

翻訳と言い換え ”(ソース)入力文:(ターゲット)入力文”の学習の有効性

村上仁一
鳥取大学工学部

murakami@tottori-u.ac.jp

概要

言い換えと翻訳は、同一言語と他言語の違いがあるが、意味が同じ文章を生成する点において同一の処理とみなせる。そこで、1つのシステムにおいて、言い換えと翻訳が同時におこなうシステムを構築することを試みた。そのとき、以下の概念を組み込んだ。

- 最も精度が高い言い換え文は入力文である。つまり、入力文と出力文が同一のとき、もっとも精度の高い言い換えになる。
- 翻訳は言い換えの一種である。

この概念から、対訳学習文に、”(ソース)入力文:(ターゲット)入力文”を追加する。その結果、高い精度で、言い換えが得られた。また高い精度の翻訳が可能になった。

1 はじめに

通常、言い換えは、同一言語において、意味が同じ文章に変換することをいう。一方、翻訳は、他言語において、意味が同じ文章に変換することをいう。両者は、同一言語と他言語の違いがあるが、意味が同じ文章を生成するため、同一の処理系とみなせる。そこで、1つのシステムにおいて、言い換えと翻訳を同時におこなうことを試みた。

ところで、最も精度が高い言い換えは、入力文である。(入力文と出力文が同一である。)このデータを学習にもちいると、言い換の精度は、大幅に高くなる。

この概念を、翻訳に組み込む。通常、対訳学習文は、日英翻訳の場合、(日本語: 英文)の組み合わせである。この対訳学習文に、(入力文(日本語): 入力文(日本語))を追加する。この結果、翻訳精度が向上する。この結果についても述べる。

2 言い換え文を利用した言い換え

言い換えは、同一言語において、意味が同じ文章に変換することをいう。システムに使用する学習データに、言い換え文を使用したとき、言い換えは、翻訳と等価な技術になる。学習データおよび入力文の例を表1に示す。

表1 言い換え文を利用した言い換え学習データ (言い換え文)

source	target
私は馬鹿だ	私は阿呆だ
彼女は馬鹿だ	彼女は阿呆だ
彼は馬鹿だ	彼は馬鹿だ ¹⁾
入力文	彼は馬鹿だ
順位	出力文(言い換え文)
1	彼は馬鹿だ
2	彼は阿呆だ

この方式の言い換え精度は高い。精度85%の言い換え文38万文を利用して、”(ソース)入力文:(ターゲット)入力文”を学習すると、出力文の第2候補の言い換え精度は90%を超える。(第1候補は入力文になる。)

3 対訳文を利用した言い換えと翻訳

翻訳は言い換えの一種であると仮定する。そして、学習データは、以下とする。

1. 対訳文
2. ”(ソース)対訳文の日本語文:(ターゲット)対訳文の日本語文”
3. ”(ソース)入力文:(ターゲット)入力文”

また、出力文を複数文とする(N-best)。この出力文は、言い換えと翻訳が混在する。学習データと入力文の例を表2に示す。

表2 対訳文を利用した言い換えと翻訳学習データ

source	target
私は馬鹿だ	I am a foolish
彼は阿呆だ	He is a foolish
私は馬鹿だ	私は馬鹿だ
彼は阿呆だ	彼は阿呆だ
彼は馬鹿だ	彼は馬鹿だ ²⁾
入力文	彼は馬鹿だ
順位	出力文(言い換え文)
1	彼は馬鹿だ
2	彼は阿呆だ
3	He is a foolish
4	彼は foolish

この方式は、対訳文と入力文を利用することで、翻訳を含む言い換え文が出力できる。

- 1) (ソース)入力文:(ターゲット)入力文
- 2) (ソース)入力文:(ターゲット)入力文

3.1 実験条件

本研究では、3章で示した言い換えおよび翻訳の精度を調査する。表3に実験に利用したデータ量を示す。

表3 実験に利用したデータ

学習データ [1]	複文 約 9 万文 + 単文 約 16 万文
入力文	複文 100 文

なお、翻訳ソフトには OpenNMT V2.3.0 を利用し、vocabulary size は 200,000 とした。その他の学習パラメータはすべて default とした。

