

AIDA: コーパスを利用した適応的辞書環境

春野 雅彦

NTT コミュニケーション科学研究所

〒239 神奈川県横須賀市光の丘 1-1

haruno@cslab.kecl.ntt.co.jp

本稿では対訳及び単言語コーパスに対する学習技術と既存の電子辞書を利用した適応的辞書環境 AIDA (Addaptive and Integrated Dictionary Agent) について述べる。利用者は AIDA の提供する 1. 品詞レベルの情報をを用いた柔軟な表現検索ならびにその対話的学習機能, 2. コーパスからのコロケーション抽出による辞書自動作成機能, 3. 辞書とコーパスを相互参照する自動インデックス機能を利用することで各人の応用分野, 好みに応じた辞書環境を作成出来る。また, AIDA のインタフェースは Netscape 上に実装されているため, ネットワークを介して多くの利用者が言語資源を共有することが可能となる。

AIDA: An Adaptive and Integrated Dictionary Agent

Masahiko Haruno

NTT Communication Science Laboratories

1-1 Hikari-No-Oka Yokosuka Kanagawa, 239 Japan

This report describes a user-oriented dictionary system AIDA (Addaptive and Integrated Dictionary Agent). The system, by using currently available large corpora and hand-compiled dictionaries, offers the following three functions; 1. flexible and learnable retrieval of expressions, 2. automatic extraction of collocations and 3. automatic linkage of dictionaries and corpora. By using these functions, users can construct their own environment according to their preference and application domain. In addition, AIDA system is implemented on Netscape and enables language resource sharing.

1 はじめに

我々が言語を利用してコミュニケーションを行なう時、様々な形で辞書を利用する。その目的は未知の単語を調べることであったり、使用する文脈が適切か確認することであったり、より良い表現を探すことであったり、また時には単なる楽しみである場合もある。特に、外国語におけるコミュニケーションを考えた場合には、言語の性質の違い、背景となる文化、慣習の違いによってますます辞書の果たす役割が大きくなっていく。

本稿ではこのような多様な用途を支援する目的で作成した適応的辞書環境 AIDA(Adaptive and Integrated Dictionary Agent) について報告する。このシステムを作成する動機となったのは外国語辞書に求められる以下の一見矛盾する要求である。

項目の簡略性と文脈依存性 辞書に記載する項目は簡潔である方が良いが簡潔過ぎるとはその用法を深く理解することは出来ない。語の用法を正確に知るには抽象的な説明と共にその用例が必要となる。

言語間の相同性と差異性 外国語辞書では2カ国語間の語彙の対応関係に重点が置かれている。しかし実際の使用の場面ではその差異を明らかにしなければならないことが多い。

語彙の一般性と専門性 語彙には用法が一般的なものから特殊な分野の専門用語まで様々なものが存在する。専門用語において必要とされるのは主に2カ国語間の対応と用法であるのに対して、一般の用語ではニュアンスや文化的背景までが必要となる。

これらの条件を1冊(枚)で満足する辞書は存在しない。現状では様々な辞書 [飛田95] を必要に応じて使い分ける必要があり、ユーザは辞書相互間のインタフェースを取り難い。また、仮に上記の条件を満たす辞書が存在したとしても、語の使用の変化、違った分野の必要性などによって辞書の中見は静的ではあり得ない。

AIDAは以上の点に鑑み、辞書を使用者が自ら使い易いよう逐次変更出来るシステムである。近年利用可能となった複数の電子化辞書と対訳、単言語コーパスに対する統計的処理を組み合わせ、同じインタフェース内での相互参照を可能としている。AIDAに実装されている主な機能は以下の通りである。

1. コーパスに対する柔軟な表現検索

AIDAで用いるコーパスの検索は形態素解析によって得られる単語、品詞情報を利用して入力文字列に構文、意味的に類似したデータを類似度の高い順に出力する。類似度の判定に用いる各属性の重みを対話的なインタラクションによって学習し、使用者の好みや分野の特徴を反映する。この検索法は基本的に1つの言語側から行なうもので、対訳、単言語双方のコーパスに対して適用される。

