自然言語処理 120－2
(1997. 7. 24)

機械翻訳における事例を用いた多義語の訳語選択手法

水野秀紀† 荒木健治†† 宮永喜一† 栃内香次†
†北海道大学大学院工学研究科
〒060 札幌市北区
北13条西8丁目
††北海学園大学工学部
〒064 札幌市中央区
南26条西11丁目
E-mail: mizuno@hudk.hokudai.ac.jp

あらまし 本稿では、英語から日本語への機械翻訳における多義性解消について、事例データに基づく手法を提案する。本手法では、すでに構文解析の済んでいる入力文の主語・直接目的語・間接目的語・補語・前置詞の各語で事例を検索し、入力文と事例との距離を算出する。検索する際に、シソーラスを用いて同義語、上位語、下位語の抽出を行なう。また距離は、シソーラスの概念階層を基に定義されている。最適解の選択は、各語の検索で求まった距離を重みづけをした線形和をとった値で行なう。実験では、事例の全くない状態から300文行なって70%以上の正解率が得られた。

キーワード 機械翻訳、多義語、事例、シソーラス、距離

Method for Example-Based Word Selection of Multi Meaning Words in Machine Translation

Hideki Mizuno†, Kenji Araki††, Yoshikazu Miyanaga†, Koji Tochinai†
†Faculty of Engineering
Hokkaido University
N13, W8, Kita-ku
Sapporo, 060 Japan
††Faculty of Engineering
Hokkai-Gakuen University
S26, W11, Chuo-ku
Sapporo, 064 Japan
E-mail: mizuno@hudk.hokudai.ac.jp

Abstract This paper proposes an example-based method for word sense disambiguation in machine translation. This system refers to examples by a subject, direct object, indirect object, complement, and preposition in a parsed input sentence, extracting synonyms, hypernyms, and hyponyms from a thesaurus, and calculates a distance between input sentence and examples. The distance is defined based on a class of concept in the thesaurus. An optimum selection is made by use of values of the weighted linear sum of distances of each words. According to experiments, this system's success rate is more than 70%.

key words machine translation, multi meaning word, example, thesaurus, distance

1 はじめに

近年、様々な機械翻訳システムが研究、開発されている。しかしながら、文に内在する意味的な多義性の解消、特に動詞の多義性解消については、様々な研究[1]、[2]が行なわれているが、未だ解決されておらず、自然言語処理分野における困難かつ重要な問題となっている。

従来、機械翻訳の手法は、文法的な解析に基づく手法[3]が主であった。この手法は、入力文の形態素解析、構文解析を行なって入力文の構文木を生成し、それを目標言語の構文木に変換して翻訳を行なう手法で、文法的に正しい翻訳結果が得られるが、膨大で複雑な文法情報と構文規則を辞書中に登録する必要があり、それらの制御が難しい。また、複雑で例外的な文には対応できないという問題点がある[4]。

このような問題点を解決するため、最近、事例データに基づく手法[4]、[5]、[6]が開発されている。この手法は、入力文と事例データとの間で何らかの距離を計算して、最も距離の近い事例を用いて解の選択を行なう手法で、事例データが解析された文の集合であるため、容易に事例の追加や削除ができ、複雑な文にも対応できる。

しかしながら、事例データに基づく手法は、大量の事例を必要とすることや、距離の定義法に多くのバリエーションが考えられ、どれが最も適切であるか分かっていないという問題点が挙げられる。そこで、我々はシソーラスによる事例の適用範囲拡大を試みている。また、距離に関しては、優先度[1]や意味距離計算[6]のようなシソーラスの概念階層における距離を基に定めている。

本稿では、英語から日本語への機械翻訳における、主に動詞の多義性の解消について、事例データに基づく手法を提案し、その有効性を確認するために行なった実験結果について述べる。

