社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－DBS－140（Ⅱ）（85）
2006／7／14

安全語のアンマスキングによる機密情報マスキングシステム

伊川 洋平，宅間 大介，金山 博
日本アイ・ビー・エム株式会社　東京基礎研究所
242-8502 神奈川県大和市下鶴間 1623-14
{yikawa, ta9ma, hkana}@jp.ibm.com

概 要

個人情報の保護，ビジネスの機密保持などの観点から，テキストデータに含まれる人名，地名，組織名などの固有名詞を伏せ字や一般語に置換する，マスキング技術が重要視されている．それに伴い，膨大なテキストデータをできるだけ人的コストをかけずにマスキングする手法に対する要求が高まっているが，従来のテキスト解析技術ではマスキングすべき表現の検出に失敗してしまうケースも多く，確実性の面で問題があった．そこで本論文では，全ての語がマスキングされている状態から，人手によって安全な語を選別してマスキングの解除を行うことで機密文書のマスキングを行う手法を提案し，実用性について検討する．

A Masking System for Confidential Documents by Unmasking Safe Words

IKAWA Yohei, TAKUMA Daisuke, KANAYAMA Hiroshi
Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
{yikawa, ta9ma, hkana}@jp.ibm.com

Abstract

In order to protect personal and confidential information such as human, place and organization names, document masking techniques are becoming important. Existing automatic masking techniques are not reliable enough since they may fail to mask out-of-vocabulary proper nouns. In this paper we propose a novel technique for document masking, an unmasking method, in which all of the words are hidden initially and a human specifies the non-confidential words to be unmasked. Our experimental results show its safety and effectiveness.

1　はじめに

個人情報の保護，ビジネスの機密保持などの観点から，機密文書に含まれる個人や企業を特定する情報を除去するマスキング技術が重要視されている．これらの情報を除去することで，機密文書を無害化して組織内で情報を共有することが可能となり，業務生産性を向上させることができる．

従来のマスキング手法では，固有名詞や，住所・電話番号のフォーマットと一致する表現などを，予め定義された辞書や規則に基づいて抽出・マスキング

を行っている．ユーザーによって辞書や規則を追加する機能はあるものの，未登録の表現を漏れなく自動検知するには至っておらず，確実性の点で問題がある．

本手法は，文書中にある固有情報を伏せ字や一般語に置換する，いわゆるマスキングによって，固有情報を混入させることなく，他者が文書の内容を閲覧できるようにする処理を行うためのものである．固有情報の典型的な例として，個人情報（氏名・住所・電話番号など，個人を特定できる情報）や企業情報（会社名・組織名・担当者名・資本金などの企業を特定できる情報）があり，これらをマスキングしたり，匿名表現や一般表現に置換した文書に変換したりすることが求められる．

この作業には確実性が要求されるため，人手で行われることが多いのが現状であるが，安全な情報共有を目的として，組織内に蓄積された膨大なテキストデータをなるべく少ない作業量でマスキングを行う手法に対する要求が高まっている．これを達成するために，固有情報である可能性がある部分をシステムが可能な限り自動判別してマスキングすることが考えられるが，静的に定義された辞書や言語的パターンのみに基づいて処理を行おうとすると，新しいパターンや未知の固有名詞などを自動検出することができない．また，テキストデータには書き誤り等のノイズが多く含まれていることもあり，テキスト解析技術では解析に失敗してしまうケースも多い．

また，マスキングすべき表現は，マスキングの目的や状況に応じて大きく変化する．マスキング後も高い情報量を維持するためには，静的な辞書を用意して固有情報を全てマスキングしてしまうのではなく，人間が目的や状況に応じてマスキングすべき固有情報かどうかを適切に判断する必要がある．

そこで本論文では，全ての語がマスキングされている状態から，人手によって安全な語を選別してマスキングの解除を行う，アンマスキング法を提案する．実験では，提案手法によるマスキングを行い，少ない作業量で目的や状況に応じて人間がマスキングすべき語を選別でき，かつ安全性の高いマスキング手法であることを示す．

本論文は以下のように構成する．まず，2 節で従来手法について説明し，3 節では提案手法について述べる．4 節では提案手法を用いた実験結果を報告し，5 節ではまとめと今後の課題について述べる．

