2003－NL－155 (7)
2003／5／26

日英特許公報を用いた対訳辞書および翻訳メモリの構築

高橋 博之[†] 川崎 立八[†] 牧田 光晴[††] 樋口 重人[††] 藤井 敦[†††,††††] 石川 徹也[†††]

† （株）クロスランゲージ
〒 169-0051 東京都新宿区西早稲田 2-20-9
†† （株）パトリス
〒 135-0043 東京都江東区塩浜 2-4-29
††† 筑波大学
〒 305-8550 つくば市春日 1-2
†††† 科学技術振興事業団 CREST

特許文には，専門用語，新語，独特の定型表現が多く現れるため，機械翻訳の精度を向上させるためには，対訳辞書と翻訳メモリの更新が不可欠である．しかし，これらの資源を人手で構築することは高価である．本研究は，同一内容に関して日米に出願された対応特許から，特許文に頻出する数字列の共起に着目して対訳文と対訳語の抽出を行い，機械翻訳用の辞書および翻訳メモリを自動構築するための手法を提案する．1995–1999 年に公開された約 3 万件の公開特許公報を対象に評価実験を行った結果，対訳文，対訳語，翻訳メモリを高精度で抽出できることが分かった．

Building Translation Dictionaries and Translation Memories Using Japanese-US Patent Corpora

Hiroyuki Takahashi[†], Tatsuya Kawasaki[†], Mitsuharu Makita[††], Shigeto Higuchi[††], Atsushi Fujii[†††,††††], Tetsuya Ishikawa[†††]

†Cross Language Inc.
2-20-9 Nishiwaseda, Shinjuku-ku, 169-0051, Japan
††PATOLIS Corporation.
2-4-29 Shiohama, Koto-ku, 135-0043, Japan
†††University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
††††CREST: Japan Science and Technology Corporation

To improve machine translation for patent documents including many technical terms, new words, and unique expressions, it is necessary to update dictionaries and translation memories. However, manual construction of these resources is expensive. We propose a method to build dictionaries and translation memories automatically from Japanese-US patent families filed for the same invention. We use occurrences of numerals across languages to align bilingual sentences, from which dictionaries and translation memories are extracted. We used approximately 30,000 Japanese-US patent pairs published in 1995–1999 and showed the effectiveness of our method by means of experiments.

1. はじめに

近年，世界的に知的財産に対する関心が高まっており，特許翻訳の需要が急増している．しかし，特許翻訳には通常の翻訳にはない専門の知識が必要であり，専門の翻訳者は不足しているのが現状である．

このような中，機械翻訳の技術が注目されており，わが国においては数社から特許専用の翻訳ソフトが市販されている．しかし，現状の機械翻訳の訳質は，調査用や翻訳者の下訳用として使う上でもまだ不十分である．特許文の機械翻訳において訳質を下げる主要な原因は，各種の専門分野の用語が頻出することと，特許特有の言い回しがあることである．前者は辞書の，後者は翻訳メモリの充実で補う必要があり，人手で収集するのはコストの面から容易ではない．

特許公報は電子化が進んでおり，計算機処理が可能である．特許は複数の国，複数の言語で同内容で出願されることが多いため，対訳コーパスとしての利用が可能であり，ここから対訳辞書や翻訳メモリを自動で構築できれば，低コストでの訳質の向上が期待できる．

本研究では同じ発明内容に基づいて出願された対応特許を利用し，ここから対訳文，対訳語を抽出し，その情報を元に機械翻訳のための対訳辞書と翻訳メモリを構築する方法を提案する．

以下，2 章で訳質向上のために必要な情報と処理について説明し，3 章で対訳辞書および翻訳メモリの構築手法の概要を述べ，4 章で対応特許からの対訳文および対訳語の抽出手法を示し，5 章で対訳文および対訳語からの対訳辞書および翻訳メモリの構築手法を示す．最後に 6 章で評価を行い本手法の有効性を示す．

2. 訳質向上のために必要な情報と処理

2.1 対訳辞書

機械翻訳では対訳辞書を用いて訳語を決定するため，辞書に登録されてない未知語は翻訳することができず，訳質の低下の原因となる．訳質改善のためには未知語の収集と訳付けの作業が必要であり，従来の人手による方法では未知語に対する網羅的な対応はコスト的に困難である．

