自然言語処理 128−18
（1998. 11. 6）

World Wide Web からの対訳データの自動収集

北 研二　山口 直宏

徳島大学 工学部

〒770-8506 徳島市南常三島町 2–1

{kita,naohiro}@is.tokushima-u.ac.jp

概要

World Wide Web(WWW) 上には、膨大なテキスト情報が蓄積されており、同一の内容を複数の言語で提供しているページも数多く存在する。これらの WWW 上の対訳ページを利用し、対訳コーパスを自動的あるいは半自動的に構築することができれば、コーパス作成に要する人的資源、時間、費用などを大幅に削減することができる。我々は、WWW から日本語と英語の対訳データを自動収集する実験的なシステムを構築した。本稿では、このシステムの概要について紹介する。

Automatically Compiling Multilingual Translations from the World Wide Web

Kenji Kita　Naohiro Yamaguchi

Faculty of Engineering, Tokushima University

2-1, Minami-josanjima, Tokushima 770-8506, Japan

{kita,naohiro}@is.tokushima-u.ac.jp

Abstract

The World Wide Web provides almost unlimited accesses to the textual documents and it also contains parallel pages in many languages. In this paper, we focus on the problem of automatically compiling multilingual translations from the Web. As a first step towards Web-based automatic multilingual corpus creation, we developed an experimental system for compiling Japanese and English translation pairs from the actual Web page documents. In this paper, we describe the system architecture as well as some experimental results.

1 はじめに

近年、対訳コーパスから言語知識データや言語モデルを自動的に構築し、これらの知識ベースやモデルに基づいた自然言語応用システムの研究が活発に行われている。このような研究の例として、異言語間情報検索 (cross-language information retrieval) [8, 13, 17]、統計的機械翻訳 (statistical machine translation) [1, 4, 6]、二言語間構文解析 (bilingual parsing) [23], 対訳用語収集 (bilingual lexicography) [10, 22] 等をあげることができる。

大規模な対訳コーパスの作成には、対訳テキストの組織的・体系的な収集活動が必要であり、多くの人的資源、時間、費用などを要する。ヨーロッパ系の言語に対しては、ELSNET(European Network in Language and Speech) による ECI 多言語コーパス (European Corpus Initiative Multilingual Corpus) 等が公開されているが [18, 19]、日本語を含むアジア系言語に対しては、大規模な対訳コーパスが作成されていないというのが現状である。

我々はインターネット、特に World Wide Web (WWW) 上に存在する文書群に着目した。WWW 上には、膨大なテキスト情報が蓄積されており、同一の内容を複数の言語で提供しているページも数多く存在する。これらの WWW 上の対訳ページを利用し、対訳コーパスを自動的あるいは半自動的に構築することができれば、コーパス作成の手間を大幅に削減することができる。

以上のような観点から、我々は WWW から日本語と英語の対訳データを自動収集する実験的なシステムを構築した。以下では、この実験システムの概要について説明する。なお、類似した研究に Resnik による Parallel Strands [20] があるが、システムの構成が我々のものとは大きく異なっている。

2 システム構成

本システムの構成を図 1 に示す。四角の箱で示された部分が処理モジュールであり、各モジュールはパイプライン的につながれている。以下で、各モジュールについて説明する。

2.1 WWW ロボット

WWW の大きな特徴は、ハイパーテキスト・リンク (hypertext link) によるページ参照である。WWW ロボット (robot, spider, or crawler) は、ハイパーテキスト・リンクを次々にたどることにより、WWW 上を自動的に巡回し、ページを収集・蓄積するプログラムである。本システムでは、NetScoop と呼ぶロボット・プログラムを使用している[1] 。

2.2 対訳候補ページ生成

対訳候補ページ生成モジュールは、与えられたページに対し、そのページの対訳となっている可能性の高いページを見つける。膨大なページを処理する必要があるため、この段階では詳細な処理は行わず、対訳の候補となるページを効率的に探すことを主眼としている。実際に対訳ページであるか否かの判断は、次の段階の文対応付けにて行う。

対訳候補ページ生成モジュールでは、まず与えられたページからハイパーテキスト・リンクを抽出する。一般に、対訳ページは、元のページと同一のドメインにあるので、同一ドメインへの参照ページに対してのみ以下の処理 (2.2.1 ~ 2.2.3) を行い、対訳ページの候補をランク付けて出力する。

