相対的意味論と機械翻訳の応用

村上仁一
鳥取大学 工学部
murakami@tottori-u.ac.jp

1 はじめに

本論文では，語彙の意味を，説明するために，相対的意味表現と，絶対的意味表現について述べる．次に，これらの方法でも，説明できない語彙に対し，２言語を利用した方法をしめす．そして，この方法を利用した翻訳方法について述べる．具体的には，”相対的意味論に基づく変換主導型統計翻訳 (TDSMT)”と，”相対的意味論に基づく統計翻訳 (RSMT)”の方法について述べる．最後に相対的意味論と NMT(Nural Network Machine Translation) の関係を考察する．

2 語彙の意味と辞書

2.1 辞書学と意味表現

辞書は，現在や過去において，使用された語彙を収集し，その品詞・意味・背景（語源等）・使用法（用例）・派生語・等を解説した書籍である．基本的には，語彙は，他の語彙を利用して，解説する．１つの辞書において，解説文には，出来るだけ統一をとるように収集される．

語彙について何らかの記述をするための言語を，メタ言語と呼ぶ．初めに，語彙とメタ言語が，同一言語の場合の例を示す．そして，語彙の意味を絶対的意味表現と相対的意味表現をもちいて述べる．

2.2 絶対的意味表現

絶対的意味表現とは，単語の語彙の意味を具体的な他の語彙で表現する．この表現方法は，基本的には，理解しやすい．以下は，絶対的意味表現の例である．

青　– 波長 400nm 近辺の光
赤　– 波長 800nm 近辺の光

多くの語彙は，絶対的意味表現で，表現できる．しかし，意味の表現が困難な語彙も多い．以下は，その例である．

右[1]
- 南を向いたとき，西にあたる方（広辞苑）
- アナログ時計の１時から５時までの表示がある側．（新明解国語辞典）
- この辞書を開いて読むとき，偶数のページのある側（岩波国語辞典）

左
- 南に向かったとき，東にあたる方（旺文社国語辞典）
- アナログ時計の文字盤に向かった時に，七時から十一時までの表示のある側．（新明解国語辞典）
- この辞典を開いて読む時，奇数ページのある側．（岩波国語辞典）

いずれの説明も，曖昧な点が残る．（例えば，“南”とは？）似た語彙として“上”“下”などがある．

2.3 相対的意味表現

相対的意味表現とは，単語の意味を，お互いに相反する単語で表現する方法である．例を以下に述べる．

右　– 左の反対
左　– 右の反対

絶対的意味表現で困難な語彙は，相対的意味表現で表現可能な場合が多い．しかし，この表現方法は，説明がループするため，“煙を蒔いたような表現”と言って嫌悪する人も多い．

また，語彙の中には，対的意味表現を用いても，説明できない語彙がある．以下は，その例である．

人間
- ひと　人類
- 遺伝的な両親が人間である人
- 考える葦

いすれの説明も，疑問点が残る．このように，“人間”の意味を説明することは，非常に困難である．[1)]

2.4 ２言語を用いた意味表現

単言語では説明するのが困難な語彙は，２言語を利用して表現することが可能な場合がある．例を以下に挙げる．

- “人間”とは“human”である．

以下に，２言語を用いた例を示す．
“右”とは“right”である．
“犬”とは“dog”である．
“動物”とは“animal”である．
しかし，“犬”には別の意味がある．
“犬”とは“spy”である．
このように，２言語をもちても，語彙に複数の意味があるとき，一意に表現できない．

2.5 ２言語を用いた相対的意味表現

単言語において，複数の意味があるとき，２言語を用いた意味表現では，１意に表現できない．このような場合．２言語の相対的意味を用いて表現できる．

“a が b ならば c は d”

これを“２言語を用いた相対的意味表現”と呼んでいる．

例を以下に示す．

1. “右”が“right”ならば“左”は“left”
2. “犬”が“dog”ならば　“猫”は“cat”

1) 個人的には，２番目の説明が好みである．

3. “犬” が “Canis” ならば “猫” は “Felis”
4. “犬” が “spy” ならば “猫” は “sex decoy”

2)

3 ２言語を用いた相対的意味表現の辞書（変換テーブル）

3.1 変換テーブルと文法

本研究では，語彙を２言語を用いた相対的意味表現に基づいて，考える．これを以下のように表現する．

- “a が b ならば c は d ただし a,b,c,d は単語”

