自然言語処理 134－20
（1999. 11. 26）

帰納的学習を用いた訳語推定手法における解析的知識の有効性について

笹岡久行† 荒木健治‡ 桃内佳雄† 栃内香次‡

†北海学園大学工学部 ‡北海道大学大学院工学研究科

あらまし

我々は、機械翻訳システムにおける辞書未登録語処理の問題の解決を目指し、帰納的学習を用いた訳語推定手法を提案し、その有効性を確認した。この手法では、単語と訳語の組みの中に存在する単位を帰納的学習を用いて獲得し、その単位を組み合わせることにより訳語を生成する。しかし、字面情報のみに基づいて帰納的学習を行うので、単位の過分割や抽出の見逃し等があり、十分な量の有効な単語片対を獲得することができなかった。そこで、本稿では字面情報に加え、既存の形態素解析システムから得られる単語あるいは形態素の区切り位置情報と品詞情報を利用した単語片対の抽出手法を提案する。さらに、本稿では本手法に基づいた実験システムを用いて行った評価実験の結果についても述べる。実験データ中で複合語名詞の評価対象を限定した場合、正推定率は66.7%であった。このことから本手法の有効性を確認した。

The Effectiveness of Prediction Method Using Inductive Learning with Information of Morphological Analysis

Hisayuki Sasaoka† Kenji Araki‡ Yoshio Momouchi† Koji Tochinai‡

† Faculty of Engineering, Hokkai-Gakuen University

‡ Graduate School of Engineering, Hokkaido University

Abstract

We have proposed the prediction method of target word using inductive learning and confirmed the effectiveness of it. In our method, the units for the prediction are extracted from the pairs of words and target words. Our system acquires the units with the only information of character strings. However, the number of effective units for the prediction is insufficient on our proposed method. In this paper, we propose the prediction method of target words using inductive learning with the units extracted by the information of the morphological analysis. In our method, this information can be given by the morphological analysis system. In this paper, we describe the results of the evaluation experiment and confirm the effectiveness of this method.

1 はじめに

我々は，機械翻訳システムが抱える辞書未登録語の問題の解決を目指し，帰納的学習を用いた訳語推定手法の提案し，その有効性を確認した [1]．辞書未登録語処理の問題とは，機械翻訳システムの辞書に登録されていない単語が翻訳対象に出現した場合，その単語の翻訳に失敗してしまい，処理ができないという問題である．我々が提案した訳語推定手法では，訳語推定に利用する単位を「単語片対」と呼ぶ．2 組の単語と訳語を抽出元とし，字面情報に基づく帰納的学習を用いて単語片対を獲得する．しかし，既に提案した訳語推定処理では，実用に耐え得る精度を達成するには至っていない [1]．

我々は，これまでに帰納的学習を用いた手法に関する種々の研究を行ってきた [2, 3, 4]．本研究においても，帰納的学習に対する基本的な考え方は同じである．つまり，実例から訳語推定のための規則を獲得し，その獲得されたものを利用して処理を進めるという考え方である．研究 [2, 3, 4] では，人間の幼児が持つ言語の獲得能力に着目し，研究が行なわれている．一方，本研究では，帰納的学習を用いたシステムをより実用的なシステムに近づけるということを目指している．

そこで，本研究では，帰納的学習を用いた訳語推定手法において，既存の形態素解析システムの解析結果から得ることができる情報を帰納的学習において利用する．字面情報に基づく単語片対の抽出に加え，既存の形態素解析システムの解析結果から得られる情報に基づく単語片対の抽出を行う．この抽出により，訳語推定処理に有効な単語片対を獲得する．本稿では，その抽出手法及びそれらの抽出結果を利用した訳語推定処理手法の提案をする．

第 2 章では，複合語及び派生語の翻訳手法等のる従来研究に言及し，その問題点について検討する．また，第 3 章では我々の訳語推定手法において利用する単語片対の定義について述べる．さらに，字面情報に基づく抽出に加え，既存の形態素解析システムから得られる情報に基づく抽出について述べ，その処理例を示す．第 4 章では，本手法を基にして作成した実験システムの処理概要について述べる．第 5 章では，本手法の有効性を確認するために行った評価実験について述べる．第 6 章では，本研究において明らかになったことについてまとめる．