2　従来手法

従来の機密情報マスキング手法は，隠蔽すべき情報を自動抽出することでマスキングを行う．機密情報を無害化するためにマスキングすべき表現は，固有表現 (Named Entity) としてこれまでにさまざまな定義がなされている．

2.1　固有表現

固有表現とは，特定の個人や組織を表す語句や，それらを類推できるような表現のことである．例えば，1999 年に開催された IREX(Information Retrieval and Extraction Exercise) [1] における固有表現抽出のコンテスト (IREX-NE) では，以下の 8 つの固有表現が定義されている．

- 組織名
- 人名
- 地名
- 固有物名
- 日付表現
- 時間表現
- 金額表現
- 割合表現

固有表現抽出は非構造データからの情報抽出の目的で重要視されてきたが，これらの固有表現をマスキングすることによって機密情報を無害化することができる．

[4] によれば，IREX-NE において精度の高かった上位システムでは，新聞記事に含まれている組織名のうち約 80% を抽出することに成功しており，テキストからの情報抽出の観点では実用的な精度といえる．しかし，残り 20% の組織名を抽出できていないため，機密情報のマスキングに適用するのは問題があることが分かる．

2.2　固有表現抽出によるマスキングの問題点

1. 抽出精度が十分ではない
   辞書登録されていない固有表現は基本的に抽出

することができず，マスキング漏れが発生する．特定の接頭語・接尾語を伴って固有表現が現れやすい（「さん」の前には人名が来る，など）といったルールによるリカバリを行っているものの，完全ではない．情報抽出の観点では，ある程度の抽出漏れは問題にならないことが多いが，機密情報のマスキングの際には，わずかな抽出漏れが機密情報の漏洩の原因となる恐れがある．

2. 目的や状況に応じてマスキングすべき表現が変化する
   マスキングすべき固有表現は，マスキングの目的や状況に応じて大きく変化する．例えば，「IBM」は固有表現だが，IBM の社内文書をマスキングする際には隠蔽しなければならない情報ではなく，マスキングせずに残しておくべき情報と考えられる．よって，静的な辞書を用意して固有表現を全てマスキングしてしまうのではなく，人間が目的や状況に応じてマスキングすべき固有表現かどうかを適切に判断する必要がある．

3 提案手法

本論文では，機密情報のマスキングを固有表現抽出のタスクではなく，情報フィルタリングのタスクとして捉え，安全性が高く，目的や状況に応じて適切にマスキングする固有表現を選別することが可能なマスキング手法を提案する．

3.1 アンマスキング法

機密情報を無害化する手段として，固有表現を検出してマスキングを行うのではなく，全ての文字列がマスキングされている状態から，人手によって安全な単語を選別してマスキングの解除を行う，アンマスキング法を提案する．

情報フィルタリングの観点から述べると，従来手法がブラックリストを整備するのに対して，提案手法はホワイトリストを整備することにより，確実に機密情報をフィルタリングしようとするものである．

本手法では，人手によって整備された安全語リストが必要となるが，以下の手順により効率よく安全語リストを作成することができる．

はじめに，マスキング対象となる全てのテキストに対して形態素解析を行い，出現する単語を集計して，単語ごとに単語名，品詞，出現頻度を記録した出現単語リストを生成する．このリストは形態素解析により自動生成を行うが，一般的な形態素解析器による分かち書きの精度は現状で約 99% であり [3]，実用上問題ない．

続いて，単語ごとに出現頻度をベースとしたスコアを算出する．単語のスコアは，例えば，出現頻度や文書全体の累計文字数（＝出現頻度 × 単語の長さ）などの値を利用したり，品詞による重み付けをすることが考えられる．このように算出されたスコアが高い単語から順番にマスクを解除してよいかどうかを判断し，安全語リストを作成する．

本手法で最大のネックとなっているのは，人手により安全語リストを整備しなければならないところである．しかし，一般に単語の出現頻度はべき乗分布になることが知られており [5]，重要度が高い単語に絞ってチェックを行うことで，文書中に出現する全ての単語をチェックしなくても，効率よくマスクを解除することができる．

また，安全語であることが明らかである単語リスト，例えば，製品名や専門用語が登録されているリストを外部から取り込むことで，安全語リストを整備する作業量を軽減することも可能である．