この辞書を “変換テーブル” と呼んでいる．
例を以下に示す．

a:犬 b:dog c:猫 d:cat
a:犬 b:Canidae c:猫 d:Felis

dog と cat は，哺乳類である．Canis は犬科で，Felis は猫科である．この例では，名詞の意味的な階層の情報が変換テーブルに含まれる．

a:彼 b:he c:彼女 d:she
a:彼 b:him c:彼女 d:her

この例では，代名詞の文法が変換テーブルに含まれる．

a:走る b:run c:歩く d:walk
a:走った b:ran c:歩いた d:walked

この例では，時制の文法が変換テーブルに含まれる．

これらの例から，変換テーブルには，文法や属性や意味の情報が含まれることがわかる．なお，文脈に応じた情報も含めることができると，考えている．

3.2 変換テーブルの作成方法

変換テーブルは，手動で作成できる．しかし，自動的に作成する方法がある．その方法の１つに，文パターンを利用して作成する方法がある．作成方法を以下に示す．

1. 学習文１を準備する．
   - 彼は山にいった．
   - He went to mountain.
2. 対訳単語を準備する．
   a: 山 b: mountain
3. 文パターンを作成する．
   - 彼は *N* にいった．
   - He went to *N*
4. 学習文２を準備する．
   - 彼は海にいった．
   - He went to sea .
5. 学習文１と文パターンと学習文２から，以下の変換テーブルを作成する．
   a:山 b: mountain c:海 d: sea

4 相対的意味論を用いた変換主導型機械翻訳 (TDSMT)[2]

変換テーブルは，文中の単語の置換可能な辞書として捕らえることができる．したがって，変換テーブルの応用として，機械翻訳がある．以下に相対的意味論を用いた変換主導型機械翻訳 (Tranfer Driven with Relative Meaning using Statistica Machine Translation 以後 TDSMT) の概念を述べる．例として，日英翻訳を考える．

4.1 TDSMT の基本概念

基本概念を，以下に述べる．

- A が B ならば C は D である．
- A,B,C,D は文である．
- a,b,c,d は単語もしくは句である．
- 入力文は C で，翻訳文は D になる．

以下に例を示す．

1. 入力文
   日英翻訳において，以下の日本文の翻訳を想定する．
   Cショウジョウハエは林檎が好きだ
2. 学習文
   学習文として 以下の対訳文を想定する．
   A私は林檎が好きだ
   BI like an apple
3. 変換テーブル
   以下の変換テーブルを想定する．
   a: 私 b: I c: ショウジョウハエ d: Fruit flies
4. 変換規則
   以下の変換規則を想定する．
   A に a を適用し，B に b を適用できたならば，C に c を適用し，D に d を適用する．
5. 出力文（翻訳文）
   以上の結果から以下の出力文（英文）を得る．
   - Fruit flies like an apple

4.2 TDSMT の一般形式

TDSMT では，以下を基本概念とする．

- A が B ならば C は D
- ただし A,B,C,D は文

C は入力文，D は出力文

A対訳学習文の日本文
B対訳学習文の英文
C入力の日本語文
D出力の英文（翻訳文）

- a が b ならば c は d
- ただし a,b,c,d は単語もしくは句

a対訳学習文の日本語単語
b対訳学習文の英語単語
c入力の日本語単語
d出力の英語単語

以下の変換規則を想定する．

- A に a を適用し，B に b を適用できたならば，C に c を適用し，D に d を適用できる．

4.3 TDSMT の例

Aツバメは矢のように飛ぶ
BSwallow flies like an arrow
C時間は矢のように飛ぶ
DTime flies like an arrow

a:ツバメ b:Swallow c:時間 d:Time

“Time flies like an arrow” の訳として “時蝿，矢を好む” がある．この訳は，非文と考えられる．しかし，なぜ，非文と考えられるのだろうか？通常は，“時蝿” が存在しない

2) なお，“犬” が “spy” ならば “猫” は “Mata Hari” は面白い

から，非文とする．しかし，存在しないことの証明は，困難である．

相対的意味論の考えかたは，“ツバメは矢のように飛ぶ” があるから，“時間は矢のように飛ぶ” と訳すことができると解釈する．

4.4 TDSMT の動作例

実際の文章に TDSMT を使って翻訳した例を以下に示す．

A風 が 北 から 南 に 変わった
BThe wind changed from north to south
C信号 が 青 より 赤 に 変わった
DThe signal changed from green to red