この問題を解決するために，コーパスからの対訳語の自動抽出手法が提案されている[1)~4)]．しかし，対訳語の自動抽出の精度は 100% ではないため，抽出した対訳語から辞書を構築する際には人手による誤りの除去作業が欠かせない．経験上，誤りの多いデータのチェックでは一件あたりの確認時間が多くかかる傾向がみられる．また誤りが多いとチェックした件数あたりの採用件数も少なくなり非効率である．したがって，効率的な辞書構築のためには，精度が高い対訳語抽出手法が必要である．

また，辞書に訳語を追加することで訳質は向上するものの，特定の専門分野でしか使われない訳語を登録すると，逆に訳質が悪化することがある．例えば “core”＝「炉心」という訳語は，原子力関連の文献では適切である．しかし，他分野では不適切であり誤訳の原因となる．

このような問題に対して，商用の機械翻訳システムでは分野ごとの専門語辞書を用意し，翻訳対象文献ごとに辞書を指定させることで対処している．しかし，人手による訳語の分野特定は，文献調査などを必要とするため時間のかかる作業である．したがって，作業の効率化のためには，訳語を自動抽出するだけでなく，同時に適切な分野も自動抽出できることが好ましい．

2.2 翻訳メモリ

翻訳メモリは複数言語間の対訳文をデータベース化しておき，入力文に対するマッチングと置き換えで翻訳する機能である．翻訳メモリに基づくシステムは類似度による柔軟なマッチング機能を持っており，登録された文とまったく同一の文でなくても翻訳文を検索することが可能である．ただし，その場合の訳文は類似文の訳でしかないので，人手による後修正が必要である．このように翻訳メモリは翻訳者のための支援ツールとしての側面が大きい．

Nagao[5)] は翻訳メモリを機械翻訳に利用する手法として，例文ベースの翻訳手法を提案している．ここでは，翻訳対象の文と類似した例文との差分を出し，対象言語側の例文の該当部分を置き換えることで訳文を生成する．この処理に必要な節や句の対応は，それぞれの構文構造の類似より自動抽出する．しかし，この手法はまだ研究途上であり，カバー率が低く，訳質も従来と同程度にとどまっているため，翻訳ソフトとして実用化はされていない．

翻訳メモリを機械翻訳に利用するための手法として，あらかじめ対応する要素を変数として記述しておくという手法がある☆．これは，データの整

☆ （株）クロスランゲージ PC-Transer シリーズ
http://www.crosslanguage.co.jp/

(a)	本発明はこの特定の適用に限定されるものでは決してない. It should be understood that the invention is in no way limited to this particular application.
(b)	より詳細には，本発明は，< $1 > に関する. More particularly, the invention concerns < $1 >.

図 1　翻訳メモリの例

備に手間がかかるものの，例文にマッチした場合は高精度の翻訳を行えるため，規則に基づく翻訳の補助機能としてすでに実用化されている．

翻訳メモリの例を図 1 に示す．(a) は対訳文をそのまま登録したもので，(b) は要素を変数化したものである．変数は < $$N$ > の形式 (N は整数) で示され，日英で対応する部分には同じ変数を用いる．

(a) は例文と同じか，あるいは句読法の違いなど微細に異なる文の翻訳にしか利用できない．(b) では変数化によって「より詳細には，本発明は，機械翻訳に関する．」「より詳細には，本発明は，マイクロアクチュエータに関する．」などより広い範囲の文の翻訳を行うことができる☆．

対訳辞書の場合と同様，翻訳メモリの構築に際してはコーパスからの対訳文の自動抽出手法が利用できる．この場合も効率的な構築のためには，高精度の対訳文抽出が必要である．

また，機械翻訳の補助としての翻訳メモリの構築を考えた場合，変数化が行われた応用の広い翻訳メモリの構築が好ましい．変数化には，文中の対応する要素を特定する必要があり，対訳語抽出の情報が利用可能である．

3. 対訳辞書・翻訳メモリ構築手法の概要

対訳辞書および翻訳メモリ構築手法の概要を図 2 に示す．同一の発明に基づく日米の対応特許を対訳コーパスとして用い，特許に頻出する数表現の共起に基づいて対訳文対を抽出する．次に対訳文対より数表現の共起位置を用いて対訳語対を抽出する．