3.2 提案手法によるマスキングの特徴

1. 従来手法よりも安全性の高いマスキング
   人手によるチェックを通過した単語のみが公開されるため，安全性の高いマスキングを行うことができる．辞書登録されていない固有表現に関しても，人手により選別することで適切にマスキングを行うことができる．たとえ形態素解析に誤りがあったとしても，最終的にその文字列を公開するかどうかを判断するのは人間である．誤りと疑われる場合は安全性を重視してマスクを解除しない，といった柔軟な対応が可能となり，安全性の高いマスキングを行うことができる．

   また，本手法において形態素解析を行う際に固

有表現を抽出して出現単語リストを作成することで，少なくとも従来手法より安全にマスキングが行えることが保証される．

2. 目的や状況に応じたマスキングが可能
必ず人手を介して単語のチェックを行うため，マスキングの目的や状況に応じて，マスキングする必要のない固有表現を適切に選別し，文書の無害性を保ったまま利用価値を高めることができる．

3.3 提案手法の適用範囲

近年，テキストマイニング技術の発達により，顧客から寄せられる問い合わせ内容を分析することで，タイムリーに顧客の声を吸い上げることが重要になっている．その際に問題となるのが個人情報などの機密情報だが，これを除去して無害化することにより，組織内でテキストマイニングシステムを活用し，得られた分析結果を安全に共有することが可能となる．

提案手法は，機密情報のマスキングに限らず，一般的なテキスト情報フィルタリング手法として適用することができる．例えば，Web のコンテンツフィルタとして利用し，有害なテキスト情報をフィルタリングすることも可能である．現在，Web のコンテンツフィルタは，人手で作成した URL リストを用いるか，ページ内のテキストを解析して，有害なページかどうかの判定を行っている [2]．しかし，有害でないページが有害ページと判定されたために，必要な情報が得られないという問題が起こる．そこで，有害と判定されたページについては，本手法により有害でないテキストのみを開示することで，必要な情報が得られる可能性を高めることができる．

また，情報を共有する際に障害となるのは，セキュリティやプライバシーの問題である．オフィス文書や電子メールなど，個人で所有しているデータを解析することで有用な知見を発見できる可能性があるにも関わらず，実際にはプライバシーの問題でデータが得られないことが多い．しかし，こういったプライバシー情報を含むデータを分析する場合に，本手法によるマスキングを行うことで，必要最小限の情報のみを開示しつつ分析を行うことが可能となる．例えば，個人で所有しているマシン内のファイルや，ネットワーク上を流れる電子メールのテキストを分析する場合にも有効であると考えられる．

4 実験

4.1 実験の概要

本手法の実用性を検証するために，人名，地名，組織名などの固有表現が含まれている，IBM の社内データを用いて実験を行った．マスキング対象の固有表現を，人名，地名，組織名，郵便番号，電話番号とし，マスキングされた文書が IBM 社内で利用されることを想定して，IBM に関する企業情報はマスキングの対象外とした．以上のポリシーに従い，マスキングすべき固有表現を人手でタグ付けしたデータを作成し，このタグ付きデータに対して提案手法によるマスキングを行った．

出現単語リストを作成する際には，IBM 東京基礎研究所で研究開発を行っている形態素解析エンジンを用いている．この形態素解析エンジンは，入力された日本語文の分かち書きを行い，品詞情報を付与して結果を出力する．付与される品詞情報には，人名や地名，組織名などの固有表現も含まれており，固有表現抽出エンジンとしての機能を兼ね備えている．ただし，本実験では未知語の辞書登録などの対象分野に即したチューニングを行っていないため，形態素解析では検出されない固有表現が少なからず存在する．

また，出現単語リストを人手でチェックする際に，単語が安全であると判定する基準は以下のようにした．ここで，名詞，固有名詞，未知語のことを名詞類と呼ぶことにする．

- 名詞類の単語は，単語名にマスキングすべき固有表現が含まれていなくても，他の単語と結びつくことにより固有表現になりうる単語に注意を払って安全かどうかを判定する．
- 数詞の単語は，電話番号，郵便番号，住所の番地，マンションの部屋番号になっている可能性があることに注意して，安全かどうかを判定する．
- 名詞類，数詞以外の単語は，形態素解析の結果が正しいと判断できる単語は安全であると判定する．

- 出現頻度が低く，文字数が少ない単語は，品詞によらず，固有表現の部分文字列になる可能性があることに注意して，安全かどうかを判定する．

この判定基準に基づいて安全語リストを作成することにより，言語処理エンジンが認識できなかったマスキングすべき固有表現や，その部分文字列を安全語リストに登録するのを防ぎ，それらがマスキングされた状態を保つことができる．