2. 対訳及び単言語コーパスからの自動表現抽出

対訳コーパスに対して形態素情報を用いた統計的手法を適用することで自動的に対訳コロケーション辞書を作成する。この機能は専門性の高い分野の対訳辞書を個々の使用者が作成する上で特に有効である。また、この手法を大量に存在する単言語コーパスに適用することで各言語のコロケーションを調べることが出来る。

3. 全ての検索を一元的に行う Netscape 上の統合インタフェース

複数の電子辞書、自動的に作成されたコロケーション辞書、コーパスの検索が同じインタフェースで行なえるだけでなく、辞書からコーパス、あるいはその逆といった様々な情報源に跨る検索がインデックスを介して高速に行なえる。また、インタフェースが Netscape 上に実装されているため、多様な端末から利用可能であり、色々なレベルでの言語資源共有が可能となる。

以下の章では AIDA の機能とその実現法について例を交えながら説明する。まず2章で AIDA の全体構成を述べた後、3章でコーパスの類似表現検索に用いる検索法、利用する属性、並びに各属性の重みを学習する手法について説明する。4章ではコーパスからの自動表現抽出法について述べる。最後に6章で関連研究に触れ、本稿をまとめる。

2 AIDA の構成

図1に AIDA の全体構成を示す。はじめに、システムが行なうデータの初期化ステップを説明する。システム

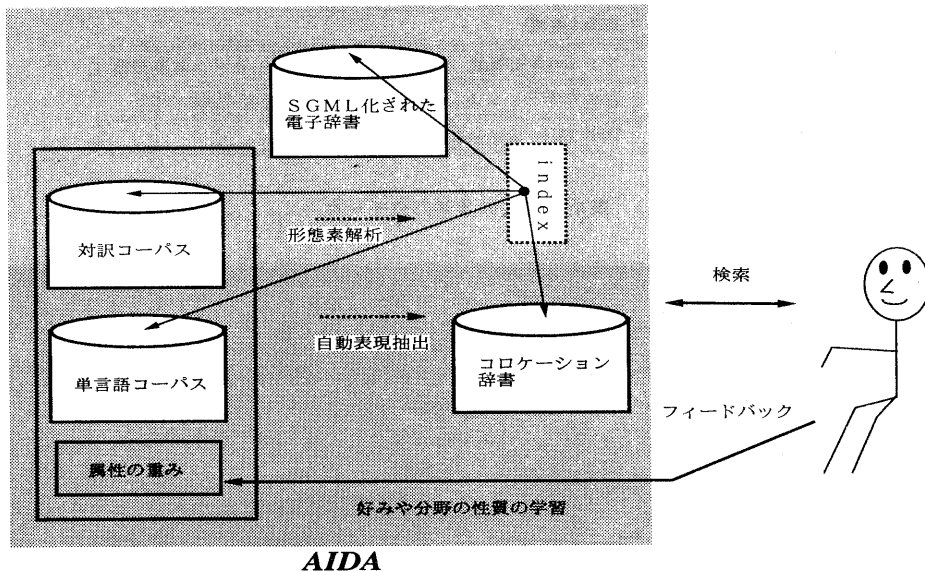


図 1: AIDA の全体構成

ムへの入力は SGML 化された電子化辞書、文対応付けの行なわれた日英対訳コーパス¹、日本語単言語コーパス、英語単言語コーパスである。上記の形の任意のデータを利用可能であるが、我々が実際に用いたデータを表 1²に示す。

これらのデータは日本語、英語とも形態素解析され[松本 他 97][Brill92]、検索の高速化のためパトリシア木の形でインデックス化される。対訳、単言語コーパスに関しては形態素解析の後、4 章に述べる手法を用いてコロケーション辞書が構成される。最後に、システムは構成された各々のデータに対応するインデックスを比較して同じ語彙に対するリンクを張りデータの相互参照を可能にする。