2.2.1 ページ記述言語の識別

WWW ページの使用言語 (日本語あるいは英語) を特定することは、対訳候補ページを絞り込む上で非常に有用である。たとえば、与えられたページが英語で記述されていれば、対訳ページは日本語で記述されていなければならないし、逆に日本語で記述されたページの対訳ページは英語で記述されていなければならない。

現在のシステムは、日本語と英語の対訳データ収集を目的としているので、ページ記述言語の識別にはきわめて簡単な方法を用いている。ページ中に「ひらがな」,「カタカナ」, あるいは「漢字」のコードがあったときは日本語ページであると判断し、それ以外は英語ページであるとみなしている。

[1] http://info.webcrawler.com/mak/projects/robots/active/html/netscoop.html

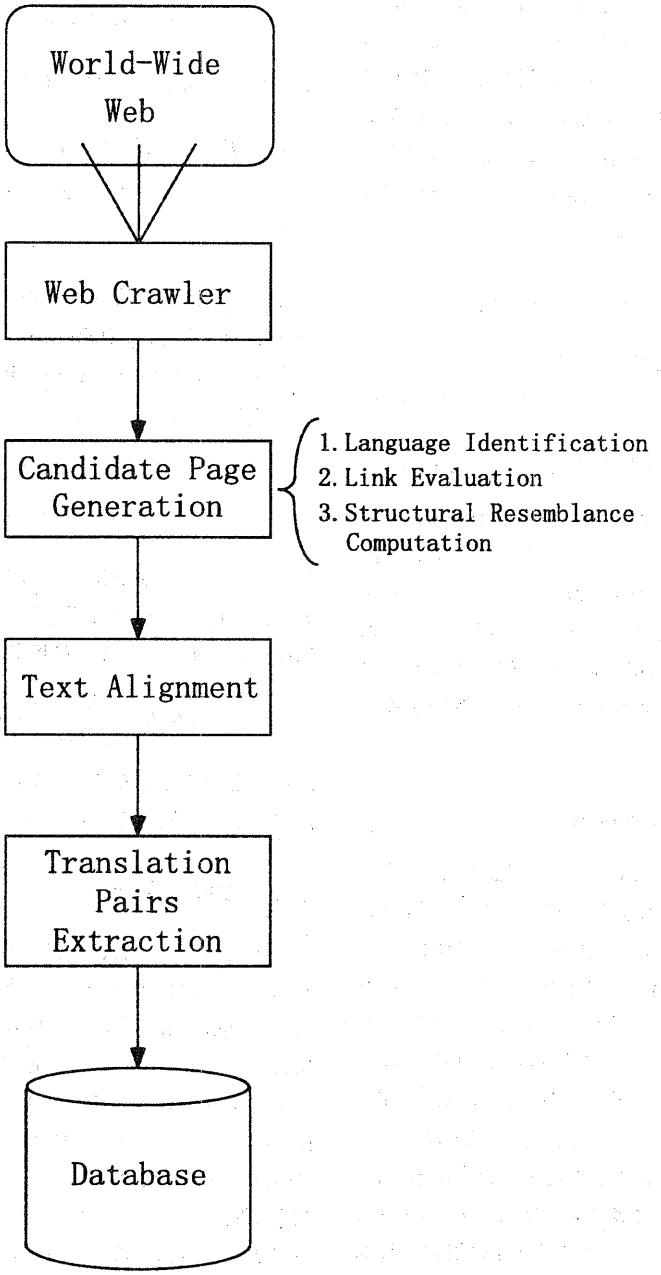
World-Wide Web
Web Crawler
Candidate Page Generation
1. Language Identification
2. Link Evaluation
3. Structural Resemblance Computation
Text Alignment
Translation Pairs Extraction
Database

図 1: システム構成

将来的には、ページ記述言語の識別に確率的言語モデル (文字 N-gram) に基づく方法を採用する予定である。確率的言語モデルに基づく言語識別の精度は非常に高く、テキスト中の単語を 20 語程度利用すればほぼ正確に言語を識別することができる [7, 9, 12, 16, 21]。多数の言語を対象とした対訳データ収集においては、確率的言語モデルに基づく方法が有効であると考えられる。

2.2.2 アンカリング・テキストの利用

WWW ページ (HTML テキスト) 中ではハイパーテキスト・リンクを表すために、次のように記述する。

```
<a href="xxx.html">English page</a>
```

