言い換えの統一的モデル — 尺度に基づく変形の利用 —

村田 真樹[†] 井佐原 均[†]

言い換えに関する研究は平易文生成，要約，質問応答と多岐の分野において重要なものであるが，本稿では言い換えの統一的モデルとして，尺度に基づく変形による手法を示し，このモデルによって種々の言い換えを統一的に扱えることを示す．このモデルでは，多様な言い換えの問題の違いを，尺度で表現することで，多様な言い換えを統一的に扱えるようになっている．本稿では具体的にこのモデルで，文内圧縮システム，推敲システム，文章語口語変換システム，RL 発音回避システム，質問応答システムを構築できることを示す．本稿の言い換えの統一的モデルは，システムの作成を効率的にしたり，言い換える原理を容易に理解させたり，多様な新たな言い換えを思いつかせる効果があり，有益なものである．

キーワード: 言い換え, 要約, 推敲, 口語文章語変換システム, *RL* 発音回避システム, 質問応答システム

Universal Model for Paraphrasing — Using Transformation Based on a Defined Criteria —

MASAKI MURATA[†] and HITOSHI ISAHARA[†]

Studies on paraphrasing are important in various research topics such as sentence generation, summarization, and question-answering. A universal model is described for paraphrasing that transforms according to defined criteria. We show that by using different criteria, we can construct different kinds of paraphrasing systems including one for compressing sentences, one for polishing the sentences up, one for transforming written language into spoken language, one for transforming English words into synonyms with the same meaning containing less “l” and “r” letters, and one for answering questions. Our model efficiently constructs systems and produces dynamic paraphrasing systems. It should prompt the creation of new paraphrasing systems in the future.

KeyWords: *Paraphrasing, Summarization, Proofreading, Transformation system for written-language-to-spoken-language, System transforming English words into synonyms containing less “l” and “r” letters, Question-answering system*

1 はじめに

言い換えに関する研究 (佐藤 1999; 山本 2001; Murata and Isahara 2001; 乾 2002) は平易文生成，要約，質問応答 (村田, 内山, 井佐原 2000; 村田 2003) と多岐の分野において重要なも

[†] 独立行政法人情報通信研究機構, National Institute of Information and Communications Technology, Independent Administrative Institution

のであるが、本稿では言い換えの研究の統一的モデルとして、尺度に基づく変形による手法を示し¹、このモデルによって種々の言い換えを統一的に扱えることを示す。このモデルでは、多様な言い換えの問題の違いを、尺度で表現することで、言い換えを統一的に扱えるようになっている。

このモデルには以下の利点が存在する。

- システム作成の効率化

本稿の言い換えの統一的モデルでは、変形の尺度や変形規則を他のものに取り替えるだけで多様な言い換えを実現することができる。システム作成では、変形の尺度や変形規則以外のモジュールは一度作成してしまえば、多様な言い換えシステムで利用することができる。すなわち、システム作成のコストを軽減できるのである。また、変形規則も複数の言い換えシステムで共用できる場合があり、その場合もシステム作成のコストを軽減できる。

- 言い換えの原理の理解容易性

本稿の言い換えの統一的モデルでは、後で述べるように変形部と評価部という二つの構成要素からなる単純なモデルだけで、多種多様な言い換えを扱うことができるようになっている。本稿のモデルは単純で理解しやすく、大雑把に言い換えをどのようにすればできるかを考えるには、このモデルを基本におくと考えやすい。

- 多様な言い換への創出

本稿の言い換えの統一的モデルでは、変形の尺度を変更することで、多様な言い換えを実現することができる。すなわち、尺度のみを考察し、新たな尺度を考えたときには、その尺度で変形を行なう新たな言い換えシステムを考えたことと等価になる。尺度のみを考察し、新たな尺度を考案することは比較的容易であるので、本稿の統一的モデルは、多様な新たな言い換をを思いつくことにも役に立つのである。

本稿ではまず、上述のような優れた利点を持つ言い換の統一的モデルについて説明する。その後で、この統一的モデルに基づいて試作した言い換システムを紹介する。紹介する言い換システムは、文内圧縮システム、推敲システム、文章語口語変換システム、RL 発音回避システム、質問応答システムである。これら多様なシステムを本稿の統一的モデルで具体的に作成できることを示すことで、本稿の統一的モデルで実際に多様な言い換の問題を扱えることを示す。

¹ 本稿は、文献(村田, 井佐原 2001a)に基づいて作成したものである。本研究の主眼になっている尺度に基づく変形については、文献(村田他 2000)の脚注6においても述べている。

図 1 言い換えの統一的モデルの模式図

2 言い換えの統一的モデル

本稿で記述する言い換えの統一的モデルは、図1の構成をしている。このモデルは、変形部 (transformation module) と評価部 (evaluation module) の二つのモジュールからなる。変形したいものが現れたときは、それを図のようにシステムに入力して、変形部で変形の候補をあげ、評価部において変形の妥当性をチェックし最も妥当であると判断されたものに変形され、それが図のように出力される。

- 変形部

変形の候補を与えるモジュールである。変形部は、人手による規則で構成してもよいし、計算機で自動獲得した規則で構成してもよいし、動的に書き換え候補を生成するものでもよいし、これらの組み合わせでもよい。

- 評価部

変形の候補の良さを、あらかじめ定めておいた尺度により評価し、最もふさわしい変形の候補を選択するモジュールである。ここで定める尺度は、扱う問題ごとに適正なものに作りかえる必要がある。

評価部で用いられる尺度の具体例として以下のものが考えられる。

- 類似度

例えば、A と B の類似度を調べたいとする。このとき、変形部の規則がすべて同義性を満足するものとする。この場合、A と B の類似度が大きくなるように、変形部の規則で A,B を変形し、A,B が良く似た状態にしてから類似度を求めると、意味が同じなのに異なる表現で記述されているような場合でも正しく類似度を計算することができる。

- 長さ

例えば、要約の一つの分野の文内圧縮のように、なるべく意味を変えずに文を圧縮したいとする。このとき、変形部の規則はすべて意味をほとんど変えずに変形するものとする。この場合、長さを尺度とし、この長さが短くなるように変形を繰り返すと文

内圧縮が実現される。

- 頻度 (または, 生起確率)

例えば, 推敲システムを考える。このとき, 変形部の規則がすべて同義性を満足するものとする。この場合, 推敲したいデータを, そのデータの生起確率が高くなるように変形すると非常に洗練された文章となる。もう少し簡単な例でこれを説明すると, 例えば, 入力したデータに「データー」とあったとしよう。また, 変形規則に「データー」を「データ」とする規則があったとしよう。このとき, 毎日新聞 (毎日新聞社 1998) などで「データー」と「データ」の数を数え, 「データ」の方が数が多い場合, 「データー」を「データ」と直すといったことである。