対訳語対を用いて対訳辞書を構築する．ここで，抽出元特許の IPC(国際特許分類) による分類コードを専門分野分類のため用いる．

対訳文対を用いて翻訳メモリを構築する．ここで，対訳語対の抽出処理を利用して変数化を行う．また，定型句の抽出を行って汎用性の高い翻訳メモリを構築する．

それぞれの処理については以下の章で具体的に説明する．

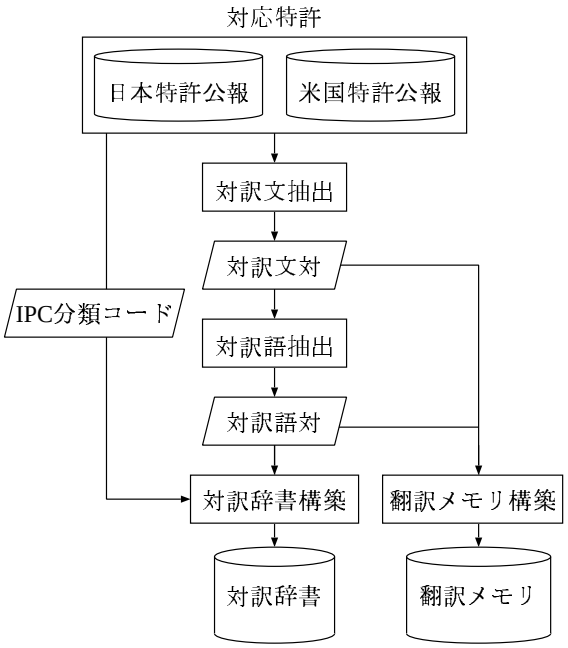

図 2　対訳辞書・翻訳メモリ構築手法の概要

4. 対訳文および対訳語の抽出

4.1 対訳文の抽出

対訳文の自動抽出法としては対訳辞書と統計を用いた手法[6] が提案されているものの，実用的な精度を達成しているとはいえない．

梶ら[3] は図の参照番号の利用可能性を示唆している．特許文では図を用いて発明の詳細を説明することが多いが，その場合，図の各要素を参照するために参照番号を用いる．対応特許では双方で同じ図を用いることが多いため，参照番号は同一となり対応する文で共起する．この手法は特許文の一部にしか利用できないものの，高精度の抽出が可能である．

梶らは対訳語の抽出手法として参照番号の利用

☆ 変数にマッチした部分は，別途，機械翻訳で訳語を生成し，その訳語で変数を置換して訳文を生成する．

表 1　特許文中の数量表現の例

分類	英語	日本語
図中の参照番号	Next, layered structure 12 is placed in ...	次に，層状構造１２を，...
図などの番号	FIG. 2/ claim 1	図２/請求項１
物理量	SiGe with 76.6% Ge has ...	Ｇｅ７６．６％を含有するＳｉＧｅの...
型番	...which are marketed by Hecon Corporation of Germany under Model No. RI41-0/3600 AR.11KB ...	...ドイツのヘーコン（Ｈｅｃｏｎ）社がモデル番号ＲＩ４１－０／３６００ＡＲ．１１ＫＢとして販売している...

数字列	文
[16, 20, 20]	The composition of layer 20 is chosen so that layer 20 has a second rate of oxidation less than the rate of oxidation of layer 16 and has desired electrical properties.
[16, 20, 20]	層２０の組成は，層２０が第１の層１６の酸化速度より遅い第２の酸化速度を有し，所期の電気特性を有するように選択する.

図 3　数字列抽出の例

を示唆したが，手法の検証は行っていない．本研究ではこの参照番号を対訳語抽出だけでなく，対訳文抽出にも利用した手法の検証を行った．

対訳文抽出への応用を考えた場合，表 1 に示すように，図中の参照番号だけでなく，図・表・請求項の番号，工業製品の型番，物理量をあらわす数量表現なども対応する文で共起するため，これらの数字も利用可能である．

そこで，数量表現一般の共起を利用した以下の対訳文抽出手法を考案した．

1. 日英の対応特許の内容を文に分割する．
2. 日英の各文から数字列を抽出する．
3. 対象外の文を除去する．
4. 日英で同じ数字列を持つ対を抽出する．

1. の文分割は句点やピリオドなどの区切り記号で行う．ただし英語の場合 “FIG.” というような略語のピリオドがあるので，略語一覧表を別に用意することで略語のピリオドでの文分割を回避する．