初期化終了後、ユーザは検索を行なうことが出来る。通常の電子辞書を検索したい場合、共通ウィンドウに文字列³を入力し辞書を選択すればその項目が表示され

¹対訳コーパスの文対応付けは辞書と統計を用いた文対応付けシステム BACCS [春野 97, Haruno and Yamazaki96] を利用して行なった。

²これらのデータの使用を許可して頂いた日経サイエンス社ならびに読売新聞社に感謝致します。

³他の電子辞書システム同様、文字列の部分一致によって候補項目を表示することも可能。

る。図 2 に英英、英和辞典で英単語 'habit' を検索した画面を示す。項目中の単語に引かれた下線はその単語が他の辞書の項目として存在することを意味する。

コロケーション辞書を検索することも出来る。図 3 にサイエンスのデータから抽出した対訳コロケーション辞書の例を示す。右側のウィンドウはコロケーションの一覧であり、その中の任意のものに対して、その表現を含む例文を表示することが出来る。左側のウィンドウは 'アルツハイマー病患者 -Alzheimer patients' に対する例文である。

次に表現検索について説明する。ユーザが表現を入力するとシステムは次章に述べる動的計画法に基づく方法で入力に類似する表現をコーパスから抽出し類似度の高い順に表示する。ユーザはその結果を見て不適切な表現が上位、或は逆に適切な表現が下位に表示されていればシステムにフィードバックする。システムはユーザからのフィードバックを用いて各属性の重みを更新することでシステムの挙動を次第に適切なものに変更する。図 4 に日本語の表現 '用地の取得を急ぐ必要がある' を対訳コーパスで検索した画面を示す。左側の画面が検索結果で右側が評価 (フィードバック) 用の画面である。

データ種類	名称 (規模)
電子化辞書	アンカー英和辞典, アンカー和英辞典, コンサイス英英辞典
対訳コーパス	日経サイエンス (65000 ペア), 読売新聞社説 (7000 ペア)
英語単言語コーパス	Wall Stree Journal(1987 ~ 1989 年度)

表 1: AIDA に実装したデータ

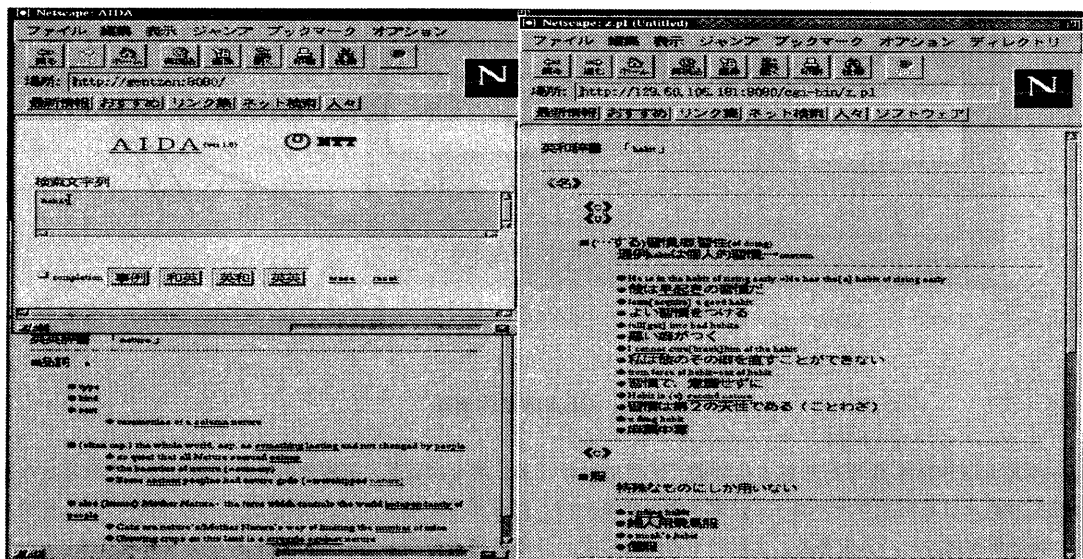


図 2: 辞書検索の画面

ここでは電子辞書、コロケーション辞書、表現検索を別々に説明したが、実際の使用場面では、対象分野、ユーザの語学力などの要因によって、これらの機能が複雑な順序で同時利用される。

