連想知識を用いた端的な要約の生成

瀧川和樹†, 村田真樹‡, 土田正明‡, De Saeger Stijn‡, 山本和英†, 鳥澤健太郎‡
† 長岡技術科学大学 電気系　{takigawa, yamamoto}@jnlp.org
‡ 情報通信研究機構 MASTAR プロジェクト 言語基盤グループ {murata,m-tsuchida,stijn,torisawa}@nict.go.jp

1　はじめに

近年、Web の発展により我々は多くの情報を容易に得られるようになった。しかしその情報量は日々増加しており、そのすべてに目を通すことは実質不可能である。この作業を効率化させるための技術にテキスト自動要約がある。この自動要約の手法は、原文からの重要文抽出や不要箇所削除による文短縮など、表層的な手法によるものが主である [1]。しかし、これらの手法は基本的に原文にあった表現から一部を選択し、要約として表示するため、元の文書の内容をなるべく損なわずに文章を圧縮するにはおのずと限界がある。元の文書にない表現も使うことで、元の文書の内容をなるべく残した適切な要約ができると考える。例えば、冗長に記述された「爆弾が爆発した。死傷者が出た。反政府運動がきっかけである。」などの文章であれば、それを端的に表現した「テロ」と要約する、また、「会社に行くために、朝起きて、歯を磨いて、朝食を食べた。書類の整理をした。」という文章であれば「出社準備」という端的な要約を生成することが望ましい。

これを実現するためには、原文にあった表層的な情報を利用するだけではなく、原文の内容を適切な表現に換言する技術も用いることが必要である。本稿では連想知識として共起情報を用いて入力文章 (1 文でも良い) を換言し、要約前の文章をなるべく適切に連想できる表現を良い要約とする手法を提案する。具体的には、入力から連想される語を共起語と仮定し、入力内に存在する各名詞と共起する語をそれぞれ取得する。得られた共起する語を要約の候補とし、以下の２つの基準を最も満たす候補を要約として出力する。

(i)　要約結果から十分に原文の内容を連想できる。
(ii)　要約結果から原文の内容にないものがなるべく連想されない。

これらの基準をどれだけ満たしているかの判断は、5 節で示す式により数値化することで行う。

本研究では出力を文とすることを目指しているが、現段階では 1 単語での出力に限っている。この 1 単語での出力は、応用として文書のカテゴライズに用いることも考えられる。

2　関連研究

換言を用いた要約の手法として、近藤ら [2] の研究がある。日本語単語辞書の語釈文を用いて複数の動作を１つの語に換言する手法や、概念辞書（上位下位辞書）を用いて複数の動作に共通する上位概念に換言する手法が提案されている。しかし、この手法では動作的な概念についての換言にとどまっているという問題がある。また、語釈文や概念辞書では換言できない事象を扱えない問題もある。本手法では共起情報を用いることで、動詞のみの換言ではなく、入力文章全体を１つの単語に置き換えることを試みている。さらに、たとえば「身長制限」という語に対しては「アトラクション」という語が共起している。このように、共起情報を用いることにより概念辞書などでは得られない語も取得でき、より柔軟な換言が期待できる。

端的な要約を行う研究と似たものに、Banko ら [3] の研究がある。Banko らは統計的翻訳の技術を用いて入力に対するヘッドラインを生成している。これも一種の端的な要約であるといえる。

また、共起情報を要約に用いる研究もある [4][5]。しかし、これらの手法はいずれも共起情報を重要箇所の特定に用いており、換言を行うことはしていない。一方、本手法では入力文章内に存在する名詞と共起する語を要約の候補として用いる点で異なる。

3　共起語の定義

本手法では連想知識として共起情報を用いる。共起情報には、その共起強度を示す相互情報量やダイス係数などの指標があるが、本手法では簡便な情報である共起頻度を用いることにした。具体的には、5,000 万の Web 文書内に存在する各文に対し、共起関係にある２つの名詞と、その共起頻度をリスト化したものを用いる。リストは、「名詞W」「Wに共起する名詞」「共起頻度」の３つの要素で構成されている。本稿では、このリストに存在する単語Wに対しスコアが上位N語以内に入っている共起する名詞すべてをWに対する共起語と呼ぶ。たとえば、W＝ “発生”としたとき、リストは表 1 のようになっている。