また, 頻度 (または, 生起確率) を調べるコーパスを種々のものに変更することで, 様々な効果を生む。例えば, 入力データが書き言葉のときに, コーパスとして話し言葉を用いると書き言葉の話し言葉への変形が実現される (村田, 井佐原 2001c)。また, 入力データが法律関係の文のときに, コーパスとして平易な文章の集合を与えておくと, 法律関係の難解な文章が平易な文章に変形されることだろう。また, ここで入力データとして適当に誰かが書いた小説をいれて, コーパスとしてシェークスピアの小説をいれると, シェークスピアの文体の小説が新たに完成することになる。また, 入力データを芥川の小説として, コーパスとして漱石の小説を用いると, 芥川の小説を漱石の文体に変形するなどということもできるだろう。

- 文としての正当性のチェック

上記の生起確率に基づく尺度は, 推敲システムでも用いることができるように文の正当性のチェックに使うことができる²。しかし, 生起確率だと尺度として強すぎる場合は以下のような尺度を用いるとよい。

- 対象としている表現が, コーパスで1回以上出現しているか否か。(これは表記誤りの検出など (竹内, 松本 1999; Murata and Isahara 2002a) によく使われる尺度である。)
- コーパスでの生起確率がある程度以上か否か。
- コーパスでの生起確率が, 環境なしでの生起確率よりも大きいかな否か。

ここで示したものは, 尺度というよりは条件のようなもので, 他の尺度と組み合わせて用いるとよい。他の尺度のところで, もしその変形において, 文としての正当性が保証されない場合は, ここで示した尺度を同時に用いるとよい。

- 変形の前後での意味の等価性

変形規則が完全に同義性を満足するということがわからない場合は, この尺度が必要となる。ただし, この尺度の構築は現時点では難しいと思われる。

² 生のコーパスが文の正当性のチェック, すなわち, 生成の研究に使えることは, 生のコーパスが照応解析に使えることを含めて文献 (村田, 長尾 1998) に記述してある。

これができるようになるまでは、変形部で利用する変形規則を完全に同義性を満足するものだけにするか、同義性を満足しない言い換えをしてしまう可能性があることを覚悟するかのいずれかである。(とはいえ、変形部で利用する変形規則に同義性を満足しないものが少々ある状況で変形の前後での意味の等価性を調べる尺度を用いなかったとしても、上述の「文としての正当性のチェック」を用いれば多くの不適切な言い換えを取り除くことができるので、工学的見地ではある程度利用可能な言い換えシステムを構築できると思われる。)

また、この項目の尺度も一つ上の「文としての正当性のチェック」と同様に、尺度というよりは条件のようなもので、他の尺度とともに用いられる。

ここにあげたもの以外にも様々な尺度が考えられる。

英語文で R や L などを含む日本人にとって発音しにくい(垣田, 小篠 1983; 小池 1994) 単語をあまり使わないという尺度も考えられる。また、丁寧な表現もしくはわかりやすさの計量的研究が十分なされれば、それも尺度とすることで丁寧な表現もしくはわかりやすい表現への自動言い換えが可能となるだろう。ただし、これは丁寧な表現もしくはわかりやすい表現のみを使ったコーパスを生起確率の算出に用いることで、先の生起確率の尺度でも扱えることである。

また、条件のような尺度には、「21 世紀」など特定の語を使うことを条件として言い換えることや、起承転結を満足する文章構成を条件として言い換えることや係り先未決定文節数を 7 程度以下とすること(村田, 内元, 馬, 井佐原 1999)を条件として言い換えることなど、様々なものが想定できる。

以降では、われわれが行なっている研究を具体的な事例として、この統一的モデルのもとでの変形操作がどのような尺度によってなされているかを、見てみよう。

3 文内圧縮システムの場合

最近是要約の研究(加藤, 浦谷 1999)が盛んになっているが、ここでは要約の一分野である文内圧縮を試みてみよう。

変形規則としては、文献(村田, 井佐原 2001b)の 3 節の研究で自動獲得した規則のうち、その文献の評価式(5)でソートした結果を上位から見て頻度が 1 の規則が現れる一つ手前までの規則を利用する。このとき規則の総数は 775 個となった。本節ではこれらを変形部の規則とする。表 1 に規則の例を示す。「 ϕ 」は空文字を意味する。この文献(村田, 井佐原 2001b; 村田, 金丸, 井佐原 2004)での研究では、同義な意味を持つ、複数の辞書の同じ項目の定義文を照合することで、ほぼ同義な表現の対を抽出している。この表現の対を変形規則に利用するのである。このシステムではこの得られた変形規則は双方向書き換え可能として利用する。

ここでは新聞記事の要約を考えることとして、評価部の尺度としては以下のものを用いることにする。

表 1 変形規則の例

・	⇔	φ
、	⇔	φ
の	⇔	が
や	⇔	・
など	⇔	φ
いう	⇔	言う
と	⇔	・
φ	⇔	ための
用いる	⇔	使う
入る	⇔	はいる
くる	⇔	来る
・	⇔	または
または	⇔	や

表 2 文内圧縮の例

良いと思われるもの
九日からの韓国訪問では昨年五月、盧泰愚（ノ・テウ）大統領来日時に合意した歴史の流れの中で解決されるよう勇気ある決断を望む米、イラクの直接対話実現に強い期待を示した。アジア・太平洋地域にも及ぶよう外交努力をしてきた。多国籍軍には十億ドルの追加的措置をとった段階だ。
良くないと思われるもの
ソ連の経済危機は天災ではなく、指導部の場当たりの対応に主要な要因がある。自由と民主主義と市場経済を求め、私たちと同じ政治経済の仕組みに向かって努力している。前村長、菊地豊氏（58）を党推薦の無所属候補として擁立することを決めた。

- 入力されたデータがより短くなるような変形を良いものとする。
- 新聞記事での出現が1個以上あることを条件とする。（文としての適切性の判定）

ここでの新聞記事は94年と95年の毎日新聞2年分とした。

厳密には以下のアルゴリズムによって実行した。

- (1) 入力として与えられるデータをJUMAN(黒橋, 長尾 1998)で形態素解析して, 形態素列に分解する。
- (2) 文頭の形態素から順に, 形態素ごとに以下の処理を行なう。

現在の形態素で始まる形態素列 S (形態素を一つも持たない場合, つまり空文字も含む) と, 変形規則 R_i の前件部の文字列 A_i が一致した場合, その後件部の文字列 B_i が, 書き換え後表現の候補となる。また, S の前接 k -gram の形態素列を $S1_i$, S の後節 k -gram の形態素列を $S2_i$ とする。

各書き換え後表現の候補 B_i に対して, 文字列 A_i から文字列 B_i になるときに短縮される文字数を数え, この値が最も大きかったときの i を m と

する.