2. では抽出する数字はローマ数字の並びとし，漢数字は無視した．漢数字は「第一の」＝“first”，「...の一つ」=”one of ... “ のように英語では数字として現れない表現に使われることが多いためである．さらに,「第１の」「...の１つ」というようにローマ数字を使って同様の表現がされる場合がある．そこで，以下の例外規則を設ける．

- 日本語で一桁の数字のあとに「つ」がある場合は抽出しない
- 日本語で一桁の数字の前に「第」がある場合は抽出しない

抽出した数字の列は昇順に並べ替えておく．すなわち，日英で比較する際に数字列の出現順序は無視する．これは日英で訳語の出現順序が入れ替わることがしばしばあるためである．

数字列抽出例を図 3 に示す．ここでは抽出された数字列を [16, 20, 20] というように表記する．日本語文での「第１」「第２」の数字は前述の条件により抽出されない．

3. では抽出対象外となる文を除去する．まず，数字がひとつも出現しない場合は本手法は適用できないので除去する．また数字が一箇所しかない場合も文対応の誤りの可能性が高いので除去する．さらに，日米それぞれの公報内で，数字列がユニークでない文を除去する．例えば [1, 8] を含む英文が “Claims 1 and 8 are ...” と “FIGS. 1 and 8 are ...” の二文あったとすると，日本語文で [1, 8] を含む文があったとしてもどちらを対応付けたらよいか判別不能なので，両方の文を除去する．

最後に 4. では日英で同じ数字列を含む文を対にして抽出する．3. で同じ数字列を持つ文を除去してあるため，対応付けの候補は高々ひとつであり，曖昧性は生じない．

4.2　対訳語の抽出

文中の数字列は対訳語の抽出にも利用可能である．

表 1 に挙げた数字のうち，図中の参照番号は多くの場合該当する要素の名称（名詞）の後に置か

れる．そこで，要素参照の数字に着目した以下の対訳語抽出手法を考案した．

1. 文中の各数字についてその前方の語（名詞）を抽出する．
2. 対象外の語を除去する．
3. 対訳語を抽出する．

1. では数字の位置を末端とする名詞があると仮定し，その名詞の先頭位置を検索する．ここで対象とする数字は 4.1 の対訳文抽出手法の 2. の条件で選択する．

名詞の抽出はそれぞれ以下のように行う．

表 2　不要語の例

品詞	例
冠詞	a, an, the
前置詞	about, after, as, at, before, by, for, from, in, of, on, to, with
疑問詞	how, what, where, which, who, why
接続詞	and, because, but, if, or, so, than, when, whether, while
be 動詞	am, are, be, was, were
代名詞	I, you, he, she, we, they, it
助動詞	can, do, may, shall, will
形容詞	all, any, less, more, some, such
副詞	ago, almost, also, ever, not

英語名詞の抽出 英語の場合，単語間にスペースをあけるので，単語の検出は容易であるが，専門用語は複数の単語の並びからなる複合名詞であることが多いので，単に一単語を抽出するだけでは正確な抽出はできない．そこで，名詞句に含まれないと思われる単語のリスト（不要語リスト）を用意して，それらを含まない最大の単語列を抽出する．使用した不要語は 143 語である．例を表 2 に示す．形容詞や副詞は名詞の一部になりにくいと思われるもののみを不要語とし，それ以外の冠詞，前置詞，疑問詞，接続詞，be 動詞，代名詞，助動詞は一般に使用されるものを全て不要語とした．

日本語名詞の抽出 単語の境界が明確でない日本語の場合は形態素解析による単語抽出が一般的である．ここでは形態素解析システム茶筌☆を用い，福井ら[2] の手法に基づく以下の規則で検出を行った☆☆．

- 名詞，未知語，接頭詞，自立動詞（体言接続特殊活用）の連続を検出する．ただし，非自立名詞，代名詞，数詞は含まない．
- 特許にしばしば見られる接頭語（概，本，各，前記）を削除する．

2. では 4.1 の対訳文抽出手法の 3. と同様に，一文中で同じ数字が複数出現してそれぞれで別の名詞が抽出された場合，その名詞は除去する．例えば「酸化物１２」と「請求項１２」が同一文中に出現した場合，数字１２に対応する名詞が二つあり一意に対応付けができないので両者を除去する．