表 1: W＝ “発生”のときの共起リスト

名詞W	Wに共起する名詞	共起頻度
発生	問題	143815
発生	損害	120965
発生	エラー	109188
発生	利用	89802
発生	地震	88296
:	:	:

このとき、N＝５とすれば、“発生”に対する共起語は「問題, 損害, エラー, 利用, 地震」となる。ただし、この共起頻度は処理負担軽減のため、近似的なものとなっている。

4　提案手法

本手法のおおまかな流れは以下の通りである。

1. 入力文章に存在する名詞を取得
2. 取得した各名詞の共起語を取得
3. 取得した共起語に対し、1 節で述べた基準 (i),(ii) に基づく評価値を計算
4. 評価値が最も高い共起語を要約として出力

4.1　入力文の名詞取得

入力文章から名詞を取得する。そのために、入力に対して形態素解析を行う。形態素解析には JUMAN(1) を用いた。形態素解析の結果、品詞が「名詞」となった単語を抽出する。さらに複合名詞を考慮するため、隣り合う形態素の品詞が両方とも名詞であった場合、これらを１つの名詞として取得する。また、日本語の場合、カタカナで書かれた言葉は外来名詞であることが多い。よって、品詞が「未定義語-カタカナ」となった語も取得する。

4.2　出力候補の取得

要約として出力する単語の候補を取得する。その候補として、入力内の名詞に対する共起語を用いる。これは、入力にあった名詞から連想されやすい名詞ほど端的な要約としてふさわしいと考えたからである。入力から得られた各名詞をWとし、共起語を取得する。本手法ではN＝50 で固定した。以下、本節で取得した単語を「出力候補」と呼ぶ。

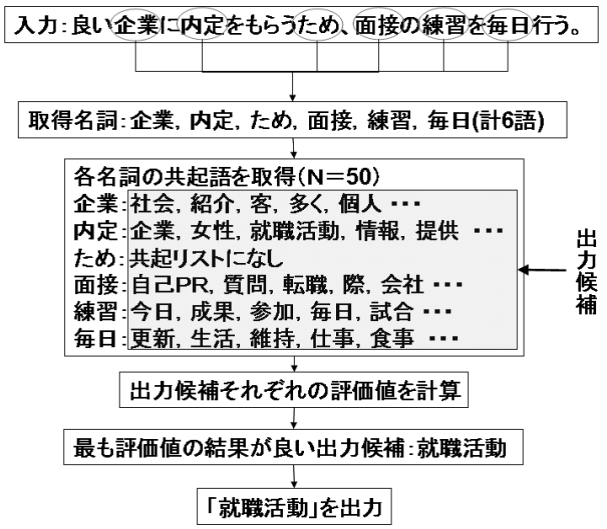

図 1: 実行例

4.3 評価値の算出

取得された出力候補の中から、実際に出力する語を選択する。そのために、1 節で述べた 2 つの基準を数値化し、それを評価値として用いる。入力の内容を正解の情報とすると、基準 (i) は、その正解の情報をどれだけ漏らさずに取り出せるかを示していると言える。よってこれは再現率に類似する。また基準 (ii) は、要約から想起されるものがどれだけ入力の内容を逸脱しないかを示しているといえる。よってこれは適合率に類似する。これらを今回の手法に当てはめると、それぞれの基準は以下のように考えることができる。

- 1 の基準は、出力候補の共起語 (＝出力から連想できる内容) が入力に含まれる名詞（＝正解の情報）を多く持つほど良いと言い換えられる。
- 2 の基準は、出力候補の共起語が入力に含まれない名詞（＝誤りの情報）をできるだけ持たないほど良いと言い換えられる。