2 システムの概要

本システムにおける多義性は、次のように定義される。

- 動詞に関しては、英和辞典[7]に複数個の意味が記されている場合。
- その他の語に関しては、シソーラスに複数個の意味が記されている場合。

本システムの構成を図1に示す。英文が入力されると、動詞が入力文と同じである事例を事例データから抽出する。そして、主語・直接目的語・間接目的語・補語・前置詞の各語に関して抽出した事例を検索し、入力文とそれぞれの事例との距離を算出する。その際、シソーラスから入力文の動詞以外の各語に関して、同義語、上位語、下位語を抽出をして検索を行なう。そして、算出された距離を基に最適解の選択を行なう。

本システムに入力される英文は、構文解析がすでに行なわれているものとするが、多義性はどの語も解消されていないものとする。構文解析については、人間が行なうため100%の正解が得られている。

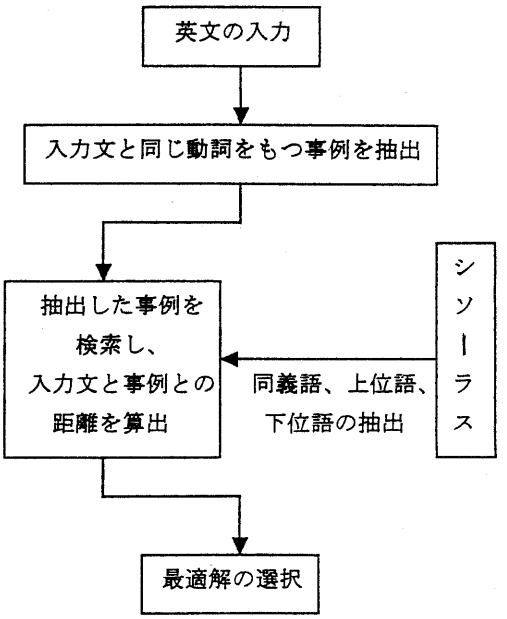

図 1: システムの構成

3 事例データ

事例データを用いる利点は、1章でも述べたように、容易に事例の追加や削除ができ、複雑な文にも対応できることである。また、一度多義性解消に失敗した文も修正して事例データに加えれば、その文に関しては

二度と間違うことはない。

本システムで用いる事例データの例を図 2 に示す。事例データには、それぞれの事例ごとに主語、動詞、直接目的語、間接目的語、補語、前置詞についてその概念番号と共に記してある。概念番号については、4 章で述べる。また、受動態については能動態に人間の手によって変換してある。その根拠は、受動態の動詞は全て be 動詞になってしまい、動詞の多義性解消には不適切だからである。図 2 の例 2 で、define を動詞とすると、直接目的語は behavior ということになる。それから、代名詞に関しては、その代名詞が指す語句を人間が解析し、その解析結果を記してある。その根拠は、シソーラスに代名詞に関する記述がないため同義語、上位語、下位語を抽出できないためであり、また、文脈に依存せずに解析ができるようにするためである。図 2 の例 3 がその例である。

例 1
The C compiler will produce an error message.
S V DO

例 2
The behavior of *printf* is defined in the ANSI
DO V P standard.

例 3
File inclusion makes it easy to handle collections
S V C DO
of *#defines* and declarations.

図 2: 事例データの例

4 シソーラス

全ての動詞の各語義に関して、十分な数の事例を用意することは困難である。そこで、各語義の上位、下位、同義関係を用いて、事例の適用範囲を広げる必要がある。そのような関係を抽出するのにシソーラスを用いる。

本システムで用いたシソーラスは、Princeton 大学で開発された WordNet である。このシソーラスの特徴は、入力語に対する同義語、上位語、下位語がその意味ごとに木構造になっていて、それぞれの語の意味も英語で出力される。3章で述べた概念番号は、出力された意味の番号を指す。WordNet の語数は、名詞が 87,641 語、動詞が 14,726 語、形容詞が 19,100 語、副詞が 5,049 語であり、合計 126,516 語である。

5 距離の定義

距離については、1章でも述べたように、シソーラスの概念階層を基に次のように定義する。

- 任意の語とそれに隣接する上位語 (下位語) との間の距離は全て等しい。
- 任意の語とその同義語との間の距離は全て等しい。
- 任意の語とそれ以外の任意の語との距離は共通する上位語までのそれぞれの距離の和とする。