3.2 実験結果

出力例文を表4に示す。また、出力文は 4-best とする。

表4 出力例

入力文	この花が咲くともう春です。
出力文	1 この花が咲くともう春です。 2 This flower is already spring . 3 When these flowers bloom , it is spring . 4 This flower is spring .
入力文	その国はこの戦いを前から予知してそれに備えてきた。
出力文	1 その国はこの戦いを前から予知してそれに備えてきた。 2 その国はこの戦いを前から予測してそれに備えてきた。 3 その国はこの戦いを前から予知してそれに備えてきた。 4 その国はこのBattleを前から予知してそれに備えてきた。
入力文	むやみに干渉しては有害無益だ。
出力文	1 むやみに干渉しては有害無益だ。 2 Do not interfere without interference without interference . 3 No interference will do more harm than good . 4 irrespective 干渉しては有害無益だ。
入力文	軍隊を出してストライキを押さえ付けた。
出力文	1 軍隊を出してストライキを押さえ付けた。 2 軍隊を出してストを押さえ付けた。 3 They put out the army and put down the strike . 4 They put out the army and suppressed the strike .

表4から、以下のことが示される。

- 第1候補は、入力文と同一になることが多い。
- 第2候補以降は、言い換え（同一言語）と翻訳（他言語）が混在する。
- 翻訳を言い換えとみなすと、高い精度で言い換え文が得られている。

3.3 言い換え精度の調査

正順逆順組み合わせ4方向、4初期値、4-best、多数決の翻訳システム（詳細は、[2]）において、テスト文100文の第2候補における言い換え文の精度を調査した。なお、入力類似文による fine tuning はおこなっていない。この結果を表5に示す。また、出力例を表10の”Pa”に示す。言い換え（翻訳を含む）の精度を人手で調査した。その結果を表5に示す。

表5 言い換え成功率

言い換え成功 95 文 (精度 95%)

3.4 対訳文を利用した言い換えと翻訳の結果

実験結果より以下のことが示される。

- 全体の評価
表5や表10から、言い換えと翻訳が高い精度で、同一システムにおいて実行できる。
- 言い換え
提案手法は、言い換え精度が非常に高い。言い換の評価は、人によって異なるため、明確には示せないが、90%を超える精度を持っている。
- 翻訳
言い換え文が英語であったとき、翻訳とみなせる。このときの翻訳精度は高い。翻訳精度も90%を超える。

3.5 対訳文を利用した言い換えと翻訳の考察

- ”(ソース)入力文:(ターゲット)入力文”の効果
3章において提案したシステムは、言い換えおよび翻訳両者において、高い精度を持つ。この原因を以下に考えている。
入力文の言い換の最大の候補は、入力文である。そのため、学習データに、”(ソース)入力文:(ターゲット)入力文”を加える。これにより、言い換の精度は高くなる。なお、”(ソース)入力文:(ターゲット)入力文”を学習しなかったとき言い換え精度は、大幅に劣化した。つまり、このデータが、言い換えおよび翻訳において、非常に有効な情報になっている。
- 翻訳精度の高い原因
このシステムにおいて、翻訳精度の高い原因を考察してみた。翻訳は、言い換の尤度を上回ったとき、出力される。つまり、精度の低い翻訳は、言い換えになる。したがって、精度の高い翻訳のみが出力される。

4 翻訳

4.1 言い換の抑制

”(ソース)入力文:(ターゲット)入力文”

3節の実験では、言い換えと翻訳が混在する。翻訳は、他言語の文生成であるため、同一言語の文生成を抑制する必要がある。また、”(ソース)入力文:(ターゲット)入力文”の学習が、言い換えおよび翻訳の精度に有効であることが示された。

そこで、通常対訳学習データに”(ソース)入力文:(ターゲット)入力文”のみを追加して、学習をおこな

う。”(ソース)入力文:(ターゲット)入力文”は、対訳文と比較すると量が少ない。そのため、出力文は、他言語のみが出力され、同一言語の言い換えは出力されない。したがって翻訳になる。

4.2 翻訳の実験の条件

翻訳システムは比較のため学習データを変えて3種類作成した。それぞれの学習データを表6に示す。なお、入力文は複文100文である。単文および複文は、文献[1]を利用した。JPACRAWL22万文は、JPACRAWLから入力文の類似文を1文につきTFを用いて1000文を収集した対訳文である。