3 コーパス中の表現検索

3.1 単語-形態素レベルの動的計画法による表現検索

表現の検索に関してはデータをどこまで処理しておくかに応じて以下の3つの選択肢がある。

1. 何もしない、すなわち文字レベルで検索する。
2. 形態素解析のみ行ない、単語と品詞レベルの情報で検索する
3. 構文解析まで行ない、構文レベルの情報で検索する

文字レベルの手法は日本語の場合、漢字の持つ情報を利用出来るという長所 [Sato92] もあるが、構文的特徴を捕まえた検索が難しい。一方、構文解析を利用すると長距離の依存関係を捉えることが出来るが、事例数が多いほど大きくない限り有効とはならない。また、現在得られる構文解析の精度はこの種のタスクに十分なものとは言えない。以上のような理由から、表現の検索は多くの場合局所的であることも考え合わせ、我々は単語-形態素レベルの情報を利用することとした。

形態素解析の行なわれた入力表現 $s(s_1, s_2, \dots, s_m)$ とコーパス中の文 $t(t_1, t_2, \dots, t_n)$ に対して、類似度 $sim(m, n)$ を動的計画法を用いて以下のように計算する。ただし $m(i, j)$ は単語 s_i と t_j の類似度への貢献を表し、次節で述べる各属性の重みの和として計算される。形態素レベルの動的計画法(サーチの幅4)を用いることで単語の

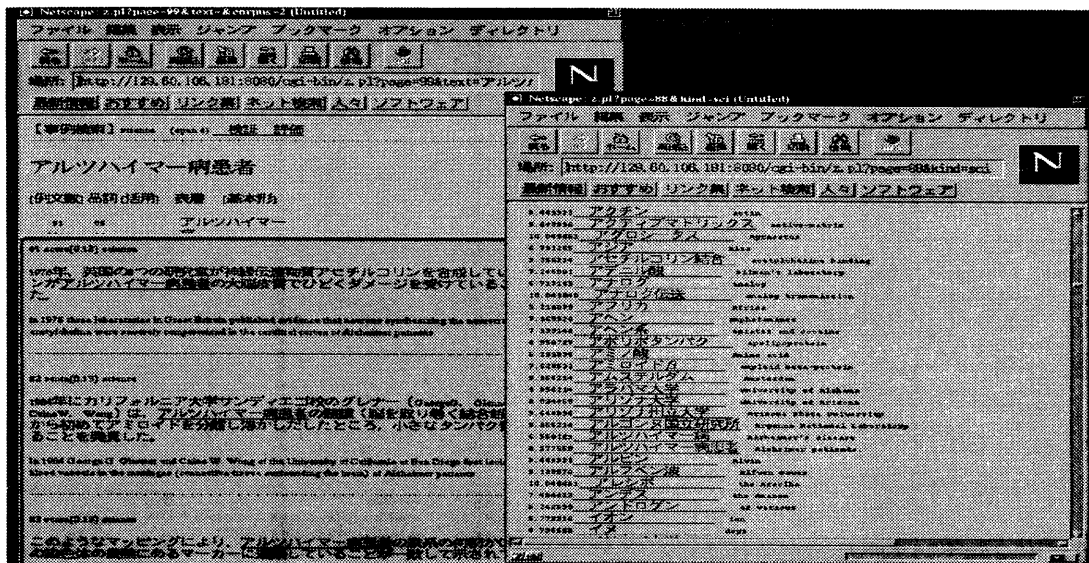


図 3: 対訳コロケーションの表示画面

一致だけでなく、局所的な構造の把握も含めた比較を高速に行なうことが可能となった。

$$sim(i, j) = \begin{cases} 0 & if(i = 0 \vee j = 0) \\ \max \begin{pmatrix} sim(i-1, j-1) + m(i, j) \\ sim(i-1, j) \\ sim(i, j-1) \end{pmatrix} & \text{otherwise} \end{cases}$$