この定義に従って、次のような例を考えてみる。図3において、A、B、C、D、E、F、G、Hはそれぞれ同義語の集合とする。Bについて考えてみると、Aは一つ上の上位語の集合であり、D、E、Fはそれぞれ一つ下の下位語の集合である。また、Cについて同様に考えてみると、Aは一つ上の上位語の集合であり、G、Hはそれぞれ一つ下の下位語の集合である。AとBとの間の距離をAB、BとDとの間の距離をBD、BとEとの間の距離をBE、BとFとの間の距離をBF、BとCとの間の距離をBC、BとGとの間の距離をBGとすると、

$$AB = BD = BE = BF \tag{1}$$

$$BC = 2AB \tag{2}$$

$$BG = 3AB \tag{3}$$

となる。

また、Bの中にp_1、p_2、p_3という語句が含まれていて、p_1とp_2との間の距離をp_1p_2、p_1とp_3との距離をp_1p_3、p_2とp_3との距離をp_2p_3とすると、

$$p_1p_2 = p_1p_3 = p_2p_3 \tag{4}$$

となる。

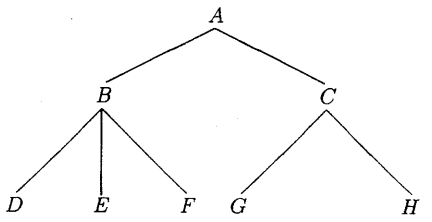

図 3: 距離

6 最適解の選択

最適解は、各事例と入力文との距離を式(5)のような重みづけをした線形和をとった値で選択する。その根拠は、重みづけをすることによって動詞の意味を決定する際にどの語とのつながりを優先するかを考慮に入れることができるからである。

$$Y = \sum_{i=1}^{N} w_i X_i \qquad (5)$$

X_i はそれぞれの語における事例と入力文との距離を表し、w_i は重み係数を表す。また N は検索範囲の語数を表す。

7 実験

7.1 実験方法

実験は、事例データの全くない状態から300文行なった。実験に使う文は、C言語のマニュアル[8]、[9]の文を使用した。その根拠は、類似の文が数多く存在し同じ表現が複数回出現するからであり、また、簡単な英語を使っているからである。主語・直接目的語・間接目的語・補語・前置詞の各語に関して、考えられる全てのパターンを検索範囲として実験を行なった。また、WordNetのindexファイル、dataファイルにより同義語・上位語・下位語の抽出の自動化を行なった。

距離は、3章の定義に従って、入力文の語そのもので見つかった場合に0、同義語で見つかった場合に0.5、1つ上(下)の上位語(下位語)で見つかった場合に1という値を与え、その後1つ上(下)へいくごとに距離は1増えていくものとした。また、見つからなかった場合を20とした。

式(5)における重み係数は、直接目的語の重み係数に2.0、主語の重み係数に1.6、補語の重み係数に1.4、間接目的語の重み係数に1.0、前置詞の重み係数に0.5という値を与えた。この値の根拠は、直接目的語に関しては、人間が英文における動詞の意味を決定する際に直接目的語に依存することが多いので最も大きい値とし、前置詞に関しては、距離が0か20であり、表層構造のみの検索なので最も小さい値とした。また、その他の語に関しては、人間が動詞の意味を決定する際に、どの語とのつながりを優先しているかを考慮に入れた。値そのものに関しては、200文の実験の後、上下に値を動かしたところ、この値が最適であると分かったので、この値に決定した。

最適解は、式(5)が最小値をとるものを選択し、選択された最適解と人間が行なった解析結果が一致した場合を正解とした。また、実験後解析した文を修正して、事例データに加えた。