2 従来の研究について

従来は，機械翻訳システムの辞書の登録語数を多くすることにより，辞書未登録語処理の問題の解決しようとした [5]．しかし，膨大な登録語数を持つ辞書を人手により作成するには，非常に大きな労力を必要とする．また，実際の翻訳対象には多くの新出語が出現するため，全ての単語を辞書に登録することは非常に困難である．

そこで，機械翻訳のための辞書を自動的に作成する幾つかの手法が提案されている [6, 7, 8]．また，辞書未登録語となるものには複合語や派生語が多いので，その翻訳手法や辞書作成手法等が提案されている．佐藤らの研究 [9] では，専門用語である複合語の翻訳手法を提案している．この手法では，翻訳対象の構成要素とその訳語の構成要素との間の対応付けがなされたものを利用し，翻訳を行う．しかし，構成要素間に対応付けがなされた単語と訳語の組を大量に用意する必要がある．また，各分野の専門用語毎にこれを用意するには，多くの人手が必要である．さらに，それらの組の中の構成要素間の対応関係の作成にはその分野に関する知識も必要となる．佐田らの研究 [10] では，英日翻訳システムにおいて，辞書に未登録である派生語の翻訳手法を提案している．この手法では，単語に接続される各接辞毎に接辞の解析規則及び訳語生成規則を予め用意し，その規則に基づいて翻訳処理を行う．しかし，このような規則に基づく手法では予め用意された規則の例外となる場合への対処や規則として規定されていない場合へ

の対処を考慮する必要がある．また，接辞を処理するための規則を作成するには大きな労力を要する．特に，訳語生成規則は翻訳対象の分野毎に異なる規則が必要と考えられるために，扱う分野毎に人手により用意する必要がある．藤井らの研究 [11] では，多言語情報検索での利用を想定した複合語翻訳のための語基辞書の作成方法を提案している．この手法では，複合語を分割する際の分割コストを計算し，そのコストが最小になる分割に従い，複合語の分割を行なう．しかし，この手法では，言語フィルターを用いて，辞書作成中に不当な分割候補を削除する必要がある．つまり，作成される辞書の精度は言語フィルターの精度に依存する．そのため，良質な言語フィルターの作成のためには大きな労力とその処理対象分野での知識が必要である．また，分割のコストを計算する際には，大量の用例を用いて計算する必要がある．田中らの研究 [12] では，対訳辞書や類義語辞書を利用し，大量の非対訳コーパスから複合名詞の対訳表現を抽出する手法を提案している．そして，出現頻度等に基づく評価関数を利用している．そのために，良質な評価関数を得るには非常に大量のコーパスを必要とする．

これらの手法に対して，我々は辞書未登録語処理の問題を解決するために帰納的学習を用いた訳語推定手法を提案した．我々が提案した手法では，単語片対という単位を利用して訳語推定を行なっている．単語片対はシステムによって単語と訳語の組から自動的に獲得される．

3 概要

3.1 本手法の考え方

我々が提案した帰納的学習を用いた訳語推定手法の基本的な考え方は，単語と訳語の中に存在する単位である単語片対を抽出し，それらを組み合わせて，翻訳を行うというものである [1]．既に我々が提案した手法では，単語と訳語の字面情報に基づいて，単語と訳語における共通部分と差異部分を単語片対としてそれぞれ抽出する．そして，抽出した単語片対を組み合わせることにより，新たな単語の訳語を生成する．この単語片対は，単語と訳語を構成する単位となりうるものである．

しかし，これまでに行った評価実験の結果から，我々が提案した手法では実用に耐え得るような精度は達成できてはいない．この原因の一つは，獲得される単語片対の過分割のためである．これは，字面情報のみに基づき単語片対という単位を抽出し，その使用頻度等によりその尤度評価を行っているので，短い単位の方が数多く利用され，そのために他のものよりも優先されるということである．これを防ぐためには，字面情報以外の情報による学習処理への制約が必要であると考えられる．また，もう一つの原因は，データスパースネスのために有効な単語片対が獲得できていないということである．文献 [13] 等で言及されているように，コーパス中には出現頻度が低い語が数多く存在する．そのような学習データから，字面情報のみに基づいて単語片対を獲得するのは大変困難である．