表6 各種翻訳システムにおける学習データ

baseline (S1)
複文 約9万文 + 単文 約16万文 + JPACRAWL 約22万文
baseline + tune (S2)
baseline のシステムに、入力文ごとに類似文を学習データからTFを用いて4文を抽出し、再学習したモデル
proposed (S3)
複文 約9万文 + 単文 約16万文 + JPACRAWL 約22万文 + 入力データの”ソース: ソース”100文
proposed + tune (S4)
proposed のシステムに、入力文ごとに類似文を学習データからTFを用いて4文を抽出し、再学習したモデル

4.3 翻訳の実験の結果

実験結果を表7に示す。また出力例を表10に示す。

表7 翻訳の実験の結果(自動評価)

	BLUE	meteor	TER	RIBES	STR ³⁾
baseline	0.269	0.552	0.533	0.824	0.07
baseline + tune	0.297	0.562	0.526	0.819	0.10
proposed	0.276	0.553	0.542	0.817	0.06
proposed + tune	0.307	0.569	0.519	0.818	0.10

4.4 翻訳の実験の考察

実験の結果から以下のことがわかる。

1. “(ソース)入力文:(ターゲット)入力文”を、単に学習データに追加するだけでは、翻訳精度に大きな差がでない。
2. しかし、“(ソース)入力文:(ターゲット)入力文”を、学習データに追加しておくことは、入力1文ごとの類似文を利用した再学習(fine tuning)において、有効である。

5 翻訳と言い換えの結合

5.1 翻訳と言い換えの結合(同一言語の出力の抑制)

3節の実験において、出力文には、日本語と英文が混在する。同一言語の出力を抑制する方法として、2つのシステムを結合し、同一言語の出力を削除し、その結果の多数決をとる方法がある。

学習データの大多数を日本語:英文とし、わずかに日本語:日本語を加えたシステムは、日英翻訳になる。学

習データを日本語:英文と、日本語:日本語を加えたシステムは、翻訳と言い換えが入り交じる。2つのシステムを結合し、同一言語の出力を削除し、その結果の多数決をとると、日英翻訳になる。具体的には、表6のproposed+tune(S4)と、表4の実験を組み合わせる。

5.2 翻訳と言い換えの結合の実験結果

実験結果を表8に示す。同時に、google との比較も行った。出力例文を表10のS5に示す。

表8 翻訳と言い換えの結合の結果

	BLUE	meteor	TER	RIBES	STR
Proposed + Baseline	0.320	0.585	0.506	0.832	0.10
google	0.260	0.554	0.577	0.820	0.05

この結果から、高い翻訳性能が得られていることがわかる。

6 考察

以上4つの実験から以下のことが考察できる。

6.1 “(ソース)入力文:(ターゲット)入力文”の有効性

”(ソース)入力文:(ターゲット)入力文”の学習は言い換えおよび翻訳ともに有効である。この理由を2つ考えられる。

1. 未知語

入力文に存在するが、学習データに存在しない単語は、未知語になる。未知語の出現頻度はジップの法則に従うため、この問題は、自然言語処理一般において、厄介な問題である。しかし、”(ソース)入力文:(ターゲット)入力文”の学習は、未知語の問題を緩和できる。

2. 対義文の抑圧

翻訳や言い換えにおいて、反対の意味を持つ文が出力されることがよくある。この原因は、反対の意味をもつ単語は、基本的に使いかたが同じになるためである(例 信号が青になった 信号が赤になった)。したがって word vector を使った場合、単語において最も近い類義語は、対義語になることがある。しかし、”(ソース)入力文:(ターゲット)入力文”の学習は、この対義文の問題を緩和できる。出力文(言い換え文や翻訳文)をみると、この対義文の生成が、大幅に低下している。

6.2 encoder-decoder model と decoder-only model

本研究では encoder-decoder モデルを利用している。しかし decoder-only モデルは encoder-decoder モデルよりも、2言語学習時における1言語出力が容易である。[4]は、fairseq を利用し、decoder-only-model で言い換えと翻訳を同一視したシステムを構築した。そして、有望な結果を得ている。

6.3 翻訳におけるソース言語の拡張性

翻訳では、対訳データが必要である。しかし対訳データは量が限定されているため、単言語データを利用したデータ拡張が研究されている。様々な方法が提案されている。しかし、ターゲット言語を拡張(増加)して、言語

