訳可能な文を図 1に示す。日英、英日とも、カバレッジは 7% 弱拡大することができた。

4.2 翻訳精度

得られた出力文の一部を、目的言語のネイティブスピーカにより評価した。“同定あり” では、同義文同定処理により新たに獲得された訳文を評価対象とした。訳文として概ね正しければ “正訳” とした。図 2に示した実験結果より、“同定あり” と “同定なし” ではほとんど翻訳精度に違いは見られない。

表 1: 翻訳可能な入力文数

翻訳	同定なし	同定あり	拡大率
日英	3,198	3,419	6.9%
英日	2,845	3,034	6.6%

表 2: 翻訳精度

		同定なし	同定あり
日英	総文数	1,552	961
	正訳文	1,395	871
	精度	89.9%	90.6%
英日	総文数	1,645	1,055
	正訳文	1,606	1,030
	精度	97.6%	97.6%

5 まとめ

本論文では、コーパスから同義表現を獲得し、それらを文間の類似度計算に利用して類似文の検索を拡大する方法について述べた。獲得する同義表現はその同義性に目的言語の性質を反映しているため、用例翻訳に適したものとなっている。

本手法を用例翻訳に適用すると、翻訳精度を低下させることなくカバレッジを拡大できることが実験により示すことができた。さらに、この効果は日英、英日の双方向の翻訳に対して確認することができた。

用例翻訳でよりよい翻訳を行うためには充分な用例を含んだ稠密なコーパスが求められるが、そのようなコーパスを用意することは容易ではない。本手法はコーパスの語彙的な稠密度を補う技術として有効であるといえる。

謝辞

本研究は通信・放送機構の研究委託により実施したものである.

参考文献

[1] 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.

[2] 下畑光夫, 隅田英一郎. パラレルコーパスからの語彙的パラフレーズ獲得. 情報科学技術フォーラム (FIT), 2002.

[3] T. Takezawa. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd LREC*, pp. 147–152, 2002.