そこで，本研究では，有効な単語片対を獲得するために，字面情報に基づく帰納的学習に加え，形態素解析結果から得られる情報に基づく帰納的学習を行なう．この情報とは，単語あるいは形態素の区切り位置の情報とそれに付与される品詞情報である．ところで，形態素解析結果から得られる情報だけを用いて訳語推定を行うことも考えられる．しかし，用意した解析的な知識に対する例外や用意した規則だけでは解析不可能な場合があるので，そのようなものだけでは十分ではないと考えられる．

3.2 単語片対抽出例

最初に，字面情報に基づく単語片対の抽出処理について説明する．この処理では，字面の共通部分と差異部分を単語と訳語からそれぞれ抽出し，それらを対として単語片対を得る．図 1に，この字面情報に基づく単語片対

の抽出例を示す．この例の抽出元では，英語側は「 electron」が共通な部分であり，それ以外が異なる部分である．日本語側は「電子」が共通な部分であり，それ以外が異なる部分である．また，単語片対中の「@1」は変数を表している．この位置に，他の文字列を代入することにより新たな文字列を生成することが可能になる．これにより，図1のような3つの単語片対を得る．

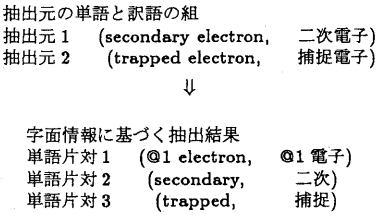

図 1: 字面情報に基づく単語片対抽出例

次に，単語および形態素の区切り位置の情報に基づく単語片対の抽出処理について説明する．英語の単語は，スペースやハイフン等を用いてその区切り位置が与えられていることが多い．しかし，日本語ではそのような形態素の区切り位置は明示的には与えられていない．そこで，複合語およびその訳語を複数の単語あるいは形態素へと分割可能な場合には分割する．そして，その分割に基づいて，共通な単語の並びと異なる単語の並び，あるいは，共通な形態素の並びと異なる形態素の並びをそれぞれ抽出し，それらを対として単語片対を得る．差異部分に複数の単語あるいは形態素が含まれる場合，その差異部分全体を一つの単位として扱っている．図 2に，この区切り位置情報に基づく抽出例を示す．図中の "||" は形態素の区切り位置を示す．この抽出例においても，先の抽出例と同様の単語片対が抽出される．

最後に，品詞情報に基づく単語片対の抽出について述べる．この処理では，既に単語あるいは形態素へと分割されたものに対して，品詞を付与し，その品詞情報を単語片対の抽出

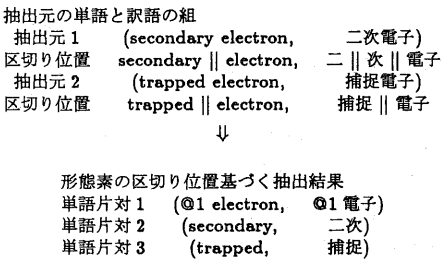

図 2: 区切り位置情報に基づく単語片対抽出例

に利用する．つまり，それぞれの品詞の並びの中から共通な品詞の並びと異なる品詞の並びを抽出し，それに対応する字面を対として単語片対を得る．図 3に，品詞情報に基づく単語片対抽出例を示す．この例においても，上述した 2 つの抽出例と同様の単語片対が抽出される．

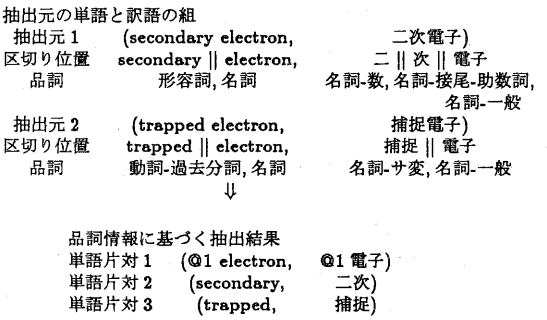

図 3: 品詞情報に基づく単語片対抽出例

4 処理過程

図 4に，実験システムの概要を示す．システムは，推定対象単語が入力されると，既に獲得している単語片対のみを利用して訳語推定を試みる．もし，獲得された単語片対を用いて推定が完了しない場合には，英和辞書の単語と訳語の組の間から新たな単語片対を抽出し，それらを用いて訳語推定を試みる．そして，訳語推定処理において複数の推定結果が生成された場合，各推定結果を構成してい