7.2 実験結果

表1: 実験結果

検索範囲	正解率	閾値より大	不能
DSCIP	77.3% (59.6%)	42文	2文
DSCI	76.7% (60.0%)	50文	1文
DSCP	77.3% (59.6%)	55文	2文
DSIP	74.8% (53.7%)	56文	2文
DCIP	75.7% (58.7%)	31文	6文
SCIP	69.4% (47.7%)	27文	10文
◎DSC	77.9% (60.0%)	52文	1文
DSI	72.6% (51.7%)	48文	2文
DSP	75.0% (53.7%)	55文	2文
DCI	73.3% (57.6%)	22文	7文
DCP	74.4% (57.8%)	27文	6文
DIP	73.0% (56.7%)	20文	7文
SCI	67.8% (42.2%)	34文	13文
SCP	68.8% (45.2%)	42文	10文
SIP	63.5% (36.4%)	35文	13文
CIP	55.8% (17.6%)	11文	50文
DS	73.0% (52.5%)	51文	2文
DC	73.3% (57.6%)	22文	7文
DI	71.1% (55.1%)	17文	9文
DP	72.8% (56.7%)	21文	7文
SC	69.3% (45.3%)	33文	12文
SI	64.9% (37.9%)	32文	16文
SP	66.2% (41.5%)	35文	11文
CI	50.0% (4.2%)	9文	65文
CP	55.8% (17.4%)	11文	50文
IP	52.5% (12.3%)	0文	58文
D	70.7% (53.6%)	16文	9文
S	63.2% (33.8%)	28文	19文
C	50.6% (4.3%)	11文	64文
I	47.8% (0.0%)	1文	72文
P	55.9% (18.5%)	56文	37文

実験結果を表1に示す。Dは直接目的語、Sは主語、Cは補語、Iは間接目的語、Pは前置詞をそれぞれ表す。()の内の正解率は、入力文の動詞に関して事例データにその意味が2つ以上登録されている場合の正

解率である。不能とは、式(5)の値の同じ事例が2つ以上あって最適解を判定できない場合である。◎は、正解率の最も高かった検索範囲を表す。

入力文と同じ動詞をもつ事例を抽出することができなかったものが117文(39.0%)あった。

8 考察

8.1 検索範囲の検討

表1より最も正解率の高い検索範囲は、直接目的語・主語・補語の組み合わせということになる。直接目的語が検索範囲になっているものは、どの場合も正解率が70%以上と高くなっていて、不能になる文も少ないことが分かる。これは、動詞の意味を決定する際に直接目的語に依存することが多く、また事例データに登録されている直接目的語の種類が他の語に比べて非常に多いためであると考えられる。一方で、間接目的語が検索範囲になっているものは、正解率が低くなっている。これは、間接目的語を入力文がもっている場合や入力文はもってないが抽出された事例がもっている場合に、距離が20になってしまって正解になるはずなのに閾値を越えてしまっているものがあるためである。検索範囲が間接目的語のみのときの表1の()内の正解率が0.0%になっているのは、間接目的語をもった事例がほとんどないために、判定不能になってしまっているためである。また、検索範囲の語数が増えると正解率も高くなり、不能になる文も少なくなることが分かる。適当な検索範囲としては、検索範囲の語数が3語以上でその中に直接目的語と主語が含まれていると良いということが分かる。

8.2 失敗の原因

検索範囲の語数が最も多い直接目的語・主語・補語・間接目的語・前置詞の組合せのときと、正解率の最も高かった直接目的語・主語・補語の組合せのときに解析に失敗した原因をそれぞれ表2、表3に示す。表2、表3において、重み係数とは、重み係数を変えることで正解になるものであり、このことは全ての重み係数を0.1から2.0に変えることで確かめることができた。一方、正解の事例の方が距離が大きいとは、重み係数を変えても正解にならないものである。また、動詞の意味のみ異なるとは、判定不能になっている複数の事例に関して、動詞以外の語についてはその概念番号までもが一致しているのに動詞の意味のみが異なっているため判定できなくなったものである。