3.2 検索に利用する属性

表 2 に日本語 (英語についても同様に定義される) の単語 s_i と t_j の類似度 $m(i, j)$ を計算するために利用する属性を示す。 $m(i, j)$ は s_i と t_j が満足する属性の持つ重みを加えたものである。図 4 に示すようにこれらの簡単な属性を用いてもかなり正確に意味、構文的に正確な検索が可能となる。他にもシソーラスやクラスタリングによる比較、単語中の漢字種などを考慮に入れることも出来るがこれらは今後の課題である。

3.3 各属性重みのオンライン学習

この節では前節で設定した各属性の重みの学習について説明する。ユーザによる検索とフィードバックは対話的に行なわれるため、全てのデータを一度に処理する

一致属性の種類	個数
単語の表層	1
単語の品詞	品詞の種類
活用語の活用形	活用形の種類
1つ前の単語 (s_{i-1} と t_{j-1}) の表層	品詞の種類 ²
表現の最後の表層	1
表現の最後の品詞	品詞の種類
表現の最初の表層	1
表現の最初の品詞	品詞の種類

表 2: 表現の比較に用いる属性

バッチ学習ではなく、逐次的に学習出来る手法が必要となる。我々は逐次学習アルゴリズム WINNOW [Littlestone88, Littlestone95] を採用する。

WINNOW は、線形分類モデルにおける重みを学習する方法である。理論的解析により、無関係属性が多く存在する場合でも、有効に働くこと、また、その誤り率の上限は、分類に関連ある属性の数に対し線形、全属性数に対して $\log$ の割合で増加することが知られている。

WINNOW は、システムが判定を誤った事例 (AIDA

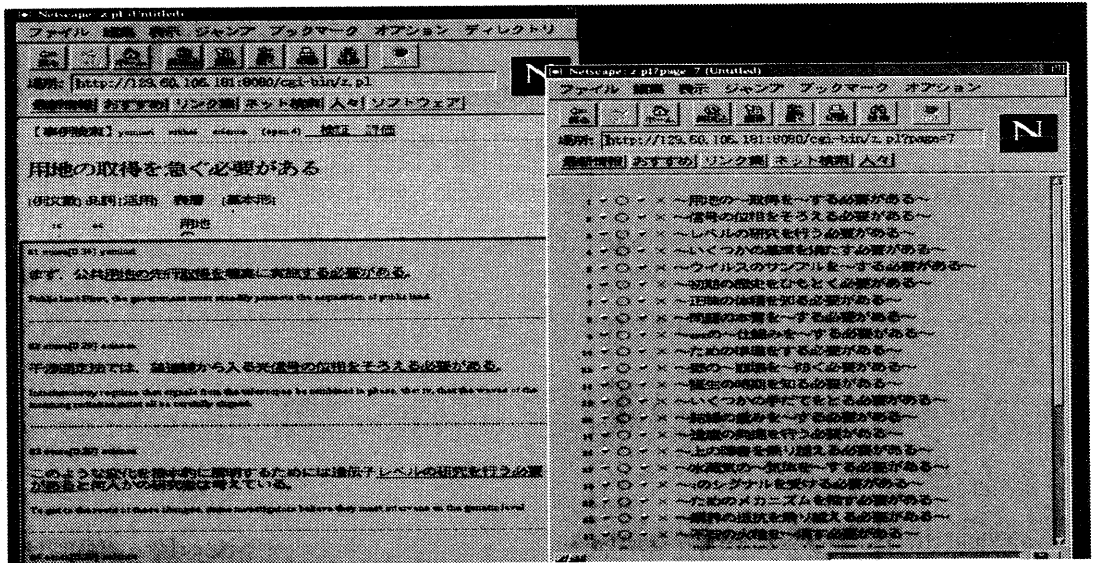


図 4: 表現の検索結果及びその評価の表示画面

の場合で考えると図 4 の評価ウィンドウでユーザが設定したデータ) が一つ入力される毎に、重み更新という形でフィードバックを行なうオンライン型の学習アルゴリズムである。