3. では数字の対応を使用して対訳対を抽出する．対訳対は同内容のものをまとめて出現頻度を集計する．このとき英語では文頭の大文字化や複数形の可能性があるので，同内容でも表記が完全には一致しないことがある．そこで，同じ日本語名詞から抽出された英語名詞を相互に比較し，大文字，小文字だけの違いの場合および複数語尾だけの違いの場合は，それぞれ小文字表記，単数形語尾として抽出する．

表 1 の例に示すように，数字は参照番号とは限らず，他の使われ方の場合には対応する名詞が数字の前にあるとは限らない．このようなケースのうち数字の直前に名詞がない場合は，名詞の検出自体が失敗するので誤った対訳語が検出されることはない．一方，日英で対応関係のない名詞が数字の直前に出現した場合には誤検出の問題があるが，このようなケースはまれであり，出現頻度が多くないので出現頻度制約で排除できる．

5. 対訳辞書および翻訳メモリの構築

5.1 対訳辞書の構築

抽出した対訳語を人手でチェックして誤抽出を除去し，正しい訳語対に品詞や活用情報などを付与して対訳辞書に登録する．

前述のように対訳語を辞書に登録する際には適切な専門分野の辞書を選択する必要がある．特許には IPC (国際特許分類) による分類コードが付与されているため，対訳語の抽出元の公報の IPC を参照することで分野の選択が可能である．

5.2 翻訳メモリの構築

抽出された対訳文は翻訳メモリ構築に利用する．例えば，図 4(a) の対訳文が抽出された場合，翻訳者が利用する翻訳メモリとしてならばこのままで利用可能である．これと完全に一致する文が別の公報で出現することはまずありえないが，類似検索を行うことで類似文の翻訳を効率的に行うこと

☆ http://chasen.aist-nara.ac.jp/

☆☆ 数字によって末尾位置が確定しているので，手法は簡略化されている．

ができる．

しかし，機械翻訳のための翻訳メモリとして使う場合には類似度の閾値を高めに設定する必要があり，このままでは実際の翻訳で使われる可能性は低い．そこで，図 4(b) のように対訳語抽出された部分を変数化することで，変数部分が他の名詞に置き換わった類似文の翻訳にも利用できるようになるが，これでも固定部分が多く，適用可能な類似文は少ない．

ここで，区切り記号である読点・カンマに着目し，区切り記号の前方のみを切り出すと，図 4(c) のような句レベルでの対訳対が得られる．これは特許公報でしばしば見られる表現であり，汎用性の高い翻訳メモリである．

そこで，本研究では翻訳メモリ構築の第一段階として，文頭の定型句に関する翻訳メモリの構築を行うこととし，以下の手法を考案した．

1. 対訳文対から対応する単語を変数化する．
2. 読点あるいはカンマより前方を抽出する．
3. 統計的手法により類似度を求め，一定値以上の対を抽出する．

1. では 4.1 の手法で抽出した対訳文から 4.2 の手法で対訳語を抽出し，対訳語および抽出に使用した数字の部分を変数で置き換える．ただし，英文で抽出名詞の直前が冠詞であった場合はそれも置き換え範囲に含める．図 4 の例では，この操作で (a) から (b) へ変換されるが，ここでは抽出された名詞 “images” の後の参照番号 “24” と前の冠詞 “an” を含む部分が変数化されている．変数は文頭から英文での出現順に $< \$1 >, < \$2 >, \ldots$ と割り当てる．

2. では日本語では読点，英語ではカンマを区切り記号として用い，文頭の句の対を抽出する．どちらか一方あるいは両方で区切り記号がなかった場合は対象外とする．さらに，翻訳メモリとして使うという性質上，変数を含まない句は抽出しない．抽出例を表 3 に示す．

文頭定型句の抽出は単語抽出の場合と異なり，誤検出の可能性が高い．定型句の後に区切り記号があるとは限らず，また定型句ではない一般の句や節を抽出する可能性があるためである．そこで，3. では統計的手法によって誤りの排除を行う．ここでは，対訳抽出でしばしば用いられる Dice 係数を用いて類似度を求め，一定の閾値以上を採用する．