3) Sentence Translation Ratio [3]

モデルを強化することで翻訳性能を向上する方法が最も良いように見える。一方、ソース言語を拡張（増加）する方法は、多くの研究において、翻訳性能がほとんど向上していない。しかし、本研究で示した、“（ソース）入力文:（ターゲット）入力文”の学習は、ソース言語を拡張（増加）する方法と言える。

6.4 言い換えと翻訳 (ターゲット言語の切り替え)

基本的には、言い換えと翻訳は、同じシステムで構築できる。そして、ターゲット言語をソース言語と同一言語にした場合、言い換えになる。またターゲット言語をソース言語と他言語にした場合、翻訳になる。言い換えもしくは翻訳に特化するために出力を選別する方法として複数ある。

1. 学習データの量による制限
本論文で提案した方法である。学習データの大多数が日英対訳データであれば、翻訳になる。学習データの大多数が日日の言い換えデータであれば、言い換えになる。
2. 出力文の単語の種類を制限
出力文において、単語の種類を制限する。入力文を日本語とし、出力文を日本語に制限すると、言い換えになる。.. 入力文を日本語とし、出力文を英語に制限すると、翻訳になる。
3. 入力に tag を付与
入力において、tag をつけて、出力を制限することで、出力文の目的を切り替える。例を表 9 に示す。

表9 TAGをつけた学習データと出力

日本文言い換え	日本文 <jp2jp> 日本文
日英翻訳	日本文 <jp2en> 英文
英文言い換え	英文 <en2en> 英文
英日翻訳	英文 <en2jp> 日本文

予備実験では、decode only model は encoder decoder model より性能が低くなった。tag を利用する方法は 1 節の方法より、性能が低くなった。節の方法は、OpenNMT の制約から実行できない。今後、これらの方法を調査していく。

6.5 人手評価

提案システム（5 章）と google との人手による対比較評価を行った。自動評価（表 8）では、提案システムは google を上回っている。しかし対比較評価では、提案システムは google を下回った。今後、評価者の問題も含め、検討していきたい。

7 おわりに

言い換えと翻訳は、同一言語と他言語の違いがあるが、意味が同じ文章を生成する点において同一の処理とみなせる。そこで、1つのシステムにおいて、言い換えと翻訳が同時におこなうシステムを構築することを試みた。その過程において、“（ソース）入力文:（ターゲット）入力文”の学習が、特に tuning において有効性が示された。そして、高い精度で、言い換えと翻訳が可能であることを示した。なお、提案した翻訳システムは、自動評価では、BLUE 値は 0.32 となり、google を超えている。

謝辞

評価者の大学院生の名村 太一と松本 武尊 両氏に深く感謝します。

参考文献

[1] 村上仁一, 藤波進. 日本語と英語の対訳文対の収集と著作権の考察. 第一回コーパス日本語学ワークショップ, pp. 119–130, 2012.

[2] 村上仁一. 多数決による自己回帰モデルに基づく機械翻訳. 言語処理学会第 30 回年次大会, No. C10-4, 2024.

[3] 村上仁一. 機械翻訳における文一致率による評価. 人工知能学会全国大会論文集 第 27 回, 2013.

[4] 名村太一, 村上仁一. 対訳文のみを用いた翻訳と言い換えのマルチタスク学習における翻訳精度. 言語処理学会第 31 回年次大会, 2025.

表 10 実験結果 出力文 まとめ

入	入力文
Pa	言い換え (3 節)
S2	baseline + tune (6 節)
S4	proposed + tune (6 節)
S5	baseline + proposed (6 節)
Go	Google 翻訳

入	～でのデモは過去の抗議デモなどちっぽけに見えてしまうほどの規模になる見込みだ。
Pa	～でのデモは過去の抗議デモなどたずさわる見えてしまうほどの規模になる見込みだ。
S2	The demonstration in is expected to be the size of the past protest demo .
S4	It is expected that will be the size of the new demo demo .
S5	A demonstration in is expected to be in the scale of the past , including protests in the past .
Go	The demonstration in is expected to be of such a scale that past protests will seem small .