評価に用いるコーパスにおいて $S1_mB_iS2_m$ の文字列の頻度を求め, この値が 1 より大きいとき, A_m を B_m に書き換え, 処理を次の形態素に移す.

ただし, k は定数である.

ここでは, 少々再現率を下げてでもよいから適合率を高める意味で k としては 2 を用いておこう.

この条件で毎日新聞の 91 年のデータの文内圧縮の実験を試みた. その結果の一例を表 2 にあげておく. 表で下線部は変形操作により消される部分を意味する. 本節の研究は圧縮ということで文を短くする操作があるため, 変形規則としては消去するパターンのものが用いられていると思われる. 「強い」や「的」などのものが消去されて正しく圧縮できたものから, 「は」や「と」を消去して意味が変わってしまう誤りもあった. また, 表の最後のデータは, 「すること」を消去したが「候補として擁立を決めた」と短い範囲で見ると正しそうに見えるが, もう少し前からみると誰それをとという部分があり「すること」を消去してはいけないとわかる. この誤りを正すには構文的な情報を反映した評価式³を用いる必要がある.

4 推敲システムの場合

ここでは推敲システムについて考えてみる. 変形規則は前節と同じ 775 個のものをを用いる. このように変形規則は複数の言い換えシステムで利用できる場合があるのである.

ここでは新聞記事の推敲を考えることとして, 評価部の尺度としては以下のものをを用いることにする.

- 入力されたデータの各部分形態素列の新聞記事コーパスでの生起確率が, より大きくなるような変形を良いものとする.

ここでの新聞記事は 94 年と 95 年の毎日新聞 2 年分とした.

ここでは, 新聞記事の推敲を考えるために生起確率をもとめるコーパスとして新聞記事コーパスを利用する. もし, 論文を推敲したいときは論文集合のコーパスを使えばよい. すなわち, 推敲したい文書と同一の種類のコーパスを利用するのである.

厳密には以下のアルゴリズムによって実行した.

- (1) 入力として与えられるデータを JUMAN で形態素解析して, 形態素列に分解する.
- (2) 文頭の形態素から順に, 形態素ごとに以下の処理を行なう.

現在の形態素で始まる形態素列 S (形態素を一つも持たない場合, つまり空文字も含む) と, 変形規則 R_i の前件部の文字列 A_i が一致した場合, その後件部の文字列 B_i が, 書き換え後表現の候補となる. また, S の前後 k -gram の形態素列を $S1_i$, S の後節 k -gram の形態素列を $S2_i$ とする.

³ この評価式としては, 例えば, 係り先がなくなる文節が生じてはいけないという条件のような尺度を利用するとよい.

表 3 推敲結果の例

良いと思われるもの
社会党の土井委員長は三十一日、神戸市の実家に帰り、三日まで実家と兵庫県西宮市の自宅で家族 と友人らと過ごす。 ソ連国民のうち「返還しなくてよい」は26%にとどまり、国民意識の面で両国間に接点がで きつつあるといえそうだ。 世界の平和と安定に貢献する上で、 支持率の回復につながったのではないか
良くないと思われるもの
日本の生活満足度は前回は8ポイント上回り、8割を超えたが、男女で差がある。 米国からイラクへの直接対話の申し入れなど、双方から平和的に解決したいという考えがにじみ出 ている。 国会移転に関する決議も、問題の解決につながる。ウマ年の昨年を象徴したのがオグリキャップ。 する

各書き換え後表現の候補 B_i に対して、新聞記事コーパスでの $S1_i B_i S2_i$ の文字列の頻度を求め、この頻度が最も大きかったときの i を m とする。新聞記事コーパスでの $S1_m A_m S2_m$ の文字列の頻度を求め、この値よりも、 $S1_m B_m S2_m$ の文字列の頻度の方が大きいとき、 A_m を B_m に書き換え、処理を次の形態素に移す。

ただし、 k は定数である。

ここでも前節と同じく、アルゴリズムでの頻度算出の環境を固定長の前後 2-gram($k=2$) としておこう⁴。

この条件で毎日新聞の 91 年のデータの推敲実験を試みた。その結果の一例を表 3 にあげておく。表の下線部は前節と同じく変形操作により消される部分を意味し、少し下に書いてある文字列はその文字列に変形されることを意味する。「や」「も」「と」「の」を補って読みやすくなったように思われるものや、逆に「も」を補ったり「で」を省略して意味が変わってしまい良くないと思われる結果があった。また、最後の行の例では、過去形の「した」を現在形の「する」に変形するというのがあったが、これは「昨年」の話で過去形であるべきで「した」を「する」にしてはいけない。これの対処としては変形規則の獲得精度をあげるか、文の妥当性の判定チェックに時制の情報も組み込むかする必要がある。

次に判定チェックに時制の情報も組み込んだ実験を行ってみた。つまり、評価部の尺度として、以下の条件も追加で利用した。

4 より良い解析をするには、各文字列の頻度の部分は、その文字列を x とするとき、与えられた入力データを環境に持つときの x が新聞記事コーパスに出る事象の確率とするとよい。また、上記アルゴリズムは環境としては前後 k 形態素のものを固定で用いるものとなっているが、可変にしたり構文的な素性、時制的な素性など広範な情報を用いて、最大エントロピー法などの強力な確率推定手法により確率を求めるようにした方がよいだろう。

表 4 時制の情報も組み込んだ推敲結果の例

時制の情報に基づいて言い換えを行なわなかったことにより, 改善されたもの
ウマ年の昨年を象徴 $\begin{matrix} \bigcirc \text{した} \\ \times \text{する} \end{matrix}$ のがオグリキャップ。
「日本の根本的な政治改革」を要求 $\begin{matrix} \bigcirc \text{する} \\ \times \text{した} \end{matrix}$ が, 小選挙区制にすりかえさせてはならない。
ブッシュ政権の支持率は8月のイラクのクウェート侵攻後, いったん低下したといわれて $\begin{matrix} \bigcirc \text{いた} \\ \times \text{いる} \end{matrix}$
が, 今回の調査では高率の支持を集めた。
時制の情報に基づいて言い換えを行なわなくなったが, 時制の情報に反して言い換えを行なってもいいとも判断できるもの
「鈴木が脅迫状を作って $\begin{matrix} \bigcirc \text{いた} \\ \bigcirc \text{いる} \end{matrix}$ のを見た」などと供述していることから警視庁に照会、裏付け捜査を進めている。