ここで、<a> タグと タグで囲まれた部分 (アンカリング・テキスト) には、一般にリンク先の内容を示すための説明文が書かれている。アンカリング・テキスト中に「日本語」や「英語」、あるいは "Japanese" や "English" 等の文字列が使われていれば、対訳ページである可能性が高いと判断することができる。

2.2.3 ページの構造的類似性

WWW ページは HTML によって記述されるが、対訳ページに対しては類似したマークアップ情報 (タグ) が用いられている場合が多い。また、対訳ページ対から対訳情報を高精度で抽出するためには、これらのページ対が構造的に類似していることが望ましい。本システムでは、ページ記述言語 HTML のマークアップ情報から、ページ対の構造的類似性を調べることにより、対訳ページ候補を絞り込んでいる。

まず WWW ページからタグの名前のみを抽出し、ページをタグの系列に変換する。次に、Broder らによる方法 [2, 3] を用いて、タグ系列間の構造的類似度を計算する。Broder らの方法では、きわめて高速にタグ系列間の類似度を計算することができる。類似度の具体的な計算方法を以下に示す。

記号系列 D 中の連続した記号列を shingle と呼ぶ。長さ w の shingle 全体の集合を w-shingling と呼び、$S(D, w)$ で表す。たとえば、

$$D = (a, b, c, a, b, c, a, b)$$

に対する 4-shingling は、次で与えられる。

$$\begin{aligned} S(D, 4) &= \{(a, b, c, a), (b, c, a, b), (c, a, b, c), \\ &\qquad (a, b, c, a), (b, c, a, b)\} \\ &= \{(a, b, c, a), (b, c, a, b), (c, a, b, c)\} \end{aligned}$$

また、shingle に出現順位を付けたものの集合をラベル付き w-shingling (labelled w-shingling) と呼ぶ。上の例では、ラベル付き w-shingling は、以下のようになる。

$$\begin{aligned} S(D, 4) &= \{(a, b, c, a, 1), (b, c, a, b, 1), (c, a, b, c, 1), \\ &\qquad (a, b, c, a, 2), (b, c, a, b, 2)\} \end{aligned}$$

このとき、w-shingling あるいはラベル付き w-shingling を用いて、系列 A および B の間の類似度 $r_w(A, B)$ を以下で定義する。

$$r_w(A, B) = \frac{|S(A, w) \cap S(B, w)|}{|S(A, w) \cup S(B, w)|}$$

w の値を大きくするほど、より正確に 2 つの系列の間の類似性を測ることができる。

2.3 文対応付け

対訳候補ページ生成により得られた候補ページに対し、リンク元のページとの間の文対応付け (sentence alignment) を行う。対訳コーパスからの文対応付けに関しては、動的計画法に基づいたアルゴリズムが開発されている [5, 11, 15]。また、井ノ上ら [14] は、2 つのコーパス間で共通に出現する英数字 (transword) を利用することにより、文対応付けの精度が向上することを示している。本システムでは、井ノ上らの方法に加え、さらにページのマークアップ情報 (タグ) を利用して、文対応付けを行っている。

文対応付けの際に、対訳候補ページが本当にリンク元のページの対訳であるか否かを判断する。あらかじめ閾値を (実験的に) 定めてあり、閾値よりも低いスコアを持つ対応付けからは対訳情報を収集しない。

http://www-a2k.is.tokushima-u.ac.jp/index.html の対訳候補
(1) http://www-a2k.is.tokushima-u.ac.jp/index-jp.html [score = 0.947215]
(2) http://www-a2k.is.tokushima-u.ac.jp/JavaPhone/index-jp.html [score = 0.034913]
(3) http://www-a2k.is.tokushima-u.ac.jp/member/sasaki/doc/hbang-jp.html [score = 0.012270]
(4) http://www-a2k.is.tokushima-u.ac.jp/visitors.html [score = 0.003928]
http://www.tokushima-u.ac.jp/index.html の対訳候補
(1) http://www.tokushima-u.ac.jp/ENGLISH/index-en.html [score = 0.336011]
(2) http://www.tokushima-u.ac.jp/INTRO/intro.html [score = 0.000000]
http://wwwsoc.nacsis.ac.jp/index.html の対訳候補
(1) http://wwwsoc.nacsis.ac.jp/index-e.html [score = 0.652363]
(2) http://wwwsoc.nacsis.ac.jp/jjap/index.html [score = 0.016682]
(3) http://wwwsoc.nacsis.ac.jp/jaeg/index.html [score = 0.000694]
(4) http://wwwsoc.nacsis.ac.jp/jsd/index.html [score = 0.000693]
http://www.monbu.go.jp/index.html の対訳候補
(1) http://www.monbu.go.jp/emindex.html [score = 0.543919]
http://www.u-tokyo.ac.jp/index.html の対訳候補
(1) http://www.u-tokyo.ac.jp/index-j.html [score = 0.426697]