入	送り状第 2929 号に従って船積みした商品の損傷のため、貴殿に迷惑をおかけし申し訳ございません。
Pa	送り状第 333 号に従って船積みした商品の損傷のため、貴殿に迷惑をおかけし申し訳ございません。
S2	We apologize for any inconvenience this may cause you
S4	We apologize for the inconvenience this may cause you for the damage of the goods shipped according to the issue of LINE .
S5	We apologize for any inconvenience this may cause you
Go	We apologize for the inconvenience caused to you due to the damage to the goods shipped according to invoice number 2929 .

入	進退伺いを出したらその儀に及ばずとのことであった。
Pa	終日伺いを出したらその儀に及ばずとのことであった。
S2	When I decided to do it , I didn't know how to do it .
S4	When I asked to hear the letter , I would like to call on it .
S5	As I asked the letter , I did not know the ceremony .
Go	I was told that if I asked them to move forward , I would not be able to do so .

入	ご存じのように、自動車業界は法律の要求を満たすことができないと主張しています。
Pa	ご存じのように、自動車産業は法律の要求を満たすことができないと主張しています。
S2	As you know , claims that the automotive industry cannot meet legal requirements .
S4	As you know , the automotive industry claim that it is not possible to meet the demands of the law .
S5	As you know , the automotive industry claims that the automotive industry cannot meet legal demands .
Go	As you know , the auto industry claims that it cannot meet the demands of the law .

入	その国はこの戦いを前から予知してそれに備えてきた。
Pa	その国はこの戦闘を前から予知してそれに備えてきた。
S2	The country predicted this battle and prepared for it .
S4	The country has predicted this battle before it has been foreseen
S5	The country has foreseen this battle from before it .
Go	The nation has foreseen this battle and prepared for it .

入	みだりに干渉しては有害無益だ。
Pa	むやみに干渉しては有害無益だ。
S2	Mere interference will do more harm than good .
S4	Your interference will do more harm than good .
S5	No interference will do more harm than good .
Go	Unnecessarily interfering is harmful and useless .

入	軍隊を出してストライキを押さえ付けた。
Pa	軍隊を出してストを押さえ付けた。
S2	We put out the army and put down the strike .
S4	He put out the army and suppressed the strike .
S5	We put out the military and put down the strike .
Go	I sent in the army to suppress the strike .

入	無実が証明されれば死んだ父も成仏できます。
Pa	無罪が証明されれば死んだ父も成仏できます。
S2	My father can also die if he is exonerated .
S4	My father 's soul is grave if he is exonerated .
S5	My father 's soul will rest in peace if he is exonerated . These flowers bloom in spring .
Go	If my innocence is proven , even my dead father can be enlightened .

入	この花が咲くともう春です。
Pa	This flower is already spring .
S2	These flowers bloom in spring .
S4	The cherry-tree blossoms in spring .
S5	These flowers bloom in spring .
Go	When this flower blooms , it's already spring .

入	天才のすることは普通の物差しでは、計れない。
Pa	天才のことは普通の yardstick では、計れない。
S2	Genius is not a normal measure to measure .
S4	The practice of genius is not to be measured by an ordinary measure .
S5	Genius can not be measured by ordinary yardstick .
Go	You can't measure what a genius does with an ordinary ruler .

入	犬がいなくなったので子供たちはすっかりしょげている。
Pa	The children are completely disheartened because the dog has gone .
S2	The children are completely disheartened because the dog has gone .
S4	The children are completely disheartened because the dog has gone .
S5	Now that the dog has left it is quite disheartened .
Go	The children are devastated because the dog is gone .

入	きみのあいまいな説明ではあの娘がだれだかわからなかった。
Pa	きみのあいまいな説明ではあの娘がだれだかわからなかった。
S2	I couldn't recognize the girl from the vague description of her you gave me .
S4	I couldn't recognize the girl from the vague description of her you gave me .
S5	I couldn't recognize the girl from the vague description of her you gave me .
Go	I couldn't figure out who that girl was from your vague description .

入	そのような危機に直面しても彼女が冷静なのに驚いた。
Pa	I was surprised for her to face such a crisis .
S2	Even with such a crisis , she was surprised in the face of such a crisis .
S4	She was surprised at her calm in the face of such a crisis .
S5	She was surprised at the face of such a crisis .
Go	I was surprised at how calm she was in the face of such a crisis