自然言語処理 115－8
（1996．5．29）

対訳テキストコーパスからの対訳語情報の自動抽出

高尾 哲康
山形県立米沢女子短期大学 社会情報学科

富士 秀　松井 くにお
(株) 富士通研究所 マルチメディアシステム研究所

概要

近年、機械翻訳システムの実用化が進んできているが、ユーザが翻訳したい文書を高品質に翻訳するためには、基本語辞書のみでは不十分であり、専門用語やユーザ用語等の対訳語情報をユーザサイドで容易に抽出・登録できることが重要である。

本論文では、統計情報と辞書情報を利用して、日英対訳テキストコーパスから、特に複合語や未登録語を中心とした対訳語情報の抽出を効率的に行なうための自動抽出手法を考案し、実際にプロトタイプを作成して評価を行ない、その結果について報告する。

Automatic Extraction of Translation Word Information from Parallel Text Corpus

Tetsuyasu Takao
Yonezawa Women's College of Yamagata Pref.
Faculty of Social and Information Studies

Masaru Fuji　Kunio Matsui
Fujitsu Laboratries Ltd.
Multi Media Systems Lab.

Abstract

Recently, machine translation has became increasingly useful in various area. It is important to extract and register translation word information of technical term and user specified term easily on user side.

In this paper, we propose a technique of automatic extraction of translation word information from Japanese and English text corpus. We use statistical and lingistic method which extracts compound word and unregisterd word. We evaluate the results of this method.

1 はじめに

機械翻訳システムにおいては、翻訳品質の向上のためには、辞書情報の充実が必要である。とりわけ、機械翻訳システムをブラウジングツールとして利用することを考えた場合、専門用語やユーザ用語等の対訳語情報が辞書情報として整備されているのといないのとでは、翻訳品質に大きく影響し、その結果、読み易さの度合がかなり違ってくる。従来の機械翻訳システムにもオプションとして専門用語辞書が提供されているが、お仕着せ的なものにとどまっており、ユーザサイドで辞書登録などを行なう作業は手間がかかりすぎる面がある。また、翻訳して読みたい文書の分野がユーザごとに非常に多岐にわたるため、機械翻訳システムを提供する側が個々のユーザごとに対処することは困難である。

本研究では、日英対訳テキストコーパスから、まずテキストのアラインメントを行なって対訳テキストの対応をとり、その後に特に複合語や未登録語を中心とした対訳語情報の抽出を効率的に行なうための自動抽出手法を考案し、実際にプロトタイプを作成して評価を行なった。

順序関係が保存された対訳テキストの文対応推定を行なうテキストアラインメント技術についてはこれまでに以下の方法などが試みられている。

1. Gale ら [1][2] による文字数に基づいて対応関係を推定する方法
2. Brown ら [3] による単語数に基づいて対応関係を推定する方法
3. Kay ら [4] によるテキスト中の単語の出現回数とその位置 (どの文に出現したか) から対応している単語を推定し、それらの単語リストから対応づけされる文を推定する方法
4. 春野ら [5] による、辞書情報と単語の位置関係から対応関係を推定する方法

また、アラインメント済みの対訳テキストコーパスから対訳語候補の情報を自動抽出する試みがいくつかなされている。山本ら [8] は、名詞連続を専門用語として抽出して、機械翻訳辞書を参照することで言語間の対応関係を推定している。ただし、抽出する名詞は辞書登録されているものに限定されるので、そこから生成可能な候補以外は抽出できないこと、文の対応を考慮していないので、精度が低くなることが欠点である。また、熊野ら [9] は対訳語候補どうしの対応を推定する際には 1 文対 1 文の対応のみを考慮しており、1 文対 2 文などの場合は正しい訳語を推定できない。そのため、対訳テキスト間の対応の正確さが対訳語抽出能力に影響を及ぼすと指摘している。

本研究では対訳テキストコーパスとして UNIX コマンドのオンラインマニュアルを利用した。その特徴として、対訳の順序関係はパラグラフレベルを含めてほとんど保存されており (意訳など、前後の文を入れ換えた訳ではない)、対訳文の対応も、1 文対 1 文が多く、一部が英語 1 文対日本語 2 文であった。英語 1 文対日本語 3 文はごくわずかであり、その他のパターンはなかった。このため、自動抽出手法としては、形態素解析結果と辞書引き結果から得られる言語情報に基づく確信度と、対訳語候補の出現位置と出現頻度から得られる統計情報に基づく確信度の 2 つを利用した。