このことから、各出力候補の共起語を取得し、それを用いて評価値を算出する。ただし、入力の名詞以外にも入力文の内容を示す名詞は存在し得る。例えば、入力にある名詞の共起語（つまり出力候補）は、入力の話題を含んでいる場合がある。これに基づくと、基準 (ii) の「入力の内容を逸脱しない」という意味では、入力にある名詞の共起語も正解とする方法も考え得る。そこで、基準 (ii) においては「入力内にある名詞のみを正解とする場合」と、「入力内にある名詞及びその共起語を正解とする場合」の 2 通りの評価値を用意した。実際の手法では、どちらかの評価値しか用いない。これらを踏まえて、出力候補を c とすると各評価値は以下の式から求められる。基準 (i) を数値化した評価値を $Recall(c)$，基準 (ii) で、入力内にある名詞のみを正解とする場合の評価値を $Precision_1(c)$，入力内にある名詞及びその共起語を正解とする場合の評価値を $Precision_2(c)$，$Recall(c)$ と $Precision_1(c)$(あるいは $Precision_2(c)$) の調和平均を $F\text{-}measure(c)$ とする。

$$Recall(c) = \frac{|RelatedWord(c) \cap InputWord|}{|InputWord|}$$

$$Precision_1(c) = \frac{|RelatedWord(c) \cap InputWord|}{|RelatedWord(c)|}$$

$$Precision_2(c) = \frac{|RelatedWord(c) \cap (Input \cup Related_i)|}{|RelatedWord(c)|}$$

$$(Input = InputWord)$$

$$(Related_i = \cup_{i \in InputWord} RelatedWord(i))$$

出力候補：c=自己PR
共起語：面接，ポイント，履歴書，書き，志望動機，自信，
例文，自己分析，私，エントリーシート，日記，考え，
記入，セオリー，自分，版，雇用条件，効果，・・・
| InputWord |：6　　| RelatedWord(c) |：50
| RelatedWord(c) ∩ InputWord |：1
| RelatedWord(c) ∩(Input ∪Related$_i$) |：32
Recall(c)=1/6=0.167　Precision(c) =32/50=0.640　F- measure(c)=0.264

出力候補：c=就職活動
共起語：学生，情報，機能，ML，企業，応援，支援，内定，
求人情報，基本，時期，個人，アドバイス，成功，
提供，面接，先輩，役，毎日，現在，・・・
| InputWord |：6　　| RelatedWord(c) |：50
| RelatedWord(c) ∩ InputWord |：4
| RelatedWord(c) ∩(Input ∪ Related$_i$) |：29
Recall(c)=4/6=0.667　Precision(c)=29/50=0.580　F- measure(c)=0.620

図 2: 評価例

$$F\text{-}measure(c) = \frac{2 \cdot Recall(c) \cdot Precision(c)}{Recall(c) + Precision(c)}$$

ここで、$InputWord$ は入力から取得されたすべての名詞の集合，$RelatedWord(x)$ は x から得られた共起語の集合，$|U|$ は集合 U の要素数とする。この式により出力候補すべての評価値を算出する。

4.4 出力候補の並び替え

出力候補を、4.3 節で求めた評価値の良い順に並び替える。そして、最も良い評価値を持つ出力候補を要約結果として出力する。本来ならば、$Recall(c)$ と $Precision(c)$ の調和平均である $F\text{-}measure(c)$ の値を優先的に用いて並び替えを行うのが一般的であるが、本手法では $Recall(c)$ や $Precision(c)$ を優先した並び替えも行い、5 節で実際にどの評価値が一番良い結果となるかを観察した。また、並び替えを行う際に優先した評価値が同じ値だった場合には他の評価値を用いてさらに並び替えを行う。これらをまとめると、本手法から考えられる出力の選択法は以下の 5 通りである。

手法 1
基準 (ii) の数値化に $Precision_1(c)$ を用いる。得られた評価値の中からどれか 1 つを優先して並び替える。

手法 2
基準 (ii) の数値化に $Precision_2(c)$ を用いる。得られた評価値の中から $Recall(c)$ を優先して並び替える。

手法 3
基準 (ii) の数値化に $Precision_2(c)$ を用いる。得られた評価値の中から $Precision_2(c)$ を優先して並び替える。