図 2: 対訳候補ページ生成の出力例

2.4 対訳情報の抽出

文対応付けでは、HTML のマークアップ情報を含んだ対応付けを行っているので、最終的にはマークアップ情報を除外し、日本語と英語の対訳文の対のみを収集し、対訳文データベースに蓄積する。

3 出力例

対訳候補ページ生成の出力例および文対応付けの出力例を、それぞれ図 2 および図 3 に示す。

たとえば、図 2 の最初の例は、http://www-a2k.is.tokushima-u.ac.jp/index.html の対訳候補ページをスコア付きで出力したものである。このページから同一ドメインへのリンク数は 22 あるが、このうち 18 個のリンクはページ記述言語の識別により対訳ページではないと判断された。残りの 4 つのリンクに対してスコアを計算したものを示している。図 2 から分かるように、対訳ページに対してはいずれも高いスコアを与えている。なお、実験では、ラベル付き 3-shingling を用いている。

4 おわりに

本稿では、World Wide Web から日本語と英語の対訳データを自動収集する実験システムの概要について述べた。現在、小規模の実験により、システムの細かいパラメータ (閾値等) をチューニング中である。今後の課題として、数千～数万程度の対訳ページ収集によるシステムの評価、日本語／英語以外の言語への拡張を行いたいと考えている。

謝辞

文対応付けプログラムを提供頂いたＫＤＤ研究所の井ノ上直己氏に感謝する。

```
dist=0.368582
==(prob: 0.000 l1:  29 l2:   0)================
(  29)<!----- Kita Lab. Home Page ----->
------------------------------------------------
================================================

==(prob: 0.000 l1:  59 l2:   0)================
(  59)<!----- Faculty of Engineering, Tokushima University, JAPAN ----->
------------------------------------------------
================================================

==(prob: 0.289 l1:   6 l2:   6)================
(   6)<HTML>
------------------------------------------------
(   6)<HTML>
================================================

==(prob: 0.289 l1:   6 l2:   6)================
(   6)<HEAD>
------------------------------------------------
(   6)<HEAD>
================================================

==(prob: 0.833 l1:  56 l2:  56)================
(  56)<link rev=made href="mailto:adm-a2k@is.tokushima-u.ac.jp">
------------------------------------------------
(  56)<link rev=made href="mailto:adm-a2k@is.tokushima-u.ac.jp">
================================================

(省略)
==(prob: 0.408 l1:  23 l2:  14)================
(  23)Welcome to Kita Laboratory
------------------------------------------------
(   4)<H1>
(  10)ようこそ！ 北研究室へ
================================================

==(prob: 0.425 l1:  23 l2:  20)================
(  23)<a href="index-jp.html">
------------------------------------------------
(  20)<A HREF="index.html">
================================================

==(prob: 0.412 l1:  44 l2:  42)================
(  44)<IMG SRC="/graph/japanese.gif" ALT="Japanese">
------------------------------------------------
(  42)<IMG SRC="/graph/english.gif" ALT="English">
================================================

==(prob: 0.289 l1:   4 l2:   4)================
(   4)</a>
------------------------------------------------
(   4)</A>
================================================

==(prob: 0.289 l1:   5 l2:   5)================
(   5)</H1>
------------------------------------------------
(   5)</H1>
================================================

==(prob: 0.289 l1:   4 l2:   4)================
(   4)<H4>
------------------------------------------------
(   4)<H4>
================================================

==(prob: 0.722 l1:  54 l2:  52)================
(  54)<IMG SRC="/graph/under_construction.gif" ALIGN="middle">
------------------------------------------------
(  52)<IMG SRC="/graph/under_construction.gif" ALIGN=middle>
================================================

==(prob: 0.358 l1:  34 l2:  19)================
(  34)THIS SITE IS ALWAYS UNDER CONSTRUCTION.
------------------------------------------------
(  19)当WWWサーバは、いつも工事中です!!
================================================

(省略)
==(prob: 0.289 l1:   7 l2:   7)================
(   7)</HTML>
------------------------------------------------
(   7)</HTML>
================================================
```