本実験により、ユーザは、翻訳したい分野の対訳テキストコーパスさえ用意できれば、対訳語情報を機械的に得ることができ、ユーザサイドにおける専門用語辞書やユーザ辞書の作成が容易にできるようになった。

2 自動抽出アルゴリズム

2.1 システムの目的

このシステムの目的は、対訳テキストコーパスを教師データとし、機械翻訳を行なったときに教師データになるべく近い訳文を生成するために必要と思われる対訳語情報を抽出することにある。ユーザはこの情報をもとに専門用語辞書やユーザ辞書を作成することになる。このシステムは機械翻訳システムのユーザが前処理段階に利用するツールとして位置づけられる。技術的には、日英対訳テキストコーパスにおいて、単語レベルのアラインメントを行なうシステムである。

2.2 システムの処理手順

1. フォーマット任意の対訳テキストコーパスにおいて、正規表現によるパターンマッチングによりパラグラフ境界と文境界を認識し、パラグラフ単位、および文レベルのアラインメントを行なう。

2. アラインメントされた対訳テキストコーパスをそれぞれ形態素解析、および対訳語辞書引きを行なう。
3. 形態素解析結果と対訳語辞書引き結果をもとに、その品詞列などから特に複合語や未登録語を中心とする対訳語候補を抽出する。
4. それぞれの言語において抽出された対訳語候補どうしで、単語・複合語・句レベルのアラインメントを行ない、品詞情報、意味情報を推定し、確信度の高い訳語対から順に出力する。

単語・複合語・句レベルのアラインメントは以下の観点から行なう。

- 対訳語候補群どうしで単語(複合語を含む自立語)の対応付けを行なう。対応付け範囲は文レベルでアラインメントされた範囲内とする。
- 対訳語表記で対応がとれなかったり、複数候補が出現した場合は、品詞情報などの文法情報、前後の対訳語ペア情報などを利用し、対訳語候補として出力する。
- 未登録語は、対訳語の対応がとれていないものとして扱う。
- 対訳語の対応はとれているが、単語辞書では頻度が低いなどで、翻訳しても訳語として選ばれない可能性があるものを検出した場合はその旨の情報を出力する。

2.3 テキストアラインメント

順序関係が保存された対訳テキストの文対応の推定を行なう。推定には統計情報や言語情報を利用する手法が考えられるが、今回は対象がUNIXマニュアルであり、日本語テキストに英語文が混じることが多く、未登録語が比較的多いと予想されるため、統計情報のみを推定に利用することにした。日本語テキスト内で英単語、算用数字がしばしば出現するが、その部分は半角文字に統一した。これは、アラインメント処理の際にテキストのバイト数を統計情報として利用することを考えてのことである。現在のところ以下の特徴がある。

1. 対応組のパターンを以下のものに限定している。
   0文対1文、1文対0文、1文対1文、1文対2文、1文対3文、2文対1文、2文対2文、2文対3文、3文対1文、3文対2文の対応組のみを考慮している。
2. 各対応組の対応確信度(距離)を、その対応が起こり得る可能性に基づいて評価する。この評価値は、まず、その対応組のテキストのバイト数の比を計算し、あらかじめコーパス全体から算出しておいた平均バイト数比の正規分布になるとみなした場合の平均値からのずれを計算することで求める。
3. ダイナミック・プログラミングの手法を用いて、全体として最も良い対応組の列を決定する。
4. 処理は2パスで行ない、最初のパスでは、パラグラフの対応をとり、第2のパスではパラグラフ内で文の対応づけを行なう。なお、これは起動オプションでパラグラフ対応アラインメントを行なわないようにすることも可能である。

UNIXマニュアルではパラグラフ間の対応がほとんどとれているため、実際のアラインメントはパラグラフ内の文の対応関係を推定するだけになるので、計算機処理量は比較的少なくなった。

2.4 対訳語候補の抽出

アラインメント済みの日英対訳データを入力とし、それぞれについて、形態素解析および辞書引きを行ない、その結果から得られる統計情報および言語情報から対訳語候補を推定する。形態素解析後、英語表記の語尾(変化部分)の補い処理を行なっている。対訳語候補は、形態素解析結果の品詞列情報から抽出する。抽出する対訳語候補は、英語文の場合は、下記の通りである。

