文書構造抽出を用いる機械翻訳システム

4K－3

野垣内 出　井ノ上 直己　鈴木 雅実
KDD 研究所

1　はじめに

翻訳を援助するための総合的なソフトウェアである翻訳支援環境の一環として、文書構造の利用が提案 [1] [2] されている。この提案は、文外情報として文章の清書情報である段落や見出しやフォント情報などを翻訳精度の向上に利用しようとするものである。

これまでの機械翻訳への文書構造の情報の利用は特定の文書清書処理に依存しているために、一般の文書やOCRから読み込んだ文書については、適用できない。この報告では、一般的な文書にも適用可能な文書構造抽出を用いた機械翻訳の方法および試行実験について述べる。

2　文書構造抽出を用いる機械翻訳

2.1　文書構造

文書構造の解析は一般的には、文書中の図がどの章に含まれるか、章ごとの構成を含む階層構造までを含むことが多いが、ここでは、文書のある箇所が「見だし」「段落」「本文」などであることと示されることを文書構造の解析と呼ぶこととする。

文書構造情報は、文書構造情報の入った文書清書用のテキスト、例えば、Tex [3]、LaTex [4] や ROFF [5] のソーステキストであれば直接得られる。しかし、一般的なテキストやOCRなどからのテキストでは、文書構造抽出を行なう必要がある。

文書構造の抽出は、これまで総合的な文書処理のためのものとして、研究されている。これらの応用のために、テキストグラマの利用、文書構造の詳細な規則化による解析 [6] [7] などが提案されている。

翻訳支援の面からは、詳細な文書構造の種類や段落間の構造などは、必要ではない。ここでは、主に位置情報から文書のブロック構造の抽出を行ない、その情報をもとに文書構造を解析する方法 [8] を用いている。

A Machine Translation System by Using a Text Structure Analysis
Izuru Nogaito, Naomi Inoue and Masami Suzuki
KDD R&D Laboratories
noga@kddnews.wlb.kddlabs.co.jp

2.2　機械翻訳と文書構造

機械翻訳の対象である文書は、計算機マニュアルやビジネスレターのように、文書形式の定まった文書であることが多い。特に計算機や通信機器のマニュアルにおいては、一定の形式を持ち、「見だし」や「個条書き」の使用の比率が高い。表1にマニュアル4種 (それぞれ5ページ相当) におけるこれらの比率を示す。ここに示した「見だし」「個条書き」は、英語において名詞と動詞が同型であることから誤訳の多い項目である。例えば、「見だし」では、名詞形となるべきものが、命令形の文の訳を得るなどの例が少なくない。

文書中のある箇所が「見出し」であることが判明していれば、その文は、名詞句となっていることが仮定できる。このように、文書構造の情報は、翻訳対象に文法的な予測を与えることで、翻訳品質の低下を防ぐことへ利用できると考えれれる。

表 1: マニュアルにおける文書構造の比率

マニュアル名	見だしの数	個条書きの数	文の数
M	23 (28%)	5 (6%)	82
O	34 (34%)	7 (7%)	105
V	15 (18%)	10 (12%)	85
X	7 (7%)	16 (15%)	107
平均	79 (21%)	38 (10%)	379

2.3　位置情報を用いる文書構造の解析

ソフトウェアおよび通信関連のマニュアルに翻訳対象を限定して、翻訳として必要な文書構造の種類を検討した。検討された文書構造の項目は、例えば、章と段落の見出しは細分化せず1項目にするなど種類は少ない。今回の規則作成では、「フッターおよびヘッダー」、「章や節の名前や見出し」、「個条書き」、「本文」を暫定的な文書構造の項目とした。

次に文書構造の解析処理の概要を示す。文書構造の解析は、テキストを行単位に解析を行ない、1行ごとにタブや複数のスペース記号で区切られた処理単位の開始位置、終了位置を得る。直前の行に同じ開始位置や終了位置のものがあれば、同じ文書構造の項目を与える。

なければ、文書構造の項目ごとの解析規則で判断を行なう。すなわち、開始位置や文頭が特定の形(数字や記号)であるか、文長、文書全体の処理単位の開始位置と終了位置の傾向などから判断する解析規則で文書構造の項目を与える。作成した解析規則の有効性を確認するために予備実験を行なった。規則はソフトウェアマニュアル(オンラインマニュアル)の数項目を参考に作成した。