なお，本実験では，提案手法の定量的な評価に主眼を置いているため，単語に固有表現が含まれているにも関わらず安全語リストに登録してしまった，といった人的ミスは起こらないものとする．

4.2 評価指標

ここで，提案手法の定量的な評価を行うために評価指標を定義する．

正解データにおいてマスキングすべき文字列で，実際にマスキングされた文字数，マスキングされなかった文字数を，それぞれ，tp，tn，正解データにおいてマスキングすべきでない文字列で，実際にマスキングされた文字数，マスキングされなかった文字数を，それぞれ，fp，fn，とする（表 1）．

このとき，マスキングの再現率 (precision) と適合率 (recall) は次のように定義される．

$$\text{再現率} = \frac{tp}{tp + tn}$$

$$\text{適合率} = \frac{tp}{tp + fp}$$

マスキングのタスクにおいて，再現率は「マスキングすべき文字列が実際にマスキングされた割合」を表し，適合率は「実際にマスキングされた文字列がマスキングすべきものである割合」を表す指標となる．

従来の固有表現抽出のタスクでは，一般に再現率，適合率による抽出精度の評価が行われている．しかし，本手法は固有表現を抽出してマスキングするものではなく，安全語をアンマスキングするものなので，安全語を基準とした指標で評価を行うべきである．

そこで，「実際にマスキングされていない文字列がマスキングすべきものでない割合」を表す指標として，可読率 (readability) を導入し，以下のように定義する．

表 1: 評価指標

	マスキングされた	マスキングされない
マスキングすべき	tp	tn
マスキングすべきでない	fp	fn

$$\text{可読率} = \frac{fn}{fp + fn}$$

また，提案手法では人手により安全語リストを整備するため，安全語リストを作成するのに必要な作業量についても評価する必要がある．そこで，作業量を表す指標としてコストを導入し，1 単語あたりを人手で安全かどうかをチェックするのにかかる作業量を 1 コストと定義する．

以下では，安全性を表す再現率，可読性を表す可読率，効率性を表すコストを用いて，提案手法の定量的な評価を行う．

4.3 コストと可読性の評価

提案手法では，あらかじめ単語を集計しておくことで，効率よく安全語リストを作成することができる．安全語リストを作成する際には，単語ごとに出現頻度をベースとしたスコアを算出し，スコアの高い順番で単語のチェックを行う．そこで，スコアに単語の出現頻度，累計文字数を用いた場合のコストと可読率の関係を調査した．

作成したタグ付きデータ 9,000 文を使用して形態素解析を行い，単語の集計を行った．この中に含まれる総単語数は 106,235，異なり語数は 5,716 であった．

単語の出現頻度，累計文字数順に単語のチェックを行い，コストが 100，500，1,000，2,000，5,000 のときの可読率を比較した．表 2 に実験結果を示す．表中のカバー率は，異なり語数に対するチェックした単語数の割合を表している．

実験結果から，頻度をベースに算出したスコア順にチェックを行うことで，効率よく可読率が上昇していることが分かる．また，可読率を 90% にするため

表 2: コストと可読率の推移

コスト	100	500	1000	2000	5000
カバー率	1.75%	8.75%	17.5%	35.0%	87.5%
出現頻度順	0.436	0.716	0.806	0.887	0.974
累計文字数順	0.464	0.732	0.823	0.899	0.974
差分	0.028	0.016	0.017	0.012	0.000

に必要なコストは 2,000 程度で，全ての単語をチェックするコストの約 35% で済むことが分かった．

スコアを出現頻度，累計文字数と変更することによる可読率の推移については，両者の間に大きな差は見られなかった．しかし，累計文字数の順にチェックを行うことにより，安全語の判定が難しい，短い単語名を持つ単語のスコアを低く抑えることができるため，効率よく可読率を上げることができると考えられる．

また，今回は評価の対象外であるが，単語をチェックする際に品詞ごとに出現単語リストを作成することで，人手でチェックする際に着目する品詞を固定して安全な単語かどうかを判定できるようになるため，作業効率を改善することができると考えられる．