- 名詞
- 名詞連続
- 形容詞+名詞(例： remote system)
- 動詞(-ed/-ing)+名詞(例： named host)
- 未登録語を含む名詞連続

日本語文の場合は、下記の通りである。

- 名詞

- 名詞連続
- 名詞＋接辞(「化」など)
- 名詞＋接辞＋名詞
- 動詞連用＋名詞(例：「切り替え文字」)
- 動詞＋｛する、した、している｝＋名詞
- 未登録語(名詞、サ変名詞)を含む名詞連続

なお、田中[7]によれば、専門用語の場合の品詞列の構成は、形容詞と名詞の組合せがほとんどであり、前置詞を含むものはごくわずかであると報告している。

それぞれの対訳語候補は、構成語(複合語の場合)、出現頻度(出現回数)、出現位置(アラインメント付けられた文のうち、何番目の文に出現したか)の情報をもつ。その後、抽出した対訳語候補群について、それぞれ候補どうしの対応の良さの評価を行なう。評価は以下のようにして行なう。

ある英語訳語候補 EW_i に対応する日本語訳語候補 JW_j を下記で示す確信度情報を計算することで評価する。i は、第 i 番目の英語対訳語候補を、j は、第 j 番目の日本語対訳語候補を意味する。

$$H(EW_i, JW_j) = f(HS(EW_i, JW_j), HL(EW_i, JW_j))$$

$HS(EW_i, JW_j)$ は、統計情報に基づく確信度であり、コーパスにおける出現頻度情報を利用し、次の式で計算する。

$$HS(EW_i, JW_j) = \frac{2freq(EW_i, JW_j)}{freq(EW_i) + freq(JW_j)}$$

$freq(EW_i, JW_j)$ は、EW_i, JW_j がアラインメントづけられた文中にそれぞれの訳語候補が出現する回数である。$freq(EW_i), freq(JW_j)$ は、それぞれの訳語候補が各言語のコーパス全体に出現する回数である。

専門用語やユーザ用語などの場合は、文章の流れに応じて訳し分けをすることはあまりない。そのため、アラインメントが正しい限り、対応する訳語どうしは対応している文中に出現する確率は非常に高くなるため、この確信度はかなり高くなることが予想される。一方、アラインメントが誤っている場合は、この確信度は低くなるので、対訳語候補の対応の信頼性が低くなる。

$HL(EW_i, JW_j)$ は、言語情報に基づく確信度であり、対訳語どうしの表層的類似度である。類似度は次の観点から考える。

1. 対訳語候補 EW_i, JW_j は、それぞれの構成語数が近いほど対応が確からしい。ただし、中点(「・」)、ハイフン(「−」)などの構成語どうしを結合するための語は構成語とはみなさない。
2. 対訳語候補 EW_i, JW_j のそれぞれの構成語間に対訳関係が多いほど対応が確からしい。
3. 対訳語候補 EW_i, JW_j のそれぞれの構成語間に対訳の順序関係があるほど対応が確からしい。ただし、英語対訳語候補に前置詞が含まれてなく、かつ日本語対訳語候補に助詞が含まれていない場合である。

$HL(EW_i, JW_j)$ は、上記の観点から計算される2つの確信度の加重平均として計算する。重み値の最適化については、実験結果から得ることにし、今回の実験ではそれぞれ 0.5 にした。

なお、未登録語のみの構成語から成る対訳語候補については、言語情報としての対訳関係は全く得られない(英単語がそのまま日本語テキストに書かれている場合を除く)が、統計情報に基づく確信度により対訳候補が抽出できる可能性がある。未登録語は、カタカナ語が比較的多く、英単語との対応が推定しやすいと考えられるため、今後はこのような情報を利用することも抽出精度向上のために必要であろう。

2.5 対訳語情報の抽出能力の評価方法

抽出された対訳語候補の適合率(抽出した対訳語候補のうち、正しいものの割合)と再現率(抽出するべき対訳語候補のうち、正しく抽出したものの割合)を計算することにより、対訳語抽出能力の評価を行なう。