WINNOW は、閾値 θ と、2 つの更新パラメータ (増進パラメータ $\alpha > 1$ 及び、降下パラメータ $0 < \beta < 1$) の計 3 つのパラメータを持ち、以下の手順で各特徴の重みベクトルを得る。

まず、前節の重みベクトル $w = (w_1, w_2, \dots, w_n)$ をを初期化する。誤り事例が入力されると、次の 2 つの戦略で重みを更新する。(1) AIDA が、不適切と判定し、ユーザが適切の場合、その中に出現する特徴の重みを、 α 倍することにより増進する ($w \leftarrow \alpha \cdot w$)。 (2) AIDA が、適切と判定し、正解が不適切の場合、その事例中に出現する特徴の重みを、 β 倍することにより減少させる ($w \leftarrow \beta \cdot w$)。この重み更新の操作を全訓練事例が収束するまで、あるいは決められた回数だけ繰り返す。WINNOW アルゴリズムを用いることで各属性の重みを、ユーザの好みや分野の特徴を反映した形で逐次学習することが可能となる。

4 コロケーションの自動抽出

(対訳) コロケーション辞書の自動抽出は以下のステップで行なう [Haruno et al.96]。

1. 図 5 の示すように単語毎にポインタを設定する⁴。
個々のポインタは、文中のその場所以降の部分文字列 (suffix) を表す。このポインタをソートすることで図 6 に示すテーブルを得る。
2. テーブル中で各単語列が何回出現したかカウントする。ただし形態素情報を用いて表現として相応しくないものは考慮しない。予め決められた回数以上出現したものを文内最長一致の原則で取り出す。
3. 上記で取り出した日英表現の相互情報量を計算することにより対訳コロケーションを抽出する。

この手法によって大規模な対訳、単言語コーパスから効率良くコロケーション抽出を行なうことが可能となる。

⁴実際には表現の先頭として相応しくない品詞はこの段階で切除する

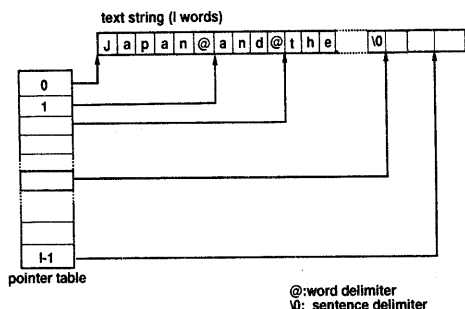


図 5: 単語レベルのソート

sent no.	adopt	coincidence	string
24		10	Japan@and@China@
105		10	Japan@and@Costa Rica
1064		16	Japan@and@the@US
3		16	Japan@and@the@US
2104		16	Japan@and@the@US
1702		16	Japan@and@the@US
1104		16	Japan@and@the@US
104		16	Japan@and@the@US

図 6: ソート結果の集計

5 関連研究

電子辞書、コーパスに関わる研究は膨大な数に登るため、ここでは我々の研究と特に関係の深いものについてのみ触れる。

表現の柔軟な検索に関しては[隅田・堤 91]と[Sato92]がある。[隅田・堤 91]は格構造を考慮に入れた検索が可能で、長い表現の検索に対し無駄な出力を減らすことが出来る。しかしながら、実際の場面で格構造レベルの一般化が有効であるためには膨大な数のデータが必要となる。また現在の構文解析の精度では現実のデータを用いた場合の信頼性に疑問が残る。一方、[Sato92]は文字の一致による動的計画法によって検索を行なっている。日本語では漢字が意味を持つため、文字レベルの検索によって'観察'から'観測'といった表現を取り出すことが可能となる。しかし実際の検索要求に多いフレーズ表現では文字だけでなく形態素情報を使った一般化が必要となる。この様な理由から、我々は局所的な構造までを考