Dice 係数を以下に示す．

$$Dice(J,E) = \frac{2f_{JE}}{f_J + f_E} \qquad (1)$$

ここで J と E はそれぞれ日本語と英語の抽出句であり，f_J と f_E はそれぞれの単体での出現頻度，f_{JE} は両者の共起頻度である．

6. 評　　価

本手法の評価実験には，福井ら[2] の提案した手法で抽出した対応特許を用いた．この手法は特許優先権主張を伴う出願制度を利用したもので，今回使用した対応特許は 1995–1999 の 5 年間に公開された日米公報より抽出した 31,045 件である．

6.1 対訳文抽出の評価

対訳文抽出の評価結果を表 4 に示す．精度と再現率は 4 件の公報 (212 文抽出) をサンプルとして評価した．再現率は高くないが，高精度での対訳文抽出に成功した．

表 4　対訳文抽出の評価結果

対訳文数	精度	再現率
1,144,676	98%	20%

本評価で発見された誤りはいずれも文分割にかかわるもので，文分割が正しければ精度は 100% であった．図 5 に誤りの例を示す．日本文の「有線電話，セルラ電話，ファクシミリ装置，パーソナル・コンピュータおよびポケットベル」に相当する部分が英文にはないので，正しい訳文対ではない．これは英文側を “:” で区切ってしまったためで，原文では “:” 以降に “wired telephone, cellular telephone, facsimile machine, personal computer and paging device.” というように文が続いている．

The method of claim 1 or 8 wherein at least one of the communications devices is a communications device selected from the group:
請求項 1 または請求項 8 に記載の方法において，該通信装置の少なくとも 1 つは，有線電話，セルラ電話，ファクシミリ装置，パーソナル・コンピュータおよびポケットベルのグループから選択される通信装置であることを特徴とする方法．

図 5　対訳文抽出の失敗例

6.2 対訳語抽出の評価

対訳語抽出の評価結果を表 5 に示す．出現頻度による制約を段階的に変化させて，抽出語数と精

(a)	Referring to FIG. 2, an image 24 is inputted into the system 10 and displayed on the display 20. 図２を参照すると，画像２４はシステム１０に入力され，表示器２０に表示される．
(b)	Referring to < $1 >, < $2 > is inputted into < $3 > and displayed on < $4 >. < $1 > を参照すると，< $2 > は < $3 > に入力され，< $4 > に表示される．
(c)	Referring to < $1 > < $1 > を参照すると

図 4　翻訳メモリの構築例

表 3　文頭定型句抽出例

出現頻度	英語	日本語
3494	< $1 > を参照すると	Referring to < $1 >
3033	< $1 > に示すように	As shown in < $1 >
1236	< $1 > に示されるように	As shown in < $1 >
658	< $1 > を参照すると	Referring now to < $1 >
576	< $1 > に示されているように	As shown in < $1 >
550	< $1 > を参照すると	With reference to < $1 >
533	< $1 > に示すように	As illustrated in < $1 >
371	次に < $1 > を参照すると	Referring now to < $1 >
366	< $1 > に示したように	As shown in < $1 >
340	< $1 > を参照して	Referring to < $1 >
309	< $1 > を参照すれば	Referring to < $1 >
255	< $1 > に示すように	Referring to < $1 >

度の変化を調べた．新語数は，比較用の辞書として (株) クロスランゲージの機械翻訳用辞書（基本語辞書＋専門語辞書 22 分野）を用い，日英辞書で日本語見出しが未登録のものを新語とした．精度はそれぞれの訳語対から 100 サンプルをランダムに抽出して行った．

出現頻度制約 5 以上までは安定した精度を保っているが，それ以下では急に精度が悪化した．これより，低頻度の単語対に誤りが集中していること，頻度制約による誤りの排除手法が有効であることが明らかになった．

表 5　対訳語抽出の評価結果

出現頻度	対訳語数	新語数 (日英)	精度
100 以上	1,620	318	90%
10 以上	26,421	15,888	86%
5 以上	62,589	42,344	86%
2 以上	192,360	137,434	75%
1 以上	500,260	351,238	57%

