気象災害ニュースの翻訳方式の検討

後藤 功雄　　田中 英輝
NHK 放送技術研究所

1 はじめに

NHK は, テレビの２ヶ国語放送で英語ニュースを放送している．この英語ニュースを効率的に制作するために自動翻訳の研究を行っている．情報発信のための翻訳は品質が高くなければならない. そこで, ドメインを限定することで高品質な翻訳を目指している．英語の気象災害ニュースは国内在住の外国人にとって重要であり，また，気象災害ニュースは内容が定型的なものが多く，技術的にも扱いやすい．そこで，気象災害ニュースを対象とした自動翻訳システムの研究開発を開始し，翻訳方式を検討したのでその検討結果を報告する．

日本語ニュースは長い複文が多い．このような文を文単位の類似用例で翻訳する場合，数が限られた用例で十分なカバレッジを得ることは困難である．本稿では，日本語の複文が英語ニュースでは複数文に翻訳される場合が多いことに着目し，用例と入力文それぞれを英文の単位で分割して処理することで, 複文を翻訳する手法を提案する．さらに基本的な内容の定型性に着目し，入力記事の分類・抽出・生成により翻訳する手法も提案する．

以下，第２章で２ヶ国語放送用の英語ニュースについて述べ，第３章で検討した翻訳方式について説明し，第４章で関連研究について述べる．

2 英語ニュース

2.1 ニュースライティング

ニュースの翻訳はニュースライティングと呼ばれ, 単なる翻訳とは異なる[1]．本節では，ニュースライティングについて説明する．

２ヶ国語放送の英語ニュースの内容は，できるだけ日本語の放送内容に忠実に沿ったものであり，かつ直訳ではなく，自然な英語ニュースでなければならない．

２ヶ国語放送のためのニュースライティングには次のような原則がある．

- できるだけシンプルな英文とし，１文の長さはあまり長くならないようにする．
- 一般的な英語ニュースのスタイル（話の展開の仕方やことばの選び方など）に準じる．
  例えば，背景から始まってから本題に入る日本語文の場合，英語では本題から始まるようにする．また，日本語では同じ内容を繰り返す場合があるが，英語では必ずしも必要ではない．
- 外国人にはバックグラウンドが必要な場合は，説明を追加する．
  その際，詳細な部分や重複する部分を省略することで日本語ニュース記事と長さを合わせる．
- 放送時に映像と同期させるという制約があるため，ニュース記事中の構成を大幅に変更することはできない．

2.2 ニュースの特徴

本節では，日本語ニュースと英語ニュースの特徴について述べる．

日英対訳の気象災害ニュース 10 記事について調査した結果を以下に示す．なお，ここでは，副詞節を含む文のみを複文とし，それ以外を単文として数えている．

- 日本語の単文と複文
  単文の数は 32 で，複文の数は 44 であった．そのため日本語ニュース文は複文が多いことが分かった．

2.1 節で述べたように，ニュースの翻訳は 1 文を 1 文に翻訳しているわけではない．日本語の 1 文はいくつの英文に翻訳される傾向があるかを調べた．

- 日本語の単文はいくつの英文に翻訳されるか．
  32 文中，29 文は 1 つの英文に，3 文は 2 つの英文に翻訳されていた．これより日本語の単文は１つの英文になることが多いことが分かった．
- 日本語の複文はいくつの英文に翻訳されるか．
  44 文中，13 文は 1 つの英文に，24 文は 2 つの英文に，6 文は 3 つの英文に，1 文は 4 つの英文に翻訳されていた．これより日本語の複文は複数の英文に翻訳されることが多いことが分かった．

英文の中には，複数の文が接続詞などで接続されて 1 つの文になっているものがある．このような英文を接続部分で分割すると構成要素の文が得られる．この構成要素の文と元々分割できない文を基底文または基底英文と呼ぶことにする．分割するのは，カンマまたは等位接続詞またはカンマ＋主節全体に係る従属接続詞による接続の場合とする．

- 日本語の複文はいくつの基底英文に翻訳されるか．
  44 文中，7 文は 1 つの基底英文に，22 文は 2 つの基底英文に，14 文は 3 つの基底英文に，1 文は 4 つの基底英文に翻訳されていた．これより日本語の複文は複数の英語の基底文に翻訳されることが多いことが分かった．

日本語複文から翻訳された基底英文に対応する日本語表現は，どのような単位であるかを調べた．

- 基底英文に対応する日本語表現には副詞節または主節がいくつ含まれるか.
  99 個の基底英文中，7 個は節数 0，82 個は節数 1，7 個は節数 2，3 個は節数 3 であった．ここで，節数が 0 であった 7 個の内訳は，連体節 3 個，連体修飾語 1 個，名詞句 3 個であった．多くの場合，日本語の 1 つの節が 1 つの英語の基底文になっていることが分かった.