- 入力されたデータの各部分の形態素列を言い換える時に, その形態素列の末尾の形態素が動詞の場合は, 時制が一致することを条件とする。

ここでは時制は JUMAN の出力で「タ形」であるものとそれ以外の場合の二種類の時制を利用した。アルゴリズムとしては (c) を以下のように変更した。

- (c) 新聞記事コーパスでの $S1_m A_m S2_m$ の文字列の頻度を求め, この値よりも, $S1_m B_m S2_m$ の文字列の頻度の方が大きく, なおかつ, 以下の条件を満足する時に A_m を B_m に書き換え, 処理を次の形態素に移す。

(条件) A_m の最後の形態素と B_m の最後の形態素がともに動詞である場合は両方ともが「タ形」か両方ともが「タ形」でない場合にこの条件を満足するとする。 A_m の最後の形態素と B_m の最後の形態素のどちらか一方でも動詞でない場合はこの条件を満足するとする。

この方法で先の実験と同じデータで新聞記事の推敲を行なってみた。先の時制が原因で誤った例(表3の最後の例)は「した」と「する」の言い換えを行なわなくなり, その部分の誤りは改善された。また同様の誤りの文も多く改善された。改善された文の例を表4に示す。表の言い換え箇所の上段は入力文での表現で, 下段は時制の制約を加えなければ出力される言い換えた表現である。また, それぞれ表現の左上の部分に, 正しい表現には○を意味が変わって言い換えとしては正しくない表現には×をつけている。この時制によるシステムの変更で出力が変わった10例をチェックしたところ, 1例だけそのように言い換えても文の意味が変わらないものであった。それ以外は時制の情報を使って言い換えを抑制する必要のある箇所であった。このことにより時制情報を利用することで言い換えシステムの性能を向上させる場合があることがわかった。

ここでは時制に関する誤りを対処するために, 時制情報に関係する制約を評価部に追加したが, これは変換規則から時制が変化する規則を取り除くことでも対処できる。

表 5 書き言葉から話し言葉への変形のための変形規則の例

、	→	φ
φ	→	の
・	→	φ
φ	→	え
φ	→	えー
の	→	φ
φ	→	を
を	→	φ
φ	→	で
φ	→	という
する。	→	いたします
対応する	→	対する

5 文章語口語変換システムの場合

ここでは書き言葉から話し言葉への言い換えを考えてみる (村田, 井佐原 2001c; Murata and Isahara 2002b).

変形規則としては, 文献 (村田, 井佐原 2001c) の研究で自動獲得した規則のうち, その文献の評価式 (4) でソートした結果を上位から見て頻度が 1 の規則が現れる一つ手前までの規則を利用する. このとき規則の総数は 240 個となった. 本節ではこれらを変形部の規則とする. 表 5 に規則の例を示す. 「φ」は空文字を意味する. この規則は双方向書き換え可能ではなく, 一方への書き換えのみ可能である. これは, この規則は書き言葉のテキストと話し言葉のテキストを照合して得られたもので, 獲得された変形規則の段階で方向性があり, 規則の左辺が書き言葉からまた右辺が話し言葉から得られた表現であるため, 書き言葉から話し言葉への言い換えの際には左辺から右辺への一方方向のみで書き換え可能となる.

次に評価に用いるコーパスであるが, ここでは話し言葉に変形したいので, 文献 (村田, 井佐原 2001c) で話し言葉データと呼んでいるものを使う. このデータは, 開放的融合研究推進制度, 話し言葉の言語的・パラ言語的構造の解明に基づく「話し言葉工学」の構築の一環として通信総合研究所と国立国語研究所で作成しているもので, 全国大会, 研究会などの学会講演データからなっている. 本稿ではそのうちの 82 編のものを用いている.

評価部の尺度としては以下のものを用いることにする.

- 入力されたデータの各部分形態素列の話し言葉データでの生起確率が, より大きくなるような変形を良いものとする.

実際に実行する厳密なアルゴリズムは前節のもの (時制情報を利用する改良をする前のもの) と等価である.

つまり, 本節の研究は, 前節の研究において新聞コーパスを用いていたところを話し言葉データに変えただけである. 前節の研究では, 入力も評価コーパスも新聞という同じものだった

表 6 書き言葉から話し言葉への変形例 (1gram の場合)

近年、パラフレーズに関する知識^え 獲得の研究が重要視されつつある。本^稿 研究^{研究} では、同義のテキストを照合し、その照合結果^をを用いてパラフレーズに関する知識を自動獲得することを試みた。この^{ような} 自動獲得の^{実験}を辞書定義文、新聞記事タイトル・本文対、講演テキストにおいて行なったところ、同義のテキストの照合による方法がパラフレーズ獲得にある程度役に立つ^{ことがわかった}。(中略)。そのシステム^{では}、基本的には、与えられた質問文^の答え^{という}が書いてありそうな文を探し出し、その答え^{です}が書いてありそうな文と質問文^の類似度^が大きくなるように双方を書き換えて照合し、答え^{です}が書いてありそうな文^での、質問文^の疑問詞に対応している箇所を答えとして出力するシステムである。

表 7 書き言葉から話し言葉への変形例 (2gram の場合)

これらは、同義な言い換えを示す、パラフレーズに関する知識がきちんとした形で整備されていない^{ことによる}。
 複数の国語辞典を用意してその定義を利用するというのが^ま考えられる。
 ある程度よさそうな同義・類義表現を抽出する^{という}ことを試みる。
 対応づけの誤りである場合もあり、同義表現としては^{あー}ふさわしくない対が多い。
 このパターンを頻度でソートした結果^{というの}を表に示す。

たため、新聞のデータをより新聞のデータらしくする、つまり、推敲の研究となっていたが、本節では入力を書き言葉、評価コーパスを話し言葉とするために、書き言葉から話し言葉への変形ということになる。

この条件で筆者の論文(村田, 井佐原 2001b)を入力として与え、話し言葉から書き言葉への変形の実験を行なった。前節のアルゴリズムの k が1のときの結果を表6に、 k が2のときの結果を表7に示す。表の下線部は変形操作により消される部分を意味し、少し下に書いてある文字列はその文字列に変形されることを意味する。もともとアルゴリズムが簡単なため、 k が1の場合では精度が悪いがそれでも、「え」をいれたり「本稿」を「本研究」と言い換えたりする結果が得られている。 k が2の場合では精度は良くほとんど誤りがなかった。「という」や「ま」や「あー」を入れていて、いかにも話し言葉にふさわしい表現になっている。しかし、変形箇所が少なく再現率が低いといった感じであった。

6 RL 発音回避システムの場合

ここでは、日本人にとって発音しにくい (垣田, 小篠 1983; 小池 1994) R や L を含む単語をあまり使わない英文に言い換えるシステムについて考えてみる. 本稿ではこのシステムのことを RL 発音回避システムと呼ぶ. 日本人が国際会議で英語で演説する際, このシステムにより R や L を含む単語をあまり使わない英文に言い換えておくと, R や L を含む単語の発音が苦手な日本人にとって話しやすい英語となる.