4.4 可読性と安全性の評価

提案手法では，固有表現抽出に対応した言語処理エンジンを用いることにより，既存の固有表現抽出技術を用いたマスキング手法よりも必ず安全性が高くなることが保証されている．しかし，その安全性がどれくらいのレベルなのかを検証するために，人手により安全語リストを作成し，マスキングを実施した．

実験データは，コストと可読性の評価の実験で用いた正解タグ付きデータ 9,000 文を使用した．マスキングに使用した安全語リスト A，B，C，D，E，F は，累計文字数をスコアとしたときのコストがそれぞれ 100，500，1,000，2,000，5,000 となる安全語リストである．これらの安全語リストを使用したときの可読率，再現率を調査した．

実験結果を表 3 に示す．表中のカバー率は，異なり語数に対するチェックした単語数の割合を表している．また，正解，一部正解，不正解の値は，それぞれ，マスキングすべき固有表現が完全にマスキングされた数，一部マスキングされた数，マスキングされなかった数を表している．また，参考データとして，出現単語リストを生成する際に使用している形態素解析エンジンで抽出された固有表現をマスキングした場合の精度を載せてある．

この可読率と再現率の推移を可視化したのが図 1 である．このグラフから，コストをかけて安全語リストを整備することにより，安全性を完全に近い水準で保ったまま可読性が向上しているのが確認できる．十分なコストを費やすことで，高い再現率を保ちつつ，従来の固有表現抽出によるマスキングとほぼ同等の可読率を実現することができる．

また，提案手法が従来の固有表現抽出によるマスキングと比べて特徴的なのは，たとえマスキングすべき固有表現のマスクがはがれてしまったとしても，完全にマスクがはがれてしまうことが少ない点である．

これは，従来手法で起こるマスキング漏れが主に固有表現の辞書登録漏れにより発生するのに対し，提案手法で起こるマスキング漏れは主に形態素解析により固有表現が分かち書きされてしまうことにより発生するためであると考えられる．

以下，提案手法でマスキング漏れが発生した固有表現の特徴を述べる．

1. 複合語により形成された固有表現
   例えば,「下鶴間自動車整備工場」という架空の組織名があったとして，形態素解析エンジンにより次のように分かち書きされた場合．

   下鶴間（地名）
   自動車（名詞）
   整備（名詞）
   工場（名詞）

   「下鶴間」は地名と判定されることでマスキングされるが,「自動車」「整備」「工場」を一般的

表 3: 可読性と安全性の評価

リスト	コスト	カバー率	可読率	再現率	正解	一部正解	不正解
A	0	0%	0	1.000	772	0	0
B	100	1.75%	0.464	0.999	767	5	0
C	500	8.75%	0.732	0.997	765	7	0
D	1000	17.5%	0.823	0.996	763	9	0
E	2000	35.0%	0.899	0.996	763	9	0
F	5000	87.5%	0.974	0.989	756	14	2
固有表現抽出精度			0.990	0.638	582	65	120

表 4: 新しく追加されたデータに対する有効性の評価

リスト	コスト	可読率	再現率	正解	一部正解	不正解
A	0	0	1.000	80	0	0
B	100	0.434	0.997	79	1	0
C	500	0.690	0.997	79	1	0
D	1000	0.774	0.997	79	1	0
E	2000	0.826	0.997	79	1	0
F	5000	0.872	0.990	79	0	1

な名詞であるとして安全語と判定すると，一部マスキング漏れとなってしまう．

2. 平仮名で表記された固有表現
例えば，下鶴間という地名が平仮名で「しもつるま」と表記されており，形態素解析エンジンにより次のように分かち書きされた場合．

し（動詞「する」連用形）
も（動詞「もつ」連体形）
つ（動詞連体形活用語尾）
るま（未知語）

このうち,「るま」は未知語のため安全語と判定しない可能性が高いが，それ以外の単語に関しては形態素解析の結果が正常であるとみなして安全語であると判定すると，一部マスキング漏れが発生してしまう．なお，実験において不正解となったのもこのパターンであった．

以上の分析により，提案手法では，一部マスキング漏れとなるケースに比べて，完全にマスキング漏れとなるケースが少なくなると考えられる．一部マスキング漏れでも固有表現の特定ができない場合も多いことから，提案手法は安全性が高いマスキング手法であると言える．