3 実験結果

3.1 文対応アラインメント結果

本実験の対象とした対訳テキストコーパスは、UNIX コマンドのオンラインマニュアルのうち、

3つのコマンド(telnet, tip, ftp)を利用した。これらのテキストについて、アラインメントした結果を、表1に示す。パラグラフレベルではほとんど対応がとれていること、英語1文に対して日本語2文または3文など、日本語訳では、文を短くしていることが特徴である。文対応のアラインメント処理については、ほとんど正解であった。抽出された対訳語候補(主に名詞連続、未登録語)のうち、確信度が0.82以上の115候補はほとんど正解である。なお、tipコマンドの対訳データから抽出された対訳候補結果の例を表3に示す。

3.2 対訳語抽出結果

アラインメント済みのテキストに対して、形態素解析、辞書引きを行ない、対訳語抽出を行なった。その結果を、表2に示す。なお、異なり形態素数には、異なり未登録語数も含まれている。

3.3 対訳語候補の品詞・意味カテゴリ推定

一般に複合語には、並列的な複合語(「入出力」など)とそれ以外の複合語がある。並列的な複合語は基本的に名詞、またはサ変名詞になる。それ以外の複合語では、その中心要素となる語を決定し、この語の品詞、意味カテゴリを複合語全体の品詞、意味カテゴリとすることにした。

1. 対訳語候補が1語で構成される場合は、辞書にある品詞、意味カテゴリを優先する。

   名詞 - 名詞　画面　`screen`　1 40

   この例では、左から順に日本語品詞、英語品詞、日本語表記、英語表記、英語名詞複数形規則番号、意味カテゴリ番号である。

2. 対訳語候補が複数語で構成される場合は、末尾の語の品詞、意味カテゴリ、格属性(用言の場合)を優先候補とする。日本語の場合で、末尾の語が接辞(「化」、「的」等)の場合は、その属性によって、サ変名詞、形容詞・形容動詞等にし、意味カテゴリは接辞の直前の語を候補とする。前置詞や助詞を含む対訳語候補については、前置詞の直前の自立語等が考えられるが、現在対訳語候補として抽出できていないため、後回しにする。

   名詞 - 名詞　データ転送統計情報
   `data transfer statistics`　1 05

3. 日英の対訳語候補いずれも未登録語のみで構成される場合は名詞とし、意味カテゴリについては、人手にて補うことにする。

   名詞 - 名詞 ＡＣＵタイプ `ACU type`　1 62

3.4 適合率・再現率

確信度による閾値による適合率・再現率の変化を図1、および適合率・再現率の相関関係を図2に示した。適合率・再現率の相関関係の図から、座標(1.0, 1.0)に最も近いものが最適であると判断した結果、確信度の閾値を0.65前後とすると、最も効率的に対訳候補が抽出できると考えられる(適合率88 %、再現率53 %)。この場合、抽出候補数249のうち、218候補が正解となった。

3.5 抽出失敗例

確信度が高かった対訳候補のうち、対訳関係が正解と認められなかった例を示す。ほとんどが、複合語の対訳ペアの未完成(複合語の構成語の一部しか対訳が対応していない)である。

1. 日本語に助詞の「の」が含まれる複合語
   - normal beautification rule ⇔ 通常の美化規則
   - multiple process ⇔ 複数のプロセス
   - default value ⇔ デフォルトの値
   - bug ⇔ 使用上の留意点

   これらの場合、「の」の直前の単語を含めて候補として抽出できていない。「の」が含まれていないならば正しく抽出できる。
2. 英語に前置詞が含まれる複合語
   - default baud rate for the connection ⇔ デフォルトの接続用ボーレート
   - syntax for variable ⇔ 変数の構文
   - amount of data ⇔ データの量
   - end of line ⇔ 行の終り
   - in second ⇔ 秒単位

表 1: 対訳テキストコーパスのアラインメント結果

データ	パラグラフ数	文数(アラインメント前)		文数(アラインメント後)	
		英　語	日本語	英　語	日本語
telnet	55	148	159	148	148
tip	87	280	339	280	280
ftp	99	304	327	304	304

表 2: 対訳テキストコーパスの形態素解析、辞書引き、対訳語抽出結果

データ	文数	総形態素数		異なり形態素数		異なり未登録語数		対訳候補抽出
		英　語	日本語	英　語	日本語	英　語	日本語	処理時間(CPU)
telnet	148	1989	2241	377	413	64	97	約 5 分
tip	280	3372	4136	593	649	109	186	約 12 分
ftp	304	4278	5520	639	777	105	212	約 15 分

- in byte ⇔ バイト単位

3. 日本語に形容詞・形容動詞活用語尾が含まれる複合語
   - printable character ⇔表示可能な文字
   - verbose message ⇔冗長なメッセージ