手法 4
基準 (ii) の数値化に $Precision_2(c)$ を用いる。得られた評価値の中から $F\text{-}measure(c)$ を優先して並び替える。$F\text{-}measure$ の値が同じだった場合には、次に $Recall(c)$ を優先して並び替える。

手法 5
基準 (ii) の数値化に $Precision_2(c)$ を用いる。得られた評価値の中から $F\text{-}measure(c)$ を優先して並び替える。$F\text{-}measure(c)$ の値が同じだった場合には、次に $Precision_2(c)$ を優先して並び替える。

普通に考えると、$Precision_1$ を用いた場合にも並び替えを行う際に 3 つの評価値をそれぞれ優先する必要がある。しかし、本手法では共起語を取得する際 N=50 に固定、つまり $|RelatedWord(x)| = 50$ で固定している。そのため、$Recall(c)$ と $Precision_1(c)$ は、ともに $|RelatedWord(c) \cap InputWord|$ のみに依存する式となり、これら 2 式に差異はなくなる。

この 2 つの評価値に差異がないとなると、これらの調和平均である $F\text{-}measure(c)$ も同様に差異はなくなる。よって、

表 2: strict による評価結果

			1 位	5 位以内	10 位以内	MRR
手法 1	$Precision_1(c)$		0.125	0.130	0.304	0.109
手法 2	$Precision_2(c)$	$Recall(c)$	0.167	0.292	0.333	0.218
手法 3		$Precision(c)$	0.000	0.167	0.250	0.060
手法 4・5		$F\text{-}measure(c)$	0.083	0.292	0.375	0.173

表 3: lenient による評価結果

			1 位	5 位以内	10 位以内	MRR
手法 1	$Precision_1(c)$		0.261	0.304	0.783	0.374
手法 2	$Precision_2(c)$	$Recall(c)$	0.333	0.583	0.750	0.447
手法 3		$Precision(c)$	0.000	0.250	0.458	0.096
手法 4・5		$F\text{-}measure(c)$	0.167	0.583	0.750	0.338

$Precision_1(c)$ を用いた場合には $Recall(c)$, $Precision_1(c)$, $F\text{-}measure(c)$ はすべて等価の評価値となり、どれか 1 つの評価値による並び替えを行うだけで良くなる。

また、基準 (ii) の数値化に $Precision_2(c)$ を用いた場合には、3 つの評価値をそれぞれ優先して並び替えを行っている。このとき、優先した評価値が同じだった場合には残り 2 つの評価値のどちらかを用いてさらに並び替えを行う。ここで $Recall(c)$ を優先したとき、$Recall(c)$ が同じ値となる出力候補があるとする。このとき、$Recall(c)$ と $Precision(c)$ の調和平均である $F\text{-}score(c)$ は、$Precision(c)$ のみに影響されることとなり、この場合の $Precision(c)$ と $F\text{-}score(c)$ は評価値としては等価となる。よって、どちらか一方で並び替えを行うだけで良い。$Precision_2(c)$ を優先して並び替えを行った場合も同様である。

例として、手法 2 による実行手順を示す。入力として「良い企業に内定をもらうため、面接の練習を毎日行う。」という文を与えたとする。この入力文からは「企業，内定，ため，面接，練習，毎日」という 6 語の名詞が得られる。次に、得られた各名詞の共起語を 50 語取得する。それぞれの名詞からは図 1 のような共起語が得られる。この共起語が出力候補となる。評価値を求めるため、各出力候補の共起語を取得する。評価値の算出例は、図 2 のようになっている。次に、得られた評価値の値が高い順に並び替える。出力候補を $Recall(c)$ 優先で並び替える場合、「就職活動」が最も高いスコアとなる。よって、この入力文の要約は「就職活動」となる。