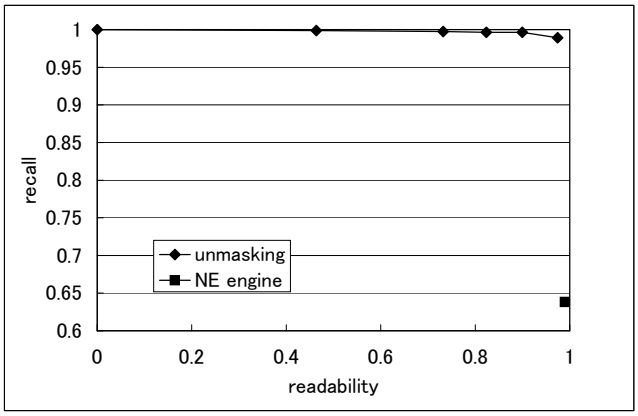

図 1: 可読率と再現率の推移

4.5 新しく追加されたデータに対する有効性の評価

最後に，新しく追加されたデータに対して，既存のデータを元に作成した安全語リストを用いたとき

の可読性，安全性を調査した．

安全語リストは，可読性と安全性の実験において使用した A，B，C，D，E，F と同じものを用いている．新しく追加するデータは，既存のものと同じ分野のデータ 1,000 行で，この中に含まれる総単語数は 11,694，異なり語数は 1,599 であった．

表 4 に実験結果を示す．実験結果から，新しく追加されたデータに対しても既存の安全語リストを用いることにより高い安全性を保っていることが分かる．

また，可読率の推移を図 2 に示す．新しく追加するデータに，既存のデータにない新しい単語が含まれることで可読性は減少するが，差分の単語のみをチェックして安全語リストを更新することにより，あまりコストをかけずに元の可読率の水準を保つことができると考えられる．

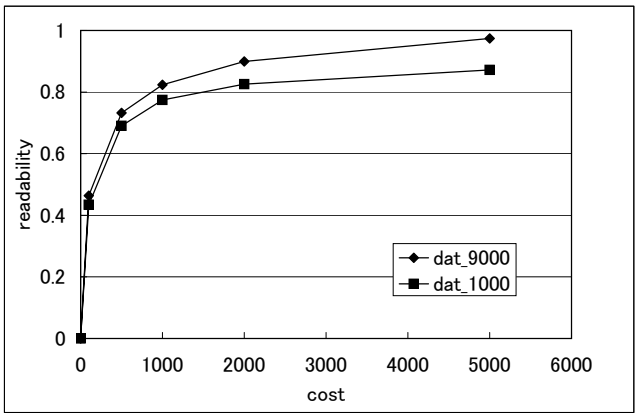

図 2: 新しく追加されたデータに対する可読率の推移

5 まとめ

本論文では，人手によって指定された安全語のマスクを解除することにより機密情報のマスキングを行うアンマスキング法を提案した．提案手法では，マスキング対象のテキストデータに出現する単語を集計することで，効率よく安全語リストを作成することができる．従来の固有表現抽出によるマスキングと比較して，人手により安全と判定された単語のみが公開されるため安全性が高く，目的や状況に応じて適切にマスキングする固有表現を選別できる点で優位性がある．

実験結果から，提案手法が効率よく安全語リストを作成でき，安全語リストを整備することで，安全性を損なわないように可読性を上げられることが分かった．また，新しく追加されたデータに対しても，既存のデータに対して作成した安全語リストが有効に働くことを確認した．

今後の課題としては，安全性を強化するために，短い単語はチャンキングした単位で集計して出現単語リストを生成する，などの対策が考えられる．また，機密情報だけでなく種々の有害なテキスト情報をフィルタリングするための手法として活用できるよう検討を行っていく．

参考文献

[1] http://nlp.cs.nyu.edu/irex/.

[2] 井ノ上 直己，帆足 啓一郎，橋本 和夫．“文書自動分類手法を用いた有害情報フィルタリングソフトの開発”，電子情報通信学会論文誌，Vol. J84-D-II, No.6, pp.1158–1166, 2001.

[3] 工藤 拓，山本 薫，松本 裕治．“Conditional Random Fields を用いた日本語形態素解析”，IPSJ SIG Technical Reports, NL-161, 2004.

[4] 関根 聡，江里口 善生．“IREX-NE の結果と分析”，言語処理学会第 6 回年次大会 予稿集, 2000.

[5] G. K. Zipf. “Human Behavior and the Principle of Least Effort”, Addison-Wesley, 1949.