使用した変換テーブルを以下に示す．

	a	b	c	d
1	風	wind	信号	signal
2	北	north	青	green
3	から	from	より	from
4	南	south	赤	red

現実の TDSMT の翻訳結果を見ると，それぞれの変換テーブルにおいて，やや奇異に思われる点がある。

1. “信号” と “風” は，あまり類似していると思えない．
2. “青” と “北” は，あまり類似していると思えない．
3. 通常，“より” は “than” と訳される．
4. 通常，“青” は “blue” と訳される．

このように，実際に利用される変換テーブルは，人間の直感と，少し異なる．

4.5 TDSMT の問題

1. 語彙数
   一般の単語辞書の見出し語は，約 10 万単語である．したがって，変換テーブルの数は 100 億単語が必要になる．（10 万 × 10 万）この数の，精度の高い変換テーブルを収集することは，困難である．
2. 直感との差
   実際に翻訳に利用される変換テーブルは，人間の直感と，少し異なる場合がある 4.4 章．

5 相対的意味論に基づく統計翻訳 (RSMT)

5.1 基本的な考え方

TDSMT は，変換テーブルの数と精度が問題になる．しかし，相対的意味論は，魅力のある考え方である．そこで，この考え方を，文に適用する．そして変換テーブルは使用しない．変換テーブルの代わりに，翻訳確率と類似度を利用する．この翻訳方式を相対的意味論に基づく統計翻訳 Relative Meaning Statistica Machine Translation (以後 RSMT) と呼んでいる．以下に RSMT の基本的な概念を述べる．

- A が B ならば C は D
- ただし A,B,C,D は文

A対訳学習文の日本語
B対訳学習文の英語
C入力日本文
D出力英文（翻訳文）

- A と C は同一言語において類似
- A と B は 2 言語の翻訳において類似
- B と D は同一言語において類似
- C と D は 2 言語の翻訳において類似

以下に例を述べる．

A私は林檎が好きだ
BI like an apple
Cショウジョウハエは林檎が好きだ
DFruit flies like an apple

この例では，以下のように解釈する．

1. “私は林檎が好きだ” と “ショウジョウハエは林檎が好きだ” は類似している．
2. “私は林檎が好きだ” に対する “I like an apple” の翻訳確率は高い．
3. “ショウジョウハエは林檎が好きだ” に対する “Fruit flies like an apple” の翻訳確率は高い．
4. “I like an apple” と “Fruit flies like an apple” は類似している．
5. 以上 4 点を考慮して，“ショウジョウハエは林檎が好きだ” を入力したとき，“Fruit flies like an apple” が翻訳文である．

5.2 実際の RSMT

5.2.1 出力英文

理想的な RSMT において，”D 出力英文” は，生成可能なすべての英文から，選択する．語彙を 10^5 と考えて，1 文は 10 単語で構成されると仮定すると，$10^{50}((10^5)^{10})$ 文から選択することになる．また，A および B は，大量の対訳学習文から，最適な文を選択する．通常は，10^6 文対あたりが想定できる．

この数字は，現実的には計算不可能である．そこで，あらかじめ，”C 入力日本文” を，“別の機械翻訳” で翻訳し，複数の出力英文を得る．これから”D 出力英文” として，RSMT を動作して A と B を選択する．この “別の機械翻訳” には，NMT や Phrase Based SMT や rule base MT などが可能である．したがって実際の RSMT は，複数の翻訳候補文から，最終的な出力英文（翻訳文）を選択する形になる．

5.2.2 文の翻訳確率 $Sim(CD_j)$

文 C と文 D_j の文の翻訳確率は，多くの計算方法が考えられる．基本的には，文の翻訳確率とみなせる．例えば NMT の翻訳確率も利用できる．本論文では，線形性を考えて，以下の式を利用している．

$$Sim(C_m D_{jn}) = \sum_{i=0}^{M}\sum_{j=0}^{N} log2(Count(C_m D_{jn})/(Count(C_m) \times Count(D_{jn})))$$

$Count(C_m D_{jn})$ 対訳データベース中において C_m と D_{jn} が同時に出現する回数
$Count(C_m)$ データベース中の C_m の出現回数
C_m 文 C の m 番目の単語
M 文 C の単語数
D_{jn} 文 D_j の n 番目の単語
N 文 D_j の単語数

5.2.3 文の類似度 $Sim(CA_i)$

文 C と文 A_j の文の類似度は，多くの計算方法が考えられる．最近は Word2vec を利用した距離尺度が多い．また BLEU を利用した距離尺度も考えられる．本論文では，線形性を考慮して，以下の式を利用している．

$$Sim(CA_i) = \sum_{i=0}^{M} CountNum(CA_i)/CountALL(A_i)$$

$CountNum(CA_i)$　文 C と文 A_iが一致する単語数

$CountALLA_i$　文 A_iの単語数

M　文 A の文数

5.3 RSMT の実働例

以下に RSMT の例を示す．C は入力文で，D は出力文（翻訳出力）である．

A信号 が 青 から 赤 に ぱっと 変わった 。
BThe traffic light jumped from green to red .
C信号 が 青 より 赤 に 変わった 。
DThe signal changed from green to red .