5 評価実験

5.1 評価方法

本手法の有効性を検討するため、人手で作成した 24 文の入力を用意し、4.4 節で述べた 5 通りの手法すべての結果を、被験者 1 名により評価した。

評価方法は、出力候補を並び替えた結果の上位 10 語までを回答とし、その回答の中に正解といえる要約結果が含まれているか、また含まれている場合、何位にその語があるかを調査した。評価基準として、strict と lenient の 2 種類を用意した。strict は、正しい回答のみを正解とする評価方法である。lenient は、正しい回答ではないが、それに近いものも正解に含める評価方法である。さらに、strict,lenient の評価結果を元にして、以下の 4 つの評価を行った。

1. 並び替えの結果、1 位の語が正解であった場合のみを正解とした際の正解率
2. 並び替えの結果、5 位以内の回答に 1 つでも正解を含めば正解とした際の正解率
3. 並び替えの結果、10 位以内の回答に 1 つでも正解を含めば正解とした際の正解率
4. MRR

MRR とは、以下の式で表される評価値である。

$$MRR = \frac{\sum_{i=1}^{M} 1/r_i}{M}$$

M は評価する対象の総数、r_i は評価対象 i がもつ最も高い正解の順位である。今回は入力を 24 文用意しているため、$M = 24$ である。また、並び替えた結果の上位 10 語を回答としているため、$1 \leq r_i \leq 10$ となる。評価対象 i が正解を 1〜10 位以外にもつ、もしくは出力候補に正解がない場合は $1/r_i = 0$ とする。

5.2 評価結果

strict による評価結果、lenient による評価結果はそれぞれ表 2, 表 3 のようになった。$Precision_2(c) - F\text{-}measure(c)$ の組み合わせを用いる場合には、さらに 2 通りの組み合わせ (手法 4 と手法 5) があったが、これらの結果は両方ともまったく同じになったため統一して記述している。下線を引いてある値は、その評価方法の中で最も良い結果を示している。

6 考察

6.1 各手法の結果

strict，lenient の 10 位以内における正解率以外では、すべての評価基準において手法 2 が最も良い結果となった。この手法による結果の例を以下に示す。「セキュリティ^s」のように付加されている “ s ” は、strict の際に正解とした単語を示す。同様に “ l ” が付加されている語は lenient の際に正解とした単語を示す。

入力 1 プライバシーを守るため、個人情報を保護するように設定を行った。

理想の正解 セキュリティ

上位 10 語

1. セキュリティ^s　2. ヘルプ　3. セキュリティー
4. ポリシー　5. 保護　6. 利用規約　7. リンク
8. 使用　9. 利用　10. 著作権

入力 2 警察が容疑者を捕まえた。引き続き拘束を行うようだ。

理想の正解 逮捕

上位 10 語

1. 身柄　2. 拘束　3. 容疑　4. 逮捕^s　5. 事情聴取
6. 疑い　7. 痴漢　8. 現行犯逮捕　9. 供述　10. 捜査

入力 3 実験の目的や理論をまとめ、図書館を利用して課題を作成する。

理想の正解 レポート作成

上位 10 語
1. 実習 ^{l}　2. 研究　3. 考察　4. 分析　5. 資料
6. 計画　7. 実験　8. サービス　9. 手法　10. 検証

入力 4　机に向かい教科書と授業のノートを開いた。今日の復習と明日の予習をする必要がある。

理想の正解　勉強

上位 10 語
1. 予習　2. 宿題 ^{l}　3. 準備　4. 理科　5. 模試
6. 教科書　7. 復習　8. 講義　9. 数学　10. 学校

入力 5　目当ての株を全部、購入した。ひとり占めの状態である。

理想の正解　買占め

上位 10 語
1. 物　2. クルマ　3. 車両　4. チェック　5. 本
6. 利用　7. 今　8. 確認　9. 車　10. 商品

入力 1 は 1 位に strict においての正解が得られた結果例、入力 2 は 5 位以内に strict においての正解が得られた結果例である。入力 3 は 1 位に lenient においての正解が得られた結果例、入力 4 は 5 位以内に lenient においての正解が得られた結果例である。また、入力 5 は 10 位以内に strict,lenient ともに正解が得られなかった結果例である。