ここで，日本語の複文が複数の英文に翻訳された場合（例 1）と 1 つの英文に翻訳された場合（例 2）の例を示す．例 2 は意訳されている部分がある.

例 1 これからあすにかけても各地で強い雨が降る恐れがあり、気象庁は今後の雨に警戒するよう呼びかけています。
Heavy rain is forecast from later tonight until tomorrow.
The Meteorological Agency has issued a heavy rain advisory.

例 2 日本海側を中心に降り続いている雪は北陸を中心に各地で記録的な大雪となり交通機関に乱れが出るなど生活にも大きな影響が出ました。
Record snowfalls in many places along the Sea of Japan, especially in Hokuriku, are affecting transportation and people's everyday lives.

3 気象災害ニュースの翻訳方式

3.1 放送用翻訳システムのねらい

全ての日本語文を自動翻訳することは難しいため，当面の目標として，翻訳システムは翻訳者の仕事の多くの部分を代わりに行うこととする．翻訳者が翻訳する前に自動翻訳するか，翻訳者と協調して自動翻訳[2]し，翻訳者は自動翻訳できない部分だけを翻訳する．これによって，翻訳者の負担を軽減する.

自動翻訳の結果に品質の低い英文が多く含まれていると，その修正作業に手間がかかる．この作業が多いと，自動翻訳は現場で受け入れられない．そのため, 大幅な修正が必要となる可能性が高い部分は, 自動翻訳せずに翻訳者に任せる．この点は，意味が通じることが重要な旅行会話や情報収集のための翻訳とは異なる.

なお, 英語ニュースは,「日本人翻訳者が翻訳→ネイティブが英語表現をチェック→デスクが内容をチェック→放送時にネイティブが日本語ニュースに同期させて英語原稿を読む」という流れで放送されている．自動翻訳の結果は放送前に内容が正しいかを人がチェックする必要があるが，内容の確認は人手で翻訳した場合でも行っている作業である.

3.2 翻訳方式の特徴

2.2 節のニュースの特徴と 3.1 節の翻訳システムのねらいを考慮して検討した 2 つの翻訳方式の特徴を以下に示す.

［基底文を用いた用例による翻訳（EBMT-BS)］

- 英文構造はなるべく生成せずに既存の英文構造をそのまま利用して，用例により翻訳する．用例により翻訳することで，意訳や発想の転換を伴う翻訳にも対応できる.
- 日本語ニュース文は長い複文が多く，日本語文単位の用例では，十分なカバレッジを得るのは困難である．そこで，英文構造を生成しなくてもよい最小単位，すなわち基底英文を単位として用例を分割して利用する．ただし，引用表現については，節を合成して英文を生成することが容易なため，分離して扱う.

［分類・抽出・生成による翻訳（CEGMT)］

- 気象災害ニュースの主な話題毎に翻訳すべき基本的な情報を定義しておき，入力記事の話題を分類することで，入力文からどのような情報を抽出すべきかを決定する．そして，それらを自動抽出する.
- 入力文から抽出された情報とテンプレートの変数との一致を確認することで，類似用例で翻訳する場合（e.g. [3]）と比べて，基本的な情報を翻訳できているか判別できる利点がある.

3.3 翻訳方式の説明

翻訳全体の流れは次の通りである.

1） EBMT-BS で翻訳する.
2） 翻訳できない場合は，CEGMT で翻訳する.
3） 最後に記事単位で英語を編集する.

なお，翻訳の実行は，全自動の処理である．ただし翻訳の事前準備は，できるだけ自動の処理とするが，翻訳の精度を高めるために必要に応じて人手で修正することも考えている.

以下，EBMT-BS，CEGMT，記事単位での英語の編集について説明する.