変形規則としては, WordNet 2.0 の名詞と動詞の同義語表現を利用した. 評価部の尺度としては以下のものを用いることにした.

- 英語文で発音しにくい R+母音や L+母音の表現を含む個数が小さいほどよいとする. (R と L は母音とくつつく場合が特に発音が難しい.)
- R+母音や L+母音の表現を含む個数が同じ場合は入力された元の表現の方がよいとし, また, 変形後の表現同士と比較では, 入力されたデータの各部分単語列の英語テキストでの生起確率がより大きくなるような変形を良いものとする.
- 英語テキストでの出現が 1 個以上があることを条件とする. (文としての適切性の判定)

英語テキストとしては BNC コーパス (Oxford University Computing Services 1995) を用いた. 本稿では, 母音の判定には文字を利用し, a,i,u,e,o,y を後ろにくつつけて持つ r,l の表現を, R+母音, L+母音の表現とした. また, 変形規則に用いる同義語表現には動詞の変化形, 名詞の複数形なども追加して用いた.

厳密には以下のアルゴリズムによって実行した.

- (1) 入力として与えられるデータをスペースで区切って単語列に分解する.
- (2) 文頭の単語から順に, 単語ごとに以下の処理を行なう.

現在の単語 S と, 変形規則 R_i の前件部の単語 A_i が一致した場合, その後件部の単語 B_i が, 書き換え後表現の候補となる. また, S の前接 k -gram の単語列を $S1_i$, S の後節 k -gram の単語列を $S2_i$ とする.

各書き換え後表現の候補 B_i に対して, B_i 中に R+母音と L+母音が含まれる頻度 $fb1$ と, 英語コーパスでの $S1_i B_i S2_i$ の単語列の頻度 $fb2$ を求め, $fb2$ が 1 以上のもので, $fb1$ の値が最も大きくその中で $fb2$ の値が最も大きかったときの i を m とする.

A_i 中に R+母音と L+母音が含まれる頻度を求め, この値よりも, B_m 中に R+母音と L+母音が含まれる頻度の方が大きいとき, A_m を B_m に書き換え, 処理を次の形態素に移す.

ただし, k は定数である.

本稿の言い換えの統一的モデルでは, 変形部と評価部を分割した構成になっており, 変形部の規則に WordNet 2.0 (Princeton University 2003) の同義語表現を利用し, 評価部で R+母音

表 8 RL 発音回避システムの例

良いと思われるもの				
We think a good	approach way	is to construct it using “X <i>no</i> Y”.		
The criteria used to	select determine	the most appropriate transformation type must be predefined.		
This figure shows the	structure composition	of the thesaurus.		
<i>length</i> d is the	length size	of a document d.		
良くないと思われるもの				
This is the	title name	of the query.		
P of d and t is the	location determination	of the first occurrence of a term <i>t</i> in the document <i>d</i> .		
This term is for weighting terms which are	followed used	by the Japanese-language particle “nado”.		

や L+母音が含まれる頻度や英語テキストでの出現頻度を利用することで、比較的容易に RL 発音回避システムを作成することができるのである。

この条件で筆者が今まで国際会議で口頭発表してきた発表の原稿を入力として与え、R や L をあまり含まない英文への変形の実験を行なった。ここでは、少々再現率を下げてでもよいから適合率を高める意味で k としては 2 を用いた。この実験の結果の一例を表 8 にあげておく。表の縦線で囲った部分が言い換えられた表現で、上の表現が下の表現に言い換えられている。それぞれ R+母音、L+母音が少なくなる表現に書き換えられている。“approach” を “way” に書き換えたり、“length” を “size” に書き換えたりして、発音しやすい語への言い換えが正しくできているものがあつた。しかし、今のところこのシステムでは、言い換えると微妙に意味が異なってしまう誤りもあつた。このシステムの今後の応用としては、今のところまだ性能が悪いので、言い換えた結果のみを出力するのではなく、言い換えの候補をいくつか尺度の値 (R+母音や L+母音の表現を含む個数など) とともにその値の順に提示し、そこでユーザに言い換えに適切な表現を選ばせるという支援システムのような形の利用が良いと思われる。

7 質問応答システムの場合

われわれの質問応答システム (村田他 2000) では、与えられた質問文の答えが書いてありそうな文を探し出し、その答えが書いてありそうな文と質問文の類似度が大きくなるように双方を書き換えて照合し、答えが書いてありそうな文での、質問文の疑問詞に対応している箇所を答えとして出力するといったことを行なう。例えば、表 9 のようなデータ (日本英語教育協会 1985) が与えられているときに「ニューヨーク州の中央部または北部に住む人たちの、最も一般的な職業は何ですか。」という質問があつたとしよう。このときこの質問文は疑問詞を X にし

表 9 データ

ふつう、日本人はアメリカ人がニューヨークに住んでいると聞くと、そのアメリカ人はニューヨーク市に住んでいるのだと思う。しかし、それはよくやる間違いで、ニューヨーク市というのは、ニューヨーク州の南の部分の大変せまい地域を占めているだけなのである。ニューヨーク市からナイアガラの滝まで車を運転して行くと約8時間かかり、そのナイアガラの施もニューヨーク州に存在している。ニューヨーク州の大部分は山や森林や原っぱ、川、湖、沼地などからできているのである。州のこの中央部や北部に住む人たちはふつう小さな町に住んでいる。そして、農業が、これらのニューヨーク州民の間では最も普通の職業で、この人たちの作る農作物で最も共通なものはトウモロコシである。