図 3: 文対応付けの出力例

参考文献

[1] Berger, A. L., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Gillett, J. G., Lafferty, J. D., Mercer, R. L., Printz, H. and Ureš, L.: "The Candide system for machine translation", *Proc. of the Human Language Technology Workshop*, pp. 157–162, 1994.

[2] Broder, A. Z., Glassman, S. C., Manasse, M. S. and Zweig, G.: "Syntactic clustering of the Web", *Proc. of the 6th International World Wide Web Conference*, pp. 391–404, 1997.

[3] Broder, A. Z.: "On the resemblance and containment of documents", *Compression and Complexity of Sequences* (SEQUENCES'97), pp. 21–29. IEEE Computer Society, 1998.

[4] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S.: "A statistical approach to machine translation", *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, 1990.

[5] Brown, P. F., Lai, J. C. and Mercer, R. L.: "Aligning sentences in parallel corpora", *Proc. of ACL-91*, pp. 169–176, 1991.

[6] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. and Mercer, R. L.: "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[7] Cavnar, W. B. and Trenkle, J. M.: "N-gram-based text categorization", *Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–169, 1994.

[8] Dumais, S. T., Letsche, T. A., Littman, M. L. and Landauer, T. K.: "Automatic cross-language retrieval using latent semantic indexing", *Proc. of AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*, pp. 18–24, 1997.

[9] Dunning, T.: "Statistical identification of language", *Computing Research Laboratory Technical Report*, MCCS 94-273, New Mexico State University, 1994.

[10] Fung, P.: "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora", *Proc. of ACL-95*, pp. 236–243, 1995.

[11] Gale, W. A. and Church, K. W.: "A program for aligning sentences in bilingual corpora", *Computational Linguistics*, Vol. 19, No. 1, pp. 75–102, 1993.

[12] Grefenstette, G.: "Comparing two language identification schemes", *Proc. of the 3rd International Conference on Statistical Analysis of Textual Data* (JADT-95), 1995.

[13] Grefenstette, G. (Ed.): *Cross-Language Information Retrieval*, Kluwer Academic Publishers, 1998.

[14] 井ノ上直己，野垣内出：「日英文対応データの自動付与方法」，情報処理学会第 45 回全国大会講演論文集 (3)，pp. 107–108, 1992.

[15] Kay, M. and Röscheisen, M.: "Text-translation alignment", *Computational Linguistics*, Vol. 19, No. 1, pp. 121–142, 1993.

[16] 北 研二：「確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築」，自然言語処理，Vol. 4, No. 3, pp. 71–82, 1997.

[17] Landauer, T. K. and Littman, M. L.: "Fully automatic cross-language document retrieval using latent semantic indexing", *Proc. of the*

Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp. 31–38, 1990.

[18] McEnery, T., Wilson, A., Sánchez-León, F., and Nieto-Serrano, A.: "Multilingual resources for European languages: Contributions of the CRATER project", *Literary and Linguistic Computing*, Vol. 12, No. 4, pp. 219–226, 1997.

[19] McKelvie, D. and Thompson, H. S.: "TEI-conformant structural markup of a trilingual parallel corpus in the ECI multilingual corpus 1", *Proc. of the Second Annual Workshop on Very Large Corpora*, pp. 7–17, 1994.

[20] Resnik, P.: "Parallel Strands: A preliminary investigation into mining the Web for bilingual text", *Proc. of the 1998 Conference of the Association for Machine Translation in the Americas*, 1998.

[21] Sibun, P. and Reynar, J. C.: "Language identification: Examining the issues", *Proc. of the Fifth Annual Conference on Document Analysis and Information Retrieval* (SDAIR), 1996.

[22] Smadja, F. and McKeown, K. R.: "Translating collocations for bilingual lexicons: A statictical approach", *Computational Linguistics*, Vol. 22, No. 1, pp. 1–38, 1996.

[23] Wu, D.: "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", *Computational Linguistics*, Vol. 23, No. 3, pp. 377–403, 1997.