3.3.1 基底文を用いた用例による翻訳（EBMT-BS）

本節では, EBMT-BS での用例の作成, 翻訳の流れ, 翻訳単位毎の変換処理について説明する.

［用例の作成］

1） 用例として記事対応と文対応がついた対訳ニュースコーパスを用意する.
2） 場所，時間，日付などの基本的な情報を表す表現を特定し，意味属性を付与する（例 3）.
3） 並列に列挙している表現を特定し，構造化する（例 4）.
4） 引用表現を特定して分離する（例 5）.
5） 英語の代名詞の照応を解析し，他の文中の表現を指していれば，具体的な表現に置き換える.
6） 日英ともに地名や時間，数値表現を抽出し，対訳辞書を利用してそれらの単語対応をつける．さらに，対訳辞書，対数尤度比[4]などの統計情報，部分的な構文情報を用いて，そのほかの日英単語対応付けを行う（例 6）.
7） 並列に列挙している表現がある場合はまとめる（例 7）.
8） 日本語 1 文に対して，英語が複数の基底文からなる用例は，英語の基底文の単位で用例を分割

する（例 8）．日本語を分割する際に，英語に合わせて主語や提題の補完が必要であれば行う．

例 3 <場所>日本海側では東北や北陸の山沿いを中心に</場所>断続的に雪が降っています.
It is snowing on and off <場所>mostly in the mountains in Tohoku and Hokuriku along the Sea of Japan</場所>.

例 4 午後六時現在の積雪量は<list><item>▼甲府市で三十八センチ，</item><item>▼福島市で二十五センチ，</item><item>▼東京の都心でも七センチ</item></list>となっています.
By six o'clock this evening, <list><item>38 centimeters of snow had piled up in Kofu</item>, <item>25 centimeters in the city of Fukushima</item>, and <item>seven centimeters in central Tokyo</item></list>.

例 5 <引用表現>ＪＲによりますと、</引用表現>東海道・山陽新幹線のダイヤの乱れはきょう一杯続く見込みだ<引用表現>ということです。</引用表現>
<引用表現>The Japan Railway Company says</引用表現> the Tokaido Sanyo Shinkansen services will be disrupted until the last train tonight.

例 6
場所　日付
(日本海側)の(大雪)は(今月十三日から)(降り始めました).
(Heavy snow) (began to fall) (along the Sea of Japan) (on Saturday).
場所　日付

例 7 午後六時現在の積雪量は<list>(<item>▼<場所>甲府市</場所>(で|でも)<数値>三十八</数値>センチ(，)?</item>)+</list>となっています。
By six o'clock this evening, <list><item><数値>38</数値> centimeters of snow had piled up in <場所>Kofu</場所></item>(, <item><数値>25</数値> centimeters in <場所>the city of Fukushima</場所></item>)*</list>.
（ここで，記号 ()|?* は正規表現を示す.）

例 8 東海道・山陽新幹線は台風のため、三回にわたって運転を見合わせた影響で、これまでに十八本の列車が運休するなどダイヤが大幅に乱れています。
↓
東海道・山陽新幹線はダイヤが大幅に乱れています。
Tokaido Sanyo Shinkansen train services have been disrupted.
東海道・山陽新幹線は台風のため、三回にわたって運転を見合わせた影響で、
The shinkansen bullet trains had to suspend operations three times today due to the typhoon.
これまでに十八本の列車が運休するなど
18 trains have been cancelled so far.

［翻訳の流れ］

1) 場所や時間などの表現を特定して意味属性を付与する．
2) 並列表現を特定し，1 つにまとめる．
3) 引用表現を特定し，分離する．
4) 様々な翻訳単位（文，1 つ以上の節やそれらに主語や提題を補完したもの，並列句）で翻訳し，翻訳単位毎のスコアから入力文単位のスコアを計算する．ここでの翻訳単位毎の翻訳処理は，次の［翻訳単位毎の変換処理］で説明する．
5) 入力文単位のスコアが最も高い英文を選択し，スコアが閾値以上の場合に翻訳処理を続行する．
6) まとめた並列表現や分離した引用表現を英文に反映させ，翻訳結果として出力する．