表 10 質問応答システムの例

類似度	種類	文
32.1	質問文	ニューヨーク州の中央部または北部に住む人たちの、最も一般的な職業はXです。
32.1	データ	そして、農業が、これらのニューヨーク州民の間では最も普通の職業で、この人たちの作る農作物で最も共通なものはトウモロコシである。
103.1	データ	そして、農業が、これらのニューヨーク州民の間では最も一般的な職業で、この人たちの作る農作物で最も共通なものはトウモロコシである。
82.5	データ	そして、農業が、これらのニューヨーク州の人たちの間では最も普通の職業で、この人たちの作る農作物で最も共通なものはトウモロコシである。
...	...	...
186.5	データ	そして、農業が、これらのニューヨーク州の人たちの間では最も一般的な職業で、この人たちの作る農作物で最も共通なものはトウモロコシである。
...	...	...
...	...	...
219.5	質問文	ニューヨーク州の中央部または北部に住む人たちの、最も一般的な職業はXである。
219.5	データ	、これらのニューヨーク州の人たちの間は最も一般的な職業は農業である。
	◆答え	= 農業
	◆補足	=、これらのニューヨーク州の人たちの間は最も一般的な職業は農業である。

表 11 質問応答システムで用いた変形規則の例

X が Y である	→	Y は X である
X の Y は Z で	→	X は Y は Z で
の間で	→	で
では	→	は
普通の	→	一般的な
州民	→	州の人たち
で、[]+\$	→	である。
である	→	です

て平叙文化され、またこの質問文と類似している文がデータから抽出され、表 10 の 1 行目の状態となる。表 11 にあげたような規則があったとすると、この規則を用いて質問文、データ双方を類似度が高くなるように書き換えていき、最終的に表のように類似度 219.5 に達して類似度がそれ以上高くなるなくなる。この状態で質問文とデータを照合すると、答えは「農業」と簡単にわかる。質問応答システムでは、類似度を尺度として言い換えを行なっていることになる。類似度が高くなるように言い換えを行なうことで答えとデータが照合しやすくなる。

ここに示したものは文献(村田他 2000)の予備実験として用いていたシステムのものをあげている。ここで用いた変形規則は人手で作成したものである。また、表 10 の最終状態のデータの日本語表現は文としておかしい。これの対処としては「文としての正当性のチェック」などの条件を尺度に組み込むとよいだろう。

本節の質問応答システムは、システムの処理内容の複雑さの都合上、本稿でいう言い換えるモデルだけでは実現できない。質問文において、疑問詞を X にして平叙文化するモジュールと、質問文とデータの類似度が向上するように言い換えるを繰り返させるモジュールが新たに必要である。しかし、質問文またはデータをデータまたは質問文に類似するように書き換える一回一回の言い換え処理は、変形規則に表 11 の規則を利用し、変形の尺度に質問文とデータの文の類似度を利用することで、本稿の言い換えるの統一モデルで扱えるものである。

ここでは、質問応答システムを例にあげたが、それに限らず類似度を計算したいときには類似度を尺度として言い換えを行なってやるとよい。例えば、情報検索などでも高度になると、上記の質問応答のようにクエリと検索される記事を類似度が高くなるように言い換えてからクエリと記事の類似度を求めるといったことも考えられる。また、照応の問題(村田, 長尾 1996)でも、「近くの大きな杉の木の根元にある穴」と「杉の木の根元の穴」の同一性もしくは包含関係が判定できないと照応を解決できないというのがあるが、類似度を尺度として両者を言い換え、例えば、「近くの大きな杉の木の根元の穴」と「杉の木の根元の穴」になった場合、後者が前者に含まれることになり後者が前者を指示可能とシステムに認識させてやることもできる。

8 関連研究との対比

本節では、本稿の言い換えモデルと他の関連研究、関連した考え方との比較を行ない、本稿の言い換えモデルと本稿のシステムの利点と欠点を考察する。

8.1 モデルの比較

本稿の統一モデル以外に考えられる一般的な言い換えるのモデルとして以下のものが考えられる。

- 言い換えの目的に応じて作成した変形規則を利用して、言い換えを実現する。(すなわち、評価部を用いず変形部のみの構成で言い換えを実現する。)

この構成の場合は変形規則にすでに言い換えの目的に応じた処理が組み込まれているため、本稿の統一的モデルにあるような言い換えの目的にあった表現を選択するための評価部が必要ないという特徴がある。

この変形部のみで構成される言い換えのモデルを利用していると考えられる研究の例としては、山崎らの言い換えによる要約の研究(山崎, 三上, 増山, 中川 1998), 佐藤の論文表題の言い換えの研究(佐藤 1999), 黒橋らの平易文への言い換えの研究(黒橋, 酒井, 鍛冶 2001)がある。山崎らの研究では、言い換え前と言い換え後の表現の対の表を作成し、この表を使って言い換えを実現している。佐藤の研究では、複数の複雑な処理と規則を組み合わせで論文表題の言い換えを実現している。黒橋らの研究では、国語辞典の定義文がその項目の見出し語を平易に表した表現であることを利用し、国語辞典の見出し語と定義文の対を変形規則のように扱って平易文への言い換えを実現している。

この変形部のみで構成される言い換えのモデルと、変形部と評価部を持つ本稿の言い換えモデルを比較してみる。この変形部のみで構成される言い換えのモデルには、評価部を必要としないという利点があるが、言い換えの問題ごとに変形部を構築する必要がある。これに対して、本稿の言い換えモデルでは、評価部を用意する必要がある代りに、言い換えの問題ごとに応じた変更を評価部のみとし、変形部を複数の言い換えシステムに利用することができる。このことによりシステム作成のコストを軽減することができる。また、このことは非常に優れた特徴であり、急遽ある特定の種類の言い換えシステムを作成する必要ができた場合、同義語・同義表現の辞書を持ってきてそれを変形規則として利用し、評価部の尺度にはその特定の種類の言い換えシステムの目的に応じたものを考えて使うだけで、その新しい特定の種類の言い換えシステムを作成できる。例えば、本稿ではRL 発音回避システムを紹介しているが、このシステムの実装は本稿のモデルに従えば極めて容易であり、英語単語の同義語辞書を変形部に利用し、発音のしにくいRやLの表現の出現頻度を評価部の尺度に利用することでこのシステムを作成できたのである。このように、本稿の統一的モデルは、変形部を複数の言い換えシステムに利用することができる特徴の他に、概ね変形部の変形規則には同義語・同義表現の辞書を利用し、評価部の尺度にはその問題に応じた尺度を考えればよいという指針まで存在するため、システム実装が容易というだけでなく、システム設計も容易という特徴がある。このことはさらに、評価部の尺度について問題に応じて新しいものを考えれば、その尺度に基づく新しい言い換えシステムを設計したことになるということにつながる。新たな評価部の尺度だけを考案することは比較的容易なので、多くの種類の新しい評価部の尺度を考えることで多様な新たな言い換えシステムを考えることができる。本稿の統一的モデルには、変形部のみで構成される言い換えのモデルに比べて上記のような強力な利点がある。