失敗の原因で最も多かった、「入力文と同じ動詞の事例はあるが正解となるべき意味のものはない」という原因は、事例不足によるものである。また、入力文と同じ動詞をもつ事例を抽出できなかったものが117文あり、これも事例不足によるものである。これらを解決するためには、事例に入力文と同じ意味をもった同じ動詞が登録されている必要があるため、かなり多量の事例を必要とすると考えられる。

表2: 失敗の原因(D S C I P)

出力	原因	文の数
不正解	入力文と同じ動詞の事例はあるが正解となるべき意味のものはない	23文
	重み係数	4文
	正解の事例の方が距離が大きい	3文
判定不能	動詞の意味のみ異なる	1文
	重み係数	1文

表3: 失敗の原因(D S C)

出力	原因	文の数
不正解	入力文と同じ動詞の事例はあるが正解となるべき意味のものはない	21文
	重み係数	4文
	正解の事例の方が距離が大きい	3文
判定不能	動詞の意味のみ異なる	1文

検索範囲が直接目的語・主語・補語・間接目的語・前置詞の組合せのときに、不正解になった具体例をそれぞれの原因ごとに次に示す。

例1 (入力文と同じ動詞の事例はあるが正解となるべき意味のものはない)

入力文: The while statement gets (受け取る) a character.

事例1: Most of the work gets (される) done in the body of the loop.

事例2: You'll get (得る) wrong answers.

例2 (重み係数)

入力文 : Certain declarations can be
made implicitly by context.
行なう

事例1 : The compilation will
make an executable file.
作る

事例2 : The ANSI standard has
made many small changes to basic
行なう types and expressions.

表4: 例2の入力文と事例との距離

語	事例1	事例2
直接目的語	6.0	7.0
主語	15.0	10.0
補語	0.0	0.0
間接目的語	0.0	0.0
前置詞	0.0	20.0
式(5)	36.0	40.0

例3 (正解の事例の方が距離が大きい)

入力文 : Each invocation gets a fresh set of
得る
all the automatic variables.

事例1 : Most of the work gets done in the body
される of the loop.

事例2 : You'll get wrong answers.
得る

事例3 : The while statement gets a character.
受け取る

事例4 : The compiler doesn't get the message.
理解する

表5: 例3の入力文と事例との距離

語	事例1	事例2	事例3	事例4
直接目的語	20.0	7.0	8.0	5.0
主語	6.0	14.0	12.0	13.0
補語	20.0	0.0	0.0	0.0
間接目的語	0.0	0.0	0.0	0.0
前置詞	20.0	20.0	20.0	20.0
式(5)	87.6	46.4	45.2	40.8

表4、表5の式(5)の値は、7.1節で述べた重み係数での値である。

例1については、不正解になったのは明らかである。例2については、7.1節で述べた重み係数では、式(5)の値は事例1の方が事例2より小さいが、前述のように重み係数を変えることで、式(5)の値は事例2の方が事例1より小さくなって、正解になることもある。例3については、事例2と事例4を比べれば分かるように、前述のように重み係数を変えても式(5)の値は、事例4の方が事例2より必ず小さくなり、正解になることはない。

8.3 動詞以外の多義性解消

表6: 動詞以外の多義性解消

検索範囲	D	S
D S C I P	42.4%	57.8%
D S C I	44.6%	61.3%
D S C P	45.1%	59.7%
D S I P	43.0%	58.9%
D C I P	42.9%	—
S C I P	—	64.7%
D S C	47.8%	63.8%
D S I	41.1%	61.2%
D S P	42.5%	57.5%
D C I	45.8%	—
D C P	44.3%	—
D I P	41.3%	—
S C I	—	61.0%
S C P	—	61.0%
S I P	—	60.8%
D S	42.7%	62.5%
D C	42.9%	—
D I	42.1%	—
D P	41.7%	—
S C	—	59.0%
S I	—	59.5%
S P	—	59.5%
D	41.1%	—
S	—	60.8%

動詞以外の多義性解消について考えてみると、表6のような正解率が得られる。これは、動詞の多義性が正しく解消された際に、動詞以外の各語に関して最適

解として選択された事例と入力文の概念番号が一致した場合を正解としている。

表6より、主語の正解率の方が直接目的語の正解率より高いことが分かる。これは、直接目的語の方が主語に比べて事例データに多くの種類の語が登録されているためであると考えられる。