［翻訳単位毎の変換処理］

1) 入力表現の述語と用例の述語が一致または類似する用例を取得する．ただし，英文の述語が用例の日本語の述語以外に対応する場合（例 9）は，その部分も入力文と一致する場合のみ取得する．
2) 入力表現と取得した用例との距離を計算し，距離が小さい上位の用例を選択する．
距離の計算方法は以下の通りである．
   - ✧ 距離は，用例の日本語表現において同じ構文となるように語順を入れ替えた場合も含めて，最小となる編集距離とする．
   - ✧ 用例で英訳時に省略されている日本語表現は，削除コストを 0 とする．
   - ✧ 意味属性が一致する表現やシソーラスで意味が近い表現は，置換コストを低くする．
3) 各用例の日本語表現を置換，削除，挿入により入力表現へ編集し，対応する英語表現もそれにあわせて日英単語対応を用いて編集する（例 10）．
英語の編集時には，日英単語対応の信頼性の高さに基づいた英語編集コストを計算する．このコストは，信頼性が高ければ小さく，低ければ大きくする．
挿入する際には，英語の構文構造を解析し，挿入する語の修飾関係が正しくなる位置のみを挿入位置の候補とする．候補が複数ある場合は，言語モデルなどの統計情報を利用して，最適な挿入位置を決定する．
4) 距離と英語編集コストに基づいて翻訳単位毎のスコアを決定する．このスコアは選択した用例中で最高のものとする．スコアを計算する具体的な評価式は，現段階ではまだ決まっていない．

例 9 発達中の低気圧が関東の南の海上を進んでいるため、
A low-pressure system is developing off the Kanto coast.

例 10
（入力文）
(気象庁によりますと)低気圧が日本付近を通過するためこれからあすにかけても北日本の太平洋側を中心にまとまった雪が降る恐れがある(ということです。)
⇩ 距離が小さくなるように入力文の引用表現以外を1つ以上の節または並列句に分割し，その類似用例を取得
（分割した入力文）　（類似用例の日本語文）
低気圧が日本付近を通過するため ⇔ 低気圧が沖縄付近を通過する
これからあすにかけても北日本の太平洋側を中心にまとまった雪が降る恐れがある ⇔ あすは北日本で大雪が降る恐れがある
（類似用例の英語文）
A low air pressure system passes close to Okinawa.
Heavy snow will fall in northern Japan tomorrow.
⇩ 入力文と異なる部分を編集
A low air pressure system passes close to Japan.
Heavy snow will fall along the Pacific Ocean in northern Japan from later tonight until tomorrow.

3.3.2 分類・抽出・生成による翻訳（CEGMT）

本節では，CEGMT での話題毎のテンプレートの作成と翻訳の流れについて説明する．

［話題毎のテンプレートの作成］

1） 気象災害ニュースの主な話題を特定する．

2） 各話題について，その話題で基本的な内容を示す定型的な数値や地名などの表現が変数となっている英語テンプレートを作成しておく（例 11）．各話題に属するテンプレートに対応する変数の集合を，その話題での基本的な情報とする．

例 11 台風の話題の場合
変数：中心付近の最大風速（単位=”メートル／秒”）
英語テンプレート : The central barometric reading is (　)<中心付近の最大風速 単位=”キロ／時”/> kilometer(s) per hour.

気象災害ニュースの主な話題を特定するため，「社会」分野のニュース 5 年分約 6 万記事をクラスタリングして観察した．その結果，主な話題は表 1 のようなものであることが分かった．

表 1 気象災害ニュースの主な話題

分類	主な話題
気象状況	台風，雨，雪，風，震度速報，余震，津波観測結果，津波予想，海外地震による津波情報
気象による影響	空の便（欠航），鉄道（運休，運転見合せ，遅れ），海の便（欠航），高速・有料道路（通行止め，通行規制）
被害状況	被害（死者人数，けが人人数，全壊した建物数，一部が壊れた建物数，床上浸水数，停電世帯数，・・・）
警報・注意報の伝達	津波警報が出る，津波警報が解除，津波注意報が出る，津波注意報が解除，大雨警報が出る，・・・

［翻訳の流れ］

1） 入力記事の話題を分類する．

2） その話題で基本的な情報が入力文中に存在するかを識別して抽出する（例 12）．

3） 情報が抽出された場合は，それらを対訳辞書で翻訳し，変数の種類が一致するテンプレートに挿入して英語を生成する（例 13）．変数が数値の場合で単位の変換が必要であれば値を変換する．

例 12 台風の接近に伴って<場所>高知県の室戸岬</場所>で<時間>午後四時五十分頃</時間>に<最大瞬間風速 単位=”メートル／秒”>三十三点三</最大瞬間風速>メートルの最大瞬間風速を観測しました。

例 13 Winds of up to about (120)<最大瞬間風速 単位=”時速”/> kilometers per hour were observed at (Cape Muroto in Koch Prefecture)<場所/> at (about 4:50pm)<時間/>.