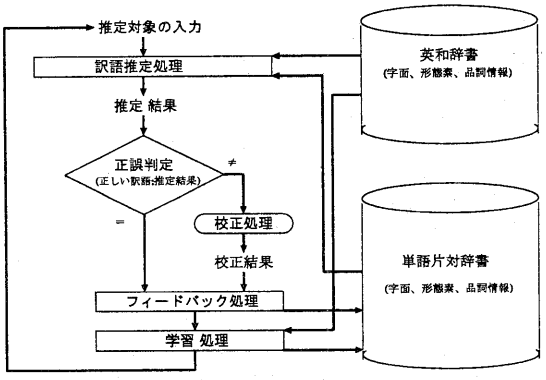

図 4: 実験システム

る単語片対が既出の単語と訳語の組に包含される回数，過去の利用状況を示す数値である出現度数，正推定度数および誤推定度数を参照し優先順位を決定する．その後，推定結果の正誤判定を行い，推定結果が誤ったものであった場合だけ，誤った推定結果に対して人手により校正処理を施す．次にフィードバック処理を行なう．この処理では，その正誤判定結果に応じて，推定結果を構成する各単位の出現度数と正推定度数あるいは誤推定度数を操作する．これにより，これ以降のシステムが持つ訳語推定能力を向上させる．そして，最後に学習処理では新たな単語片対の抽出を行う．まず，学習処理では処理対象であった単語と正しい訳語の組を単語片対辞書に登録する．そして，この対と初期辞書の見出し語と訳語の組とを抽出元として新たな単語片対の獲得を行なう．さらに，単語片対辞書中に登録された単語片対同士からも新たな単語片対の抽出を行なう．

5 評価実験

5.1 実験方法

本手法を元に作成した実験システムを用いて行った評価実験の結果について述べる．実験データとしては，インターネット上のホームページから取り出した大学の講座名や専門分野名等の英語の複合名詞とその日本語の訳語の組 50 組を用いた．

そして，形態素解析結果を得るために用いた既存の形態素ツールは，英語に関しては「Tagger」 [14] を，日本語に関しては「茶筅」 [15] を，初期状態の文法設定のままそれぞれ利用した．また，単語片対抽出のために利用する英和辞書は「gene」 [16] を利用した．この電子化された辞書の見出し語と訳語の組に対して，上述のツールを用いて形態素解析結果を付与した．

推定結果は，推定が完了したものと未了であるものに分類した．さらに，推定が完了したものの中で,「優先順位 10 位以内に文脈に適合した訳語と一致する推定結果が存在するもの」を正推定,「推定を完了したが，優先順位 10 位以内に文脈に適合した訳語と一致する推定結果が存在しないもの」を誤推定と分類した．

5.2 実験結果

表 1に，本手法における推定完了と推定未了のものの数と割合を示し，さらに，比較のために従来手法 [17, 18] における実験結果も併せて示した．また，表 2では，本手法と従来手法における正推定と誤推定の数と割合を示した．この結果において，本手法の方が従来手法より高い精度で訳語推定が行われていることが確認される．この結果から，本手法が従来手法に比べて有効に働くと考えられる．

6 考察

6.1 実験データについて

文献 [1] において述べたように本手法では要素合成原理に基づいて訳語を生成している．そのために，単語と訳語の構成要素の間に対応を持つ場合には，非常に有効に働く．そのような対応を持つ単語と訳語の組は，実験データ 50 組において 40 組存在した．実際の例としては,「laser engineering, レーザ工学」等があった．また，そのような実験データの基準に該当しない例としては,「electric power laboratory,

表 1: 本手法と従来手法の推定完了と推定未了

	推定完了数	率 [%]	推定未了数	率 [%]	データ数	率 [%]
本手法	26	52.0	14	48.0	50	100.0
従来手法	22	44.0	28	56.0	50	100.0

表 2: 本手法と従来手法の正推定と誤推定

	正推定数	率 [%]	誤推定数	率 [%]	推定完了数	率 [%]
本手法	14	53.8	12	46.2	26	100.0
従来手法	6	27.3	16	72.7	22	100.0

電力システム研究室」等があった．この例の中では，日本語の「システム」に関する部分には対応関係がなかったので，実験データとはしなかった．そこで，もう一度実験データを評価すると，実験データの総数が 40 であるために本実験での推定完了率は，65.0%となる．