話を少し戻して先にあげた三つの研究を本稿の統一的モデルで実現するにはどのようにすればよいかを考えてみる。山崎らのニュース文要約の研究(山崎他 1998)では、研究の基本方針として、「冗長な表現を短縮し、略語があれば略語に言い換える、無くても理解できる部分は削除する、強調表現や文の接続を表す語句も削除する」というものが記述されてある。本稿の統一的モデルだと、変形部の変形規則に略語とその略語の略さない表現の対を追加して、強調表現や接続表現に関する変形規則などを追加し、尺度に本稿でも用いた文の長さが短いほどよいというものを使うとよい。佐藤らの論文表題をわかりやすく言い換える研究(佐藤 1999)の場合は、語尾の動詞性名詞を格関係も考慮して動詞化する規則を評価部の変形規則に利用して、尺度に文末は動詞がよいというものや、平易な文章を集めてその平易な文章での出現頻度が多い表現ほどよいというものを使うとよい。また、わかりやすくした結果の論文表題も多数収集されれば、そこでの頻度が多い表現ほどよいという尺度も利用できる。黒橋らの平易文への言い換えの研究(黒橋他 2001)の場合では、国語辞書から取り出した見出し語と定義文の対から変形規則を作成しそれを変形部の変形規則に追加して、尺度には平易な文章を集めてその平易な文章での出現頻度が多い表現ほどよいというものを使うとよい。このように、これらの三つの研究は本稿の統一的モデルでも扱うことができる。

なお、変形部と評価部を分離した構成を持つ構成・方法については、いくつかの文献(山本 2001; 乾 2002; 藤田, 乾 2003)でも述べられている。例えば、藤田らの論文(藤田, 乾 2003)では言い換えレイヤと目的適合性レイヤを分離して扱う方法が述べ、このことによりテキストの評価基準(本稿でいうところの尺度)を取り替えることで様々な用途の違いを吸収でき、汎用的な枠組みを提供するということが述べられている。山本や乾の文献(山本 2001; 乾 2002)には評価部の尺度を考える際に役に立つ、言い換えるの基準や考え方(文献では換言因子とも呼ばれている)が数多く記載されている。役に立つ貴重なものと思われる。しかし、これらの文献では本稿で紹介しているような、実際にテキストの評価基準を取り替えて様々な言い換えシステムを構築してその動作例を示すというようなことはしていない。その意味では本稿は評価部の尺度を種々のものに取り替えることで、様々な言い換えシステムを構築できることを実際に示しており、その面で価値がある。

8.2 モデルの個々の要素の検討

本節で本稿のモデルの個々の要素である、変形部と評価部について議論する。

まず、変形部についてであるが、本稿で実際に示したシステムでは、変形部には主に Perl の `s` コマンドでも扱えるような、文字列の変換のみしか扱わなかった。文内圧縮システムでも指摘したが、構文的な情報を反映した変形をしないと正しく言い換えが実現できない場合がある。今後はこのような構文的な情報を反映した変形を実現する必要がある。乾らのグループが開発している言い換えエンジン KURA(高橋, 岩倉, 飯田, 乾 2001)は、構文構造を持つ表現の言

い換えを実現できるもので、その有用性は大きいと考える。今後は、KURA を利用した研究やシステム開発も行なってみたいと考えている。KURA は修正・棄却規則を具備し、不適切な言い換えが生じることを防ぐ機能も有している。

次に評価部について議論する。言い換えは「場面」「話し手(または書き手)」「聞き手(または読み手)」などの状況によって変わるもので、それぞれに応じた言い換えを実現する必要がある。ここではそれらの状況を本稿の評価部で扱うことができるかを考えてみたい。

「場面」「話し手」「聞き手」の情報は、それぞれ本稿のモデルの評価部で扱うことができる。例えば、「場面」がニュース報道であり、ニュース報道的な文章に言い換えたい場合は、ニュース報道のテキストを評価部のコーパスとして用い、ニュース報道のテキストによく現れる表現に変更することで、ニュース報道的な文章に言い換えることができる。また、「場面」が悲しい状態であり、悲しい状態に適した文章に言い換えたい場合は、悲しい状態のテキストを収集しそれを評価部のコーパスとして用い、悲しい状態のテキストによく現れる表現に変更することで、悲しい状態に適した文章に言い換えることができる。また、「話し手」または「書き手」を次に考えると、2節にも書いていたようにシェークスピアや漱石のテキストを評価部に利用することで、シェークスピアや漱石の文体に変更することができると思われる。また、ニュースのアナウサーのような「話し手」の文章に変更したい場合は、ニュースのアナウサーのテキストを集めてそれをコーパスとして利用することでニュースのアナウサーのような文章に言い換えることができる。また、「聞き手」または「読み手」が小学生の場合、小学生向けのテキストを集めそれを評価部に利用することで、小学生向けの文章に言い換えることができる。また、「聞き手」または「読み手」がある特定の分野の専門家で、その特定の分野の文章に言い換えた方がその専門家にとって読みやすくそのように言い換えたい場合は、その特定の分野の文章を集めてそれを評価部に利用することで、その特定の分野の文章に言い換えることができる。また、「話し手」(または「書き手」)と「聞き手」(または「書き手」)の間の関係が上下関係で敬語表現などを使う必要がある場合も、その「話し手」と「書き手」の間の関係と同じテキストを集めてそれを評価部に用いることで、その関係にふさわしい文章に言い換えることができる。上述のように本稿のモデルでは、それぞれに適したテキスト集合を収集してそれを評価部に用いることで、「場面」「話し手」「聞き手」の情報を比較的簡便に扱うことができるのである。

ここでは、主にテキストコーパスでの頻度を尺度とする方法を示してきたが、この方法の他に、評価部の尺度としては、心理実験や計量的な研究をして、単純な頻度ではない尺度も構築することができると、それを評価部の尺度として用いることもできる。

9 おわりに

本稿では言い換えの統一的モデルとして、尺度に基づく変形による手法を記述した。また、様々な尺度を設定することで、文内圧縮システム、推敲システム、文章語口語変換システム、