3.3.3 記事単位での英語の編集

文単位では正しく訳せていても，重複する表現が多いと自然な英語にならない．そこで，代名詞化（例 14）や文の接続（例 15）を行う．

例 14 Typhoon number eleven was spotted 230 kilometers east-south-east of Tanega-shima Island.
Typhoon number eleven is proceeding north at a speed of 15 kilometers per hour.
↓
Typhoon number eleven was spotted 230 kilometers east-south-east of Tanega-shima Island.
It is proceeding north at a speed of 15 kilometers per hour.

例 15 The low-pressure system is still developing.
The low-pressure system has brought gusty winds to Kanto.
↓
The low-pressure system is still developing, and has brought gusty winds to Kanto.

4 関連研究

日本語長文を短文に分割することで，日英翻訳の精度を向上させる手法が提案されている[5]．この手法は，生成する英語を考慮せずに，日本語側の情報だけで文を分割する．我々の手法は，生成する英語を考慮して入力文を分割する．

独立した英文が連続したり，複数の英文が接続詞でつながれた発話を入力とした場合に，用例の単位を考慮して発話を分割して翻訳する手法がある[6]．我々は，入力として日本語の複文を対象としている．また，用例を英文に合わせて分割して利用する．

入力文と部分的に一致する用例を組み合わせて英文を生成する手法[7]や複文をパターンにより翻訳する手法[8]がある．これらの手法では，日本語複文を複数の英文に翻訳することは考慮されていない．

また，生成は変数の値を入力とするため，日本語ニュースが入力の場合，生成では翻訳できない．パターンによる翻訳（e.g. [8]）では，入力文と一致するパターンが存在する場合のみ，入力文中で変数として扱う部分が決まる．我々の CEGMT では，翻訳元言語側のパターンを必要とせず，話題の分類と情報抽出により，入力文中の基本的な情報を特定する．

5 おわりに

日本語ニュースを２ヶ国語放送用の英語ニュースへ自動翻訳する方式について検討した．日本語ニュースには複文が多いことを示し，複文を翻訳する方式として基底文を用いた用例による翻訳方式を提案した．また，話題毎の基本的な内容の定型性に着目した，分類・抽出・生成による翻訳方式も提案した．

現在は，翻訳システムで利用する気象災害ニュースの対訳データベースを構築中である．今後は，システムを構築して評価を行う予定である．

参考文献

[1] NHK「ニュース７」「ニュース９」の２カ国語放送の制作現場を拝見！，翻訳辞典 2000 年度版，アルク，pp.67-74, 1999.
[2] 熊野ほか,「翻訳部品の配置による翻訳作業」を目指した翻訳統合環境の提案，言語処理学会第 13 回年次大会，2007.
[3] Sumita, Example-based machine translation using DP-matching between word sequences, 39th ACL workshop on DDMT, pp.1-8, 2001.
[4] Melamed, Models of Translational Equivalence among Words, Computational Linguistics, Vol.26, No.2, pp.221–249, 2000.
[5] 金ほか，日英機械翻訳のための日本語長文自動短文分割と主語の補完，情報処理学会論文誌，Vol.35, No.6, pp.1018-1028，1994.
[6] Doi et al., Splitting Input for Machine Translation Using N-gram Language Model Together with Utterance Similarity, IEICE Trans. Inf. & Syst., Vol.E88-D, No.6, pp.1256-1264, 2005.
[7] 荒牧ほか, 用例ベース翻訳の確率的モデル化, 自然言語処理, Vol.13，No.3，pp.3-19，2006.
[8] 池原ほか，非線形な表現構造に着目した重文と複文の日英文型パターン化，自然言語処理，Vol.11，No.3，pp.69-95，2004.