   これらの場合、「表示可能文字」、「冗長メッセージ」などとなっていれば正しく抽出できる。
4. 受動態が含まれるもの
   - typed character ⇔ タイプされた文字
5. 日本語(または英語)で訳語表記が一貫していない場合英語の"kill character"に対して、日本語側が、「kill 文字」、「キル文字」、「抹消文字」と複数の訳語が存在する場合、うまく候補として抽出できても確信度が低くなる。
6. 意訳の場合
   - see also ⇔ 関連項目

4　まとめ

対訳テキストコーパスさえあれば、機械翻訳システムの形態素解析機能と辞書引き機能を利用することにより、専門用語やユーザ用語の抽出を自動的に行なうことができ、人手をあまりかけずにユーザ独自の辞書を整備することが容易になることがわかった。対訳語候補抽出失敗例を調査すると、ひとつの複合語としてうまくとれてきていない場合がまだ見受けられるため、より詳細な分析と改良が必要である。前置詞や助詞を含む対訳語候補抽出能力の向上については、対訳語の確からしさのパラメータとして、意味カテゴリの類似性なども確信度計算に利用できる可能性がある。テキストアラインメントについては、今回は文対応レベルまでを統計情報算出のベースとしたが、今後は精度向上や句レベルの対応など、対応レベルの精密化するために、より信頼性の高い統計情報が得られるように改良していくことや、言語情報の利用等が考えられる。

今後は、このようにして抽出した対訳語情報のうち一定の閾値以上のものを実際に辞書に登録し、さらに対訳語候補情報の抽出を繰り返し行なうというアニーリング法により、閾値をどのように制御すれば最も効率的かを探ることを考えている。また、機械翻訳システムにおける翻訳品質にどのほどの効果があるかなどを評価することも重要となる。

参考文献

[1] Gale, William A. and Church, Kenneth

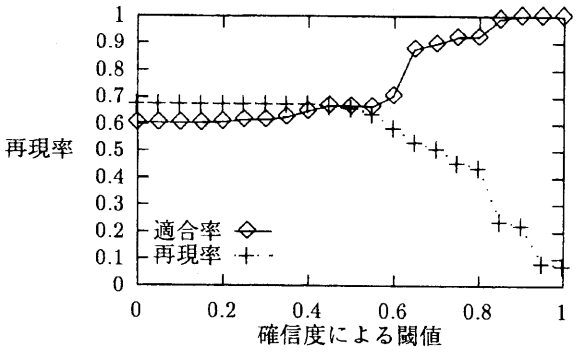

図 1: 確信度による閾値ごとの適合率・再現率

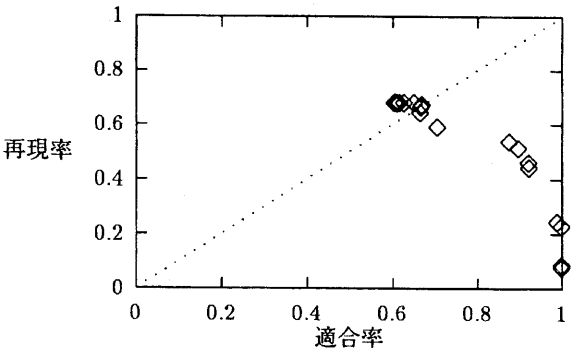

図 2: 適合率・再現率の相関関係

W., A Program for Aligning Sentences in Bilingual Corpora, Proc. of ACL'91, pp.177–184, (1991).

[2] Gale, William A. and Church, Kenneth W., A Program for Aligning Sentences in Bilingual Corpora, Computational Linguistics, 19(1), pp.75–102, (1993).

[3] Brown, Peter F., Lai, Jennifer C. and Mercer Rocer Rovert L., Aligning Sentences in Parallel Corpora, Proc. of ACL'91, pp.169–176, (1991).

[4] Kay, Martin and Röscheisen, Martin, Text-Translation Alignment, Computational Linguistics, 19(1), pp.121–142, (1993).

[5] 春野 雅彦, 山崎毅文, 「辞書と統計を用いたアラインメント」、情報処理学会自然言語処理研究会, 112-4, (1996).

[6] Fung, P. and Church, K., K-vec: A New Approach for Aligning Parallel Texts, COLING-94, pp.1096–1102, (1994).