- $sim(AC) = -1.807355$
- $sim(BD) = -3.321930$
- $trans(CA) = -75.601800$
- $trans(BD) = -1.041353$

5.4 RSMT の長所

RSMT の長所を挙げる．

1. 修正の容易さ
   翻訳した場合，その翻訳に近い翻訳文が表示される．そのため，誤った翻訳が出力されても，人手で用意に変更箇所が解る．
   A彼 は 我 を 通した 。
   BHe had his own way .
   C彼女 は 我 を 通した 。
   DShe had his own way .
   正解文は “She had her own way” である．この例文では “his” を “her” に書き換えるのは容易と考えている．
2. N-BEST
   RSMT では 1 位候補が間違っていても， 2 位候補が正しい場合が多い．例を以下に示す．
   Cぶらんこ が 揺れ て いる 。
   D_1The swing is flicking .
   A_1炎 が 揺れ て いる 。
   B_1The flame is flickering .
   D_2The swing is swining .
   A_2ハンモック が 揺れ て いた 。
   B_2The hammock was swinging .
   この例では正解が第 2 候補にある．
3. 人手評価と自動評価の差
   自動評価では，評価できない良さがある．人手評価では，通常の NMT を超える場合が多い．

5.5 RSMT の問題点

以下に RSMT の問題点を示す．

ある単語と，ある単語が共起する確率が高いとき，この単語を削除することは，困難である．例を以下の述べる．

A彼 は 我 を 通した 。
BHe had his own way .
C彼女 は 我 を 通した 。
DShe had his own way .

この例では “我” を “his” と翻訳している．学習データに “我” を “his” の共起の確率が高い場合，この “我” を “her” と翻訳することは困難である．

A炎 が 揺れ て いる 。
BThe flame is flickering .
Cぶらんこ が 揺れ て いる 。
DThe swing is flicking .

この例では “ぶらんこ” が “揺れている” を “flicking” と翻訳している．学習データに “揺れている” と “flicking” の共起の確率が高い場合，“揺れている” と “swinging” と訳することは，困難である．

6 考察 - 相対的意味論と NMT

NMT は，現在，尤も普遍的に使われている翻訳システムである．この翻訳システムを，相対的意味論的に考えると，A と B が複数ある場合の C と D の関係として捕らえることが可能である．NMT の学習データは A,B に相当する．C は入力文で，D は出力文（翻訳文）である．そして，文の類似確率と翻訳確率は，非線形関数であり，Nural Network で類推している．つまり，5.1 章における，A,B を明示しない翻訳システムである．

例を以下に示す．学習データを 2 文とする．C は入力文で，D は出力文である．

A_1彼は山にいった．
B_1He went to mountain
A_2彼女は海にいった
B_2She went to sea
C彼は海にいった
DHe went to sea

この例では以下のように考える．

“彼は山にいった．” と “He went to mountain”, と “彼女は海にいった” と “She went to sea” が，存在するから，“彼は海にいった” は “He went to sea” と翻訳できる。

この考え方は，NMT における，別の見識を持てる．

7 まとめ

本論文では，相対的意味論に基づく統計翻訳 (RSMT) の考え方をまとめた．RSMT は，文の相対的意味論を利用している．そして “A が B ならば C は D” を想定している．そして，文の翻訳確率と文の類似度を利用して，出力文（翻訳文）を選択する．ただし，文の翻訳確率と文の類似度には，多くの計算方法がある．また，相対的意味論は，NMT の別の見方が可能になる．また，多くの分野で，応用が可能である．これらの分野を追求していきたい．

参考文献

[1]三浦 しをん. 舟を編む. 光文社 ISBN-10 4334927769, 2011.
[2]安場裕人, 村上仁一. 相対的意味論に基づく変換主導型統計機械翻訳～未知語の出力. 言語処理学会第 25 回年次大会, No. A5-4, 2019.