RL 発音回避システム, 質問応答システムといった多様なシステムを構築できることを具体的に示した。

本稿の言い換えの統一的モデルでは, 変形の尺度や変形規則を他のものに取り替えるだけで多様な言い換えを実現することができるので, 尺度や変形規則以外の部分を複数の言い換えシステムで利用することができ, システム作成のコストを軽減する効果がある。また, 本稿の言い換えの統一的モデルは, 変形部と評価部という二つの構成要素からなる単純なモデルだけで, 多種多様な言い換えを扱えるようになっているため, 本稿のモデルは単純で理解しやすく, 大雑把に言い換えるの原理を考察するには役に立つモデルである。また, 新たな尺度を考えたときには, その尺度で変形を行なう新たな言い換えシステムを考えたことと等価になるため, 多くの新たな尺度を考案することで多様な新たな言い換えシステムを思いつくことにも役に立つのである。

われわれは本稿で示した言い換えの統一的モデルを多くの人に知ってもらって, 効率よく多くの言い換えの研究がなされることを切に希望する。

謝辞

独立行政法人情報通信研究機構の和泉絵美氏には 6 章の研究に対して有益なコメントと手助けをしていただきました。ここに感謝いたします。また, 言い換えに関する研究の創設と発展および本特集号に尽力されている, 京都大学佐藤理史助教授と奈良先端科学技術大学院大学乾健太郎助教授をはじめとする方々に感謝いたします。

参考文献

- 藤田篤, 乾健太郎 (2003). “語彙・構文的言い換えにおける変換誤りの分析.” 情報処理学会論文誌, **44** (11).
- 乾健太郎 (2002). “言語表現を言い換える技術.” 言語処理学会第 8 回年次大会チュートリアル資料, pp. 1-21.
- 垣田直巳, 小篠敏明 (1983). 英語の誤答分析. 大修館書店.
- 加藤直人, 浦谷則好 (1999). “局所的要約知識の自動獲得手法.” 言語処理学会誌, **6** (7).
- 小池生夫 (1994). 第二言語習得研究に基づく最新の英語教育. 大修館書店.
- 黒橋禎夫, 長尾真 (1998). 日本語形態素解析システム JUMAN 使用説明書 version 3.6. 京都大学大学院工学研究科.
- 黒橋禎夫, 酒井康行, 鍛冶伸裕 (2001). “国語辞典に基づく文章理解とパラフレーズ.” 言語処理学会第 7 回年次大会ワークショップ論文集.
- 毎日新聞社 (1998). “毎日新聞 1991-1998.”.

- 村田真樹, 長尾真 (1996). “名詞の指示性を利用した日本語文章における名詞の指示対象の推定.” 言語処理学会誌, **3** (1).
- 村田真樹, 長尾真 (1998). “表層表現と用例を用いた照応省略解析手法.” 言語理解とコミュニケーション研究会 NLC97-56.
- 村田真樹, 内元清貴, 馬青, 井佐原均 (1999). “日本語文と英語文における統語構造認識とマジカルナンバー 7 ± 2 .” 言語処理学会誌, **6** (7), pp. 61-71.
- 村田真樹, 内山将夫, 井佐原均 (2000). “類似度に基づく推論を用いた質問応答システム.” 自然言語処理研究会 2000-NL-135, pp. 181-188.
- 村田真樹, 井佐原均 (2001a). “言い換えの統一的モデル — 尺度に基づく変形の利用 —.” 言語処理学会第7回年次大会ワークショップ論文集.
- 村田真樹, 井佐原均 (2001b). “同義テキストの照合に基づくパラフレーズに関する知識の自動獲得.” 情報処理学会 自然言語処理研究会 2001-NL-142.
- 村田真樹, 井佐原均 (2001c). “話し言葉と書き言葉の diff.” ワークショップ「話し言葉の科学と工学」.
- 村田真樹 (2003). “質問応答システムの現状と展望.” 電子情報通信学会学会誌, **86** (12), pp. 959-963.
- 村田真樹, 金丸敏幸, 井佐原均 (2004). “複数の辞書の定義文の照合に基づく同義表現の自動獲得.” 言語処理学会誌, **11** (5).
- Murata, M. and Isahara, H. (2001). “Universal Model for Paraphrasing — Using Transformation Based on a Defined Criteria —.” In *NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications*.
- Murata, M. and Isahara, H. (2002a). “Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples.” *IEICE Transactions on Information and Systems*, **E85-D** (9), pp. 1416-1424.
- Murata, M. and Isahara, H. (2002b). “Automatic Extraction of Differences between Spoken and Written Languages, and Automatic Translation from the Written to the Spoken Language.” In *LERC 2002*.
- 日本英語教育協会 (1985). 20 日完成 英検 2 級二次試験対策 (面接テスト).
- Oxford University Computing Services (1995). “British National Corpus.”.
- Princeton University (2003). “WordNet 2.0.”.
- 佐藤理史 (1999). “論文表題を言い換える.” 情報処理学会論文誌, **40** (7).
- 高橋哲朗, 岩倉友哉, 飯田龍, 乾健太郎 (2001). “KURA: 統一的かつ宣言的記述法に基づく言い換え知識の開発環境.” 電子情報通信学会 思考と言語研究会 TL01-12.
- 竹内孔一, 松本裕治 (1999). “統計的言語モデルを用いた OCR 誤り修正システムの構築.” 情報

処理学会論文誌, 40 (6).

山本和英 (2001). “換言処理の現状と課題.” 言語処理学会第7回年次大会ワークショップ論文集.

山崎邦子, 三上真, 増山繁, 中川聖一 (1998). “聴覚障害者用字幕生成のための言い替えによる
ニュース文要約.” 言語処理学会 第4回年次大会.

略歴

村田 真樹: 1993年京都大学工学部卒業. 1995年同大学院修士課程修了. 1997年同大学院博士課程修了, 博士(工学). 同年, 京都大学にて日本学術振興会リサーチ・アソシエイト. 1998年郵政省通信総合研究所入所. 現在, 独立行政法人情報通信研究機構主任研究員. 自然言語処理, 機械翻訳, 情報検索, 質問応答システムの研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 電子情報通信学会, 計量国語学会, ACL, 各会員.

井佐原 均: 1978年京都大学工学部電気工学第二学科卒業. 1980年同大学院修士課程修了. 博士(工学). 同年通商産業省電子技術総合研究所入所. 1995年郵政省通信総合研究所. 現在, 独立行政法人情報通信研究機構けいはんな情報通信融合研究センター自然言語グループリーダー. 自然言語処理, 機械翻訳の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 日本認知科学会, ACL, 各会員.

(2004年1月8日 受付)

(2004年4月30日 再受付)

(2004年7月5日 採録)