次に，誤推定結果となったものの実験データについて考察する．本手法では，実験データや初期辞書から帰納的学習を用いて訳語推定に利用する単語片対を獲得している．そのために，実験データや初期辞書に推定対象の単語と訳語の組を構成するような抽出元が存在しない場合，本手法を用いて訳語推定を行うことは不可能である．このような単語と訳語の組は，実験データ 50 個中に 14 個存在した．特に，誤推定結果の中には 5 個存在した．このような単語は本手法を用いるには不適切な単語とみなし，評価の対象から外すと，21 個の推定完了したものの中で 14 個の推定に成功していることになる．そのために，本手法を用いた純粋な正推定率は約 66.7%となる．

このように，本手法では単語と訳語の各構成要素に対応があり，その構成要素を他の単語と訳語の組から獲得できるような環境では比較的高い正推定率が得られることが確認できる．このように，本手法は上述した条件を満たすような単語の訳語推定に利用することが望まれる．

6.2 区切り位置情報に基づく抽出の有効性

区切り位置情報に基づく抽出の有効性について考察する．従来手法では誤推定であった推定結果が，本手法では正推定となった推定結果の個数は 4 個であった．図 5にそのような正推定の処理例を示す．この推定処理において，単語片対 1「system, システム」は英和辞書に予め登録されていた．また，単語片対 2「@1 control, @1 制御」は，英和辞書の見出し語と訳語の組「advanced control, 先行制御」と「automatic control, 自動飛行制御」の間から図 6のような区切り位置情報に基づいた抽出から得られている．さらに，単語片対 3「@1 laboratory, @1 研究室」は「applied electric power laboratory, 電力応用研究室」と「airborne laser laboratory, 機上レーザー研究室」の間から図 7のような区切り位置情報に基づいた抽出から得られている．

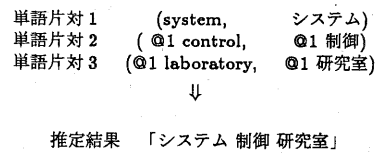

図 5: 正推定例 1

図 6: 区切り位置情報に基づく抽出例 1

また，従来手法では推定未了であった推定

抽出元の単語と訳語の組

抽出元 1	(applied electric power laboratory, 電力応用研究室)
区切り位置	applied || electric || power || laboratory, 電力 || 応用 || 研究室
抽出元 2	(airborne laser laboratory, 機上レーザー研究室)
区切り位置	airborne || laser || laboratory, 機上 || レーザー || 研究室

⇓

単語片対 1	(@1 laboratory,	@1 研究室)
単語片対 2	(applied electric power,	電力応用)
単語片対 3	(airborne laser,	機上レーザー)

図 7: 区切り位置情報に基づく抽出例 2

処理が本手法では正推定となった推定結果の個数は 4 個であった．図 8にそのような正推定の例を示す．この推定処理において，単語片対 1「power @1, パワー@1」は図 9に示すように区切り位置情報に基づいて抽出された．また，単語片対 2「electronics, エレクトロニクス」は，英和辞書の見出し語と訳語として予め登録されていたものである．さらに，単語片対 3「@1 laboratory, @1 研究室」は上述した正推定の処理例で示した図 7のように抽出されている．

単語片対 1	(power @1,	パワー@1)
単語片対 2	(electronics,	エレクトロニクス)
単語片対 3	(@1 laboratory,	@1 研究室)

⇓

推定結果　「パワーエレクトロニクス研究室」

図 8: 正推定例 2

抽出元の単語と訳語の組

抽出元 1	(power balance,	パワーバランス)
区切り位置	power || balance	パワー || バランス
抽出元 2	(power conveyer,	パワーコンベヤ)
区切り位置	power || conveyer	パワー || コンベヤ

⇓

単語片対 1	(power @1,	パワー@1)
単語片対 2	(balance,	バランス)
単語片対 3	(conveyer,	コンベヤ)

図 9: 区切り位置情報に基づく抽出例 3

6.3　品詞情報に基づく抽出の有効性

本稿における実験では，品詞情報に基づく抽出により得られた単語片対が有効に働く例は確認されなかった．この原因の一つに実験データの少なさがあると思われる．また，この他に複合名詞の品詞の並びの類似性に原因があると考えられる．田中らの研究 [12] における複合名詞の品詞の構成に関する調査から，複合名詞の品詞の並びには多様性が非常に少ないことが分かる．そのために，品詞情報に基づく単語片対の抽出では，他の情報に基づく単語片抽出に比べて獲得される単語片対の数が少ないために，有効性を確認することはできなかったと考えられる．今後は，この品詞情報の有効な利用方法について検討する予定である．