6.3 翻訳メモリ抽出の評価

抽出した対訳文 1,144,676 対を用いて文頭定型表現に関する翻訳メモリの抽出精度を評価した．

表 6　翻訳メモリの抽出精度

Dice 係数	件数	精度
0.1 以上	91	97%
0.03 以上	441	87%
0.01 以上	1,943	64%

まず，自動抽出の精度を調べた．表 6 に結果を示す．ここでは定型表現を抽出するために，少なくとも片方が頻度 100 以上の対に限定し，Dice 係数の閾値を段階的に変化させ，それぞれ 100 件のサンプル調査 (Dice 係数 0.1 以上の場合は全数調査) で抽出精度の変化を調べた．Dice 係数の制約を厳しくすることで精度が向上しており，Dice 係数の効果が明らかとなった．

次に，今回抽出した翻訳メモリを翻訳に使用した場合の訳質の評価を行った．Dice 係数で上位 100

表 7 翻訳メモリの訳質評価

分類	割合	例 (「原文」→ “翻訳メモリ”/” 機械翻訳”)
(a) 翻訳メモリの方がはるかによい	29%	「$に戻ると」 → “Returning to $”/”When $ is returned to”
(b) 翻訳メモリの方がよい	35%	「まず$を参照すると」 →”Reffering first to $”/”At first when $ is refferd to”
(c) 同等の訳質	11%	「$に関しては」 →”with respct to $”/”As for $”
(d) 同訳/近似訳	25%	「$ に示したように」 →”As shown in $“/”As indicated in $“

件の翻訳メモリ（抽出ミスは除去した）について同じ句を機械翻訳した場合との訳質比較をおこなった☆．今回は日英翻訳で評価した．表 7 に結果を示す．

ここで，(a) は機械翻訳では誤訳となった場合，(b) はどちらでも意味は取れるが翻訳メモリの方が自然な場合，(c) はほぼ同等の訳の場合，(d) は全く同じ訳か訳語の微妙な違いだけの場合である．(a) と (b) をあわせて 64 % の例で訳質が改善しており，本手法で抽出した翻訳メモリが訳質の改善に有効であることが示された．

次に，今回抽出した翻訳メモリが実際の公報の翻訳においてどの程度使用されるかを調べた．今回のデータ抽出に使用したものとは別の日本公報公報 12 件，2191 文を用いて翻訳メモリにマッチする率を調べた．翻訳メモリは訳質評価と同じ 100 件を用いた．表 8 に結果を示す．これより，原文の 3.9 % が翻訳メモリとマッチし，表 7 の評価よりそのうちの 64% で訳質が改善することから，全体で 2.5 % の文で訳質が改善されると考えられる．これは効果的な訳質改善手法が乏しい機械翻訳においては十分効果的といえる．

表 8 翻訳メモリの評価 (マッチ率)

文数	マッチ数	マッチ率
2191	86	3.9%

7. おわりに

特許文中の数字列の共起に着目した対訳文・対訳語抽出手法およびそれらの情報を用いた対訳辞書・翻訳メモリの構築手法を提案した．

本研究の文対応の手法は精度が極めて高いという特長がある一方で，再現率は低く特許コーパスの情報を十分に利用しているとはいえない．今後は対応文の位置をキーにしてその間にある文の対応を取る手法の研究を行う予定である．

また，今回提案した数表現の対応による対訳文，対訳語の抽出手法は対応特許の抽出にも利用できる可能性がある．この手法の検討および評価も今後の重要な課題である．

☆ ここでは比較用の機械翻訳システムとして (株) クロスランゲージの PAT-Transer V5 を用いた．

参 考 文 献

1) Smadja, F., Hatzivassiloglou, V. and McKeown, K.R.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, Vol. 22, No. 1 (1996).
2) 福井雅敏, 樋口重人, 藤井敦, 石川徹也: 日米対応特許コーパスを用いた対訳抽出手法, 情報処理学会自然言語処理研究会 145-4 (2001).
3) 梶博行, 相薗敏子: 共起語集合の類似度に基づく対訳コーパスからの対訳語抽出, 情報処理学会論文誌, Vol.42, No.9, pp.2248–2258 (2001).
4) 北村美穂子, 松本裕治: 対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌, Vol. 38, No. 4, pp. 727–736 (1997).
5) Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, pp. 173–180 (1984).
6) 春野雅彦: 辞書と統計を用いた対訳アライメント, 情報処理学会論文誌, Vol. 38, No. 4, pp. 719–726 (1997).