慮に入れて、形態素解析⁵をしたデータに対して検索を行なった。もちろん我々の方法に文字レベルの一致やソーラスの情報を属性として与えることは容易である。

辞書とコーパスを組み合わせる研究として[Klavans90]と[Nerbonne96]がある。[Klavans90]は少数の動詞に限り既存の対訳辞書に、コーパスから得られる新しい用法、頻度情報などを付加する手法を提案している。[Nerbonne96]はコミュニケーションを支援することに徹し、仏蘭辞書、フランス語コーパスを同一画面で単語検索するシステムである。これらに対し、AIDAでは辞書とコーパスのそれぞれの特徴にあった処理を行ない、それらを有機的に結合した。この点でAIDAはコビルド英語辞典[COBUILD95]を参考にした。コビルドと異なるのは個々のユーザレベルで辞書とコーパスの選択、加工を可能としたこと、2ヶ国語の環境においてこそ例文が大切であると考えたことの2点である。

今後AIDAに付加しなければならない機能として意味的曖昧性の解消がある。統計的手法(例えば[Luk95])によってコーパス中の単語が辞書のどの用法に対応するかを明らかにすることで、辞書とコーパスをより密接に結び付けることが出来ると思われる。

6 結論

本稿では適応的な辞書環境AIDAについて述べた。形態素解析と簡単な学習処理を組み合わせることで既存の電子辞書とコーパスを効果的に組み合わせることが出来た。今後は実際の使用場面を調査しより柔軟な機能を付加する予定である。

参考文献

- [Brill92] Eric Brill. A simple rule-based part of speech tagger. In *Proc. Third Conference on Applied Natural Language Processing*, pages 152-155, 1992.
- [COBUILD95] COBUILD. *COBUILD English Dictionary*. Collins, 1995.
- [Haruno and Yamazaki96] Masahiko Haruno and Takafumi Yamazaki. High-Performance Bilingual Text

⁵形態素解析の誤りに関しては約4%あるが、この誤りは検索要求に対して一貫性を持って起こるので大きな問題とはならなかった。

- Alignment Using Statistical and Dictionary Information. In *Proc. 34th ACL*, 1996.
- [Haruno *et al.*96] Masahiko Haruno, Satoru Ikehara, and Takefumi Yamazaki. Learning Bilingual Collocations by Word-level Sorting. In *Proc. 16th COLING*, pages 525–530, 1996.
- [Klavans90] Judith Klavans and Evelyne Tzoukermann. The BICORD System. In *Proc. 13rd COLING*, pages 174–179, 1990.
- [松本 他 97] 松本 裕治 他. 日本語形態素解析システム『茶筌』使用説明書. 奈良先端科学技術大学院大学, 1997.
- [Littlestone88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
- [Littlestone95] Nick Littlestone. Comparing several linear-threshold learning algorithm on tasks involving superfluous attributes. In *Proc. 12nd International Conference on Machine Learning*, pages 353–361, 1995.
- [Luk95] Alpha K. Luk. Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions. In *Proc. 33rd ACL*, 1995.
- [Nerbonne96] John Nerbonne and Petra Smit. GLOSSERuG: in Support of Reading. In *Proc. 16th COLING*, pages 830–835, 1996.
- [Sato92] Satoshi Sato. CTM: an example-based translation aid system. In *Proc. 14th COLING*, pages 1259–1263, 1992.
- [隅田・堤 91] 隅田 英一朗 and 堤 豊. 翻訳支援のための類似用例の実用的検索法. 電子通信学会論文誌, J-74(D-II)(10):1437–1447, 1991.
- [春野 97] 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, 38(4):719–726, 1997.
- [飛田 95] 飛田 茂雄. 役に立つ辞書をさがす本. バベル・プレス, 1995.