7　おわりに

帰納的学習における解析的知識の利用の有効性を確認するために，本研究では，帰納的学習を用いた訳語推定手法において，形態素解析結果から得られる 2 種類の情報に基づいた単語片対抽出手法を新たに提案した．この 2 種類の情報とは，単語および形態素の区切り位置の区切り位置の情報とそれらに付与される品詞情報であった．そして、今回提案した訳語推定手法では，字面情報と形態素解析結果から得られる 2 種類の情報に基づく単語片対抽出から得られたものを利用して訳語推定を行った．本手法に基づく実験システムを用いて行った評価実験では，評価対象を本手法に適する複合名詞に限定した場合，正推定率が約 66.7%であった．この結果から本手法の有効性を確認した。今後は，さらに品詞情報の有効な活用方法について検討する予定である．そして，大量の実験データを用いた評価実験を行う予定である。

謝辞

本研究の一部は文部省科学研究費（No. 09878070, No.10680367 ）および北海学園大学ハイテク・リサーチ・センター研究費による補助のもとに行われた。

参考文献

[1] 笹岡久行, 荒木健治, 桃内佳雄, 栃内香次, "帰納的学習を用いた訳語推定手法の派生語および複合語における有効性の評価", 信学論 (D-II), vol.J81-D-II, No.9, pp.2146 – 2158, 1998.

[2] 荒木健治, 栃内香次, "帰納的学習による語の学習および確実性を用いた語の認識", 信学論 (D-II), J75-D-II, No.7, pp.1213 – 1221, 1992.

[3] 荒木健治, 高橋祐治, 桃内佳雄, 栃内香次, "帰納的学習を用いたべた書き文のかな漢字変換," 信学論 (D-II), J79-D-II, No.3, pp.391 – 402, March 1996.

[4] 内山智正, 荒木健治, 宮永喜一, 栃内香次, "帰納的学習による機械翻訳手法の評価実験", 情処自然言語処理研究報告, NL 93-4, pp.23–30, 1993.

[5] 野村浩郷, 田中穂積 共編, "機械翻訳", bit 別冊, 共立出版, 1988.

[6] 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", 情処自然言語処理研究報告, NL94-12, pp.85–92, 1993.

[7] 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", 情処学論, vol.35, no.11, pp.2283–2290, 1994.

[8] 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", 情処学論, vol.38, no.4, pp.727–736, 1997.

[9] 佐藤理史, "アナロジーによる機械翻訳", 認知科学モノグラフ 4, 共立出版, 1997.

[10] 佐田いち子, 吉見毅彦, 九津見毅, 岡本純子, "英日機械翻訳システムにおける派生語推定とその訳語付与の改良", 言処学会, 第 5 回年次大会, pp15–16,March 1999.

[11] 藤井敦, 石川徹也, "日本語複合語の自動分割と日英語基対訳辞書の作成", 情処自然言語処理研究報告, NL128-10, pp67–72, Nov. 1998.

[12] 田中貴秋, 松尾義博, "対訳関係のないコーパスからの複合名詞の対訳の獲得", 言処学会, 第 5 回年次大会, pp29–32,March 1999.

[13] 北研二, 中村哲, 永田昌昭 共著, "音声言語処理" 電子情報通信学会, 1996.

[14] E. Brill. "Some advances in rule-based part of speech tagging." Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa., 1994.

[15] 松本裕治, 北村啓, 山下達雄, 今一修, 今村友明. "日本語形態素解析システム『茶筅』version2.0 使用説明書." Technical Report NAIST-IS-TR99008, 奈良先端科学技術大学院大学,1999.

[16] 久保正治, 英和・和英電索辞典 gene, 技術評論社,1995.

[17] 笹岡久行, 荒木健治, 桃内佳雄, 栃内香次, "帰納的学習を用いた訳語推定手法における単語片対の抽出元の選択数に関する性能評価", 言処学会, 第 5 回年次大会, pp357–360,March 1999.

[18] Hisayuki Sasaoka, Kenji Araki, Yoshio Momouchi and Koji Tochinai "Constraint on Sources of Units for Prediction Method of Target Word Using Inductive Learning", In *Proceedings of PACLING 1999*, pp235–243, Aug. 1999.