また、検索範囲別に見てみると直接目的語に関しては直接目的語・主語・補語の組合せが最も正解率が高く、主語に関しては主語・補語・間接目的語・前置詞の組合せが最も正解率が高い。主語に関して2番目に正解率が高いのが直接目的語・主語・補語の組合せであることから、動詞以外の多義性解消についても直接目的語・主語・補語の組合せが最も良い結果が得られるということが分かる。

8.4 他研究との比較

SENA[1]では、2000個の事例をあらかじめ集めて実験を行なっていて、動詞をtakeに絞っているにもかかわらず、50個実験を行なった段階でも正解率が70%を越えていない。

MBT1[4]では、300個の翻訳例を生成し188個の未知の入力に関して実験を行なっている。正解率は、85.9%と高いが、この手法は翻訳例からシソーラスを自動合成しているので、シソーラスの不備による失敗が多い。同様に、既存のシソーラスを用いたMBT1b[4]では、我々の手法と同様事例不足が一番の失敗の原因になっている。ただ、MBT1bでもあらかじめかなりの翻訳例を生成している。

我々の手法は、事例の全くない状態から実験を行なって70%以上の正解率を得られている点から、分野がC言語のマニュアルに限定されているが、これらの手法より有効な手法であると考えられる。

9 おわりに

8章で述べたように、本手法は事例データに基づく手法の大きな課題となっている、非常に多くの事例を必要とするという問題点を解決するものではない。しかしながら、事例の全くない状態から実験を行なって70%以上の正解率が得られ、有効な手法であることが分かった。また、本手法に適当な検索範囲が検索範囲の語数3語以上でその中に直接目的語と主語が含まれていると良いということが分かった。

本手法の最大の問題点は、事例に入力文と同じ意味の同じ動詞が含まれていなければ正解になることはないということである。これを解決するためには、大量の事例の追加や新たなアルゴリズムの追加などが必要である。

それから、今回用いた距離の値や重み係数の値が最適な値であるとは限らない。今後これらの最適値を求めていく必要がある。それと同時に、式(5)の検討も必要である。また、事例が300文と少ないのでさらに実験を行なって事例の追加をしていく必要がある。

謝辞

本研究では、Cognitive Science Laboratory, Princeton Universityで開発されたフリーソフトWordNetを使用しました。ここに、Cognitive Science Laboratory, Princeton Universityに感謝致します。

参考文献

[1] 浦本 直彦:“制約と事例による優先度を組み合わせた英文の多義性の解消”, 情報処理学会研究報告, 92-NL-90, pp.65-72 (1992-7).

[2] 藤井 敦, 乾 健太郎, 徳永 健伸, 田中 穂積:“動詞多義性解消における格要素の貢献度について”, 情報処理学会研究報告, 96-NL-111, pp.55-62 (1996-1).

[3] 野村 浩郷編:“言語処理と機械翻訳”, 講談社 (1991).

[4] 佐藤 理史:“MBT1:実例に基づく訳語選択”, 人工知能学会誌 Vo.6, No.4, pp.592-600 (1991-7).

[5] 赤間 清:“帰納的学習システムLS/1による翻訳の学習”, 人工知能学会誌 Vol.2, No.3, pp.341-349 (1987-9).

[6] 古瀬 蔵, 隅田 英一郎, 飯田 仁:“経験的知識を活用する変換主導型機械翻訳”, 情報処理学会論文誌 Vol.35, No.3. pp.414-425 (1994-3).

[7] 小西 友七編:“ジーニアス英和辞典”, 大修館書店 (1989).

[8] Brian W.Kernighan and Dennis M.Ritchie: “The C programming language. Second Edition”, Prentice-Hall (1988).

[9] B.W. カーニハン, D.M. リッチー著, 石田 晴久訳:“プログラム言語C第2版”, 共立出版 (1989).