[7] 田中 康仁, 「パラレルコーパスからの専門用語の抽出」, 情報処理学会自然言語処理研究会, 88-7, (1992).

[8] 山本 由紀雄, 坂本 仁, 「対訳コーパスを用いた専門用語対訳辞書の作成」, 情報処理学会自然言語処理研究会, 94-12, (1993).

[9] 熊野 明, 平川 秀樹, 「対訳文書からの機械翻訳専門用語辞書作成」, 情報処理学会論文誌, Vol.35, No.11, pp.2283–2290, (1994).

表 3: 対訳語候補抽出結果

確信度	品詞対応	日本語	英語	属性	意味カテゴリ
1.00	名詞 - 名詞	端末回線	terminal line	1	11
1.00	名詞 - 名詞	システム名	system name	1	58
1.00	名詞 - 名詞	画面	screen	1	40
1.00	サ名 - 動詞	プリント	print	がをにでから	259
1.00	形容 - 形容	オープン	open	でから	407
1.00	名詞 - 名詞	ローカル端末	local terminal	1	24
1.00	名詞 - 名詞	回線速度	line speed	1	62
1.00	名詞 - 名詞	入力文字	input character	1	58
1.00	サ名 - 動詞	生成	generate	がをにへでよりから	257
1.00	名詞 - 名詞	エラーメッセージ	error message	1	61
1.00	名詞 - 名詞	環境変数	environment variable	1	50
1.00	名詞 - 名詞	デフォルトホスト	default host	1	13
1.00	名詞 - 名詞	データ転送	data transfer	1	59
1.00	名詞 - 名詞	美化スイッチ	beautification switch	1	24
1.00	名詞 - 名詞	時間	amount of time	7	62
1.00	名詞 - 名詞	ブール変数	Boolean variable	1	50
0.94	名詞 - 名詞	ＵＮＩＸシステム	UNIX system	1	20
0.93	名詞 - 名詞	ｔｔｙ標準エラー	tty standard error	1	55
0.93	名詞 - 名詞	リモート回線出力	remote line output	1	05
0.93	名詞 - 名詞	リモート回線入力	remote line input	1	05
0.93	名詞 - 名詞	パーソナル記述ファイル	personal description file	1	23
0.92	名詞 - 名詞	ロック・ファイル	lock file	1	23
0.92	名詞 - 名詞	ローカル・システム	local system	1	20
0.92	名詞 - 名詞	ローカル・プロセス	local process	1	287
0.92	名詞 - 名詞	ファイル・システム	file system	1	20
0.92	名詞 - 名詞	エスケープ・シグナル	escape signal	1	24
0.92	名詞 - 名詞	エスケープ・シーケンス	escape sequence	1	62
0.91	名詞 - 名詞	排他的オープンｉｏｃｔｌ（2）コール	exclusive-open ioctl(2) call	1	05
0.90	名詞 - 名詞	冗長モード	verbose mode	1	59
0.90	名詞 - 名詞	変数	variable	1	67
0.90	名詞 - 名詞	ｔｂ能力	tb capability	1	66
0.90	名詞 - 名詞	標準消去	standard erase	1	279
0.90	名詞 - 名詞	リモート側	remote side	1	40
0.90	名詞 - 名詞	ｐｕｔコマンド	put command	1	58
0.90	名詞 - 名詞	奇数パリティ	odd parity	1	72
0.90	形容 - 形容	ローカル	local	から	407
0.90	名詞 - 名詞	対話式セッション	interactive session	1	0
0.90	名詞 - 名詞	初期化ファイル	initialization file	1	20
0.90	名詞 - 名詞	偶数パリティ	even parity	1	72
0.90	名詞 - 名詞	エコーチェック応答	echo check response	1	256
0.90	名詞 - 名詞	接続メッセージ	connection message	1	61
0.90	名詞 - 名詞	ＴＡＢ文字	TAB character	1	58
0.90	名詞 - 名詞	ＳＰＡＣＥ文字	SPACE character	1	58
0.90	名詞 - 名詞	Ｃシェル	C shell	1	24
0.88	名詞 - 名詞	美化規則	normal beautification rule	1	58
0.88	名詞 - 名詞	大文字マッピング・モード	upper case mapping mode	1	59
0.86	名詞 - 名詞	ファイル転送コマンド	file transfer command	1	58
0.86	名詞 - 名詞	ＡＣＵタイプ	ACU type	1	62