作成者の異なる7種のソフトウェアのマニュアルと2種の通信機器(モデム)のマニュアルについて解析実験を行なった。各マニュアルとも5ページ相当を無作為に選んだ。解析実験結果を表2に示す。文書構造数は、人手により文書に文書構造を与えた際の数、正解析数は、人手により与えた文書構造項目と解析結果の一致数、構造解析率はこれらの比である。

表 2: 文書構造解析実験

マニュアル名	文書構造数	正解析数	構造解析率
A	46	43	93%
C	91	64	70%
CD	83	77	93%
CL	53	48	91%
M	44	39	89%
O	60	52	87%
S	53	52	98%
V	44	42	95%
X	44	42	95%
平均	58	51	90%

2.4 文法項目の指定

文書構造解析の情報を機械翻訳システムへ文法項目の指定の形式で与える。機械翻訳システムは、文法項目(名詞句や命令形など)を指定するとシステムは、その部分の文法項目が指定と一致するものを優先的に処理を行なう。文書構造結果は、文書構造名から文法項目へ変換され、入力された文に付加される。図1に文法項目の指定の例を示す。個条書きの例では、前半が記号や数字である、後半は文となることを指示している。見だしの例では、全体で名詞句となることを指示している。

文書構造名と文法項目指定の対応は、現在、「見だし」「目次」「ヘッダーおよびフッター」はそれぞれ、名詞句」、「個条書き」は、「記号や数字と文」、「本文」は「文」となっている。

2.5 検討

翻訳試行では、文書構造を抽出することで翻訳品質の向上がみられる。この向上は、翻訳対象に含まれる「見だし」「個条書き」「目次」などの含まれる比率に大きく依存している。特に「見だし」「個条書き」の多く含まれるものほど、翻訳品質は向上する。

「個条書き」

原文

2. Move the program bianary to pro.ow2.

文法項目指示

[+NUMB 2.]

[+TEXT Move the program bianary to pro.ow2.]

「見だし」

原文

Copying Other Window

文法項目指示

[+NP Copying Other Window]

図 1: 文法項目の指示

実験に用いた機械翻訳システム[9]では、解析部分においてすべての解析結果が得られる。この場合、文法項目の指定はすべての可能性から文法項目に一致するものを取り出すことに対応している。また、指定した文法項目が誤りなどで、その文法項目で解析ができない場合には、もっとも文法的な解釈が得られる。したがって、文書構造解析の誤りの影響は少ない。

3 おわりに

文書構造解析を行なう機械翻訳の方法の概要を述べた。文書構造の抽出は、文書全体とその部分の位置関係をもとに推定を行なう。また文法構造抽出結果は、文法項目に変換されて、機械翻訳システムの解析において、その部分の文法項目として優先して解釈を行なう。

今後、計算機や通信機器のマニュアルの他、ビジネスレター、電子メイル、電子ニュースなどにも対象を拡げて実験を行なう。

参考文献

[1] 中村順一: "機械翻訳のソフトウェア環境", 情処,Vol.26,No.10,pp.1184-1190(1985-10)

[2] 西野文人. 他: "機械翻訳使用のための総合的環境", 情処研報, 自然言語処理 85-13(1991-09)

[3] Donald E.Knuth: "The TEXBook",Addison-Wesley Publishing(1983)

[4] Leslie Lamport: "LATEX: A Document Preparation System",Addison-Wesley Publishing(1986)

[5] Sun Microsystems: "Using NROFF & TROFF",Sun Microsystems(1990)

[6] 山田 満: "文書画像の ODA 論理構造化文書への変換方式", 信学論 (D-II),J76-D-II,11,pp.2274-2284(1993-11)

[7] 土井美和子. 他: "文書構造抽出技法の開発", 信学論 (D-II),J76-D-II,9,pp.2042-2052(1993-09)

[8] 野垣内 出. 他: "機械翻訳のための文書構造の解析", 信学全国大会,D-118,6,pp.6-118(1994-03)

[9] 榊 博史: "コンピュータ翻訳技術", 電子情報通信学会,(1993-12)