評価は 1 人の被験者が主観で行ったため、その主観による影響を受けている可能性がある。そこで、評価データの一部に対しもう一人被験者を加え、計 2 名で評価を行った。その結果をもとに κ 値を算出した結果、strict の場合は $\kappa = 0.775$, lenient の場合は $\kappa = 0.746$ となった。κ 値は、0.6 を超えれば良い値だとされているため、主観評価の影響は小さいといえる。

6.2　$Precision_1(c)$ と $Precision_2(c)$

$Precision_1(c)$ と $Precision_2(c)$ を比較すると、$Precision_2(c)$ を用いた手法のどれかが、ほとんどの場合 $Precision_1(c)$ より良い結果となっている。lenient の 10 位以内における正解率の評価結果のみ $Precision_1(c)$ の方が良い場合もあるが、出力の結果を見ると以下のような問題がみられた。例えば、「食事制限をし、毎日ジョギングなどの運動を行う。」という入力を与えたときの手法 1, 手法 2 の上位 5 位までの結果は以下のようになっている (それぞれ、Re:$Recall(c)$, Pre:$Precision(c)$, F:$F\text{-}measure(c)$)。

・手法 1

1	ダイエット	Re:0.667	Pre:0.040	F:0.075
2	消費エネルギー	Re:0.667	Pre:0.040	F:0.075
3	脂肪燃焼	Re:0.667	Pre:0.040	F:0.075
4	プロモデル	Re:0.667	Pre:0.040	F:0.075
5	減量	Re:0.667	Pre:0.040	F:0.075

・手法 2

1	ダイエット	Re:0.667	Pre:0.440	F:0.530
2	無理	Re:0.667	Pre:0.392	F:0.494
3	筋トレ	Re:0.667	Pre:0.392	F:0.494
4	有酸素運動	Re:0.667	Pre:0.380	F:0.484
5	脂肪燃焼	Re:0.667	Pre:0.360	F:0.468

両者とも、1 位には理想の正解といえる「ダイエット」が得られている。しかし、手法 1 の結果を見ると、1 位から 5 位までの評価値がすべて同じとなっている。一方、手法 2 の結果は、Recall の値はすべて同じだが、その他の評価値に差があるため同率で 1 位ということにはなっていない。これは、$Precision_2(c)$ を用いたすべての手法に共通した結果であった。実際に要約として出力する際には 1 つの単語のみに絞る必要がある。そのため、手法 1 のように同率となる語が多くなる手法は良いとは言えない。このことから、$Precision_1(c)$ より $Precision_2(c)$ のほうが優位であるといえる。

6.3　評価値の優先順位

評価値として、$Recall(c)$,$Precision(c)$,$F\text{-}measure(c)$ を用意した。3 つの評価値を出力候補の並び替えに用いた結果、どの結果でも $Recall(c)$ を優先した場合が最も良い結果となった。このことから、1 節で述べた (ii) の基準より (i) の基準のほうが出力候補の選択の際には優位に働くと考える。もしくは、$Precision_2(c)$ の式が、(ii) の基準を正確に数値化できていなかった可能性がある。この点については、今後改善を加える必要がある。

7　今後の課題

今後は、基準 (i),(ii) をより正確に数値化できる式の考案や並び替えを行うための条件の追加・変更等により、要約精度を向上を目指す。そのために考えられる方法として、本手法では共起語の取得する数 N を 50 で固定しているが、この N が最適となる数を求めることがあげられる。次に、今回は共起語として得られた語は一律同じ重みで扱ってきたが、共起頻度や順位などにより重みを加えることも考えられる。また、共起語はあくまで連想知識の一例として用いているので、この共起語をたとえば近藤ら [2] のように語釈文や、単語類似度により代用し、それぞれの結果の比較も行いたい。

また、現在のところ要約は 1 単語のみに限っている。しかし、その出力される単語が入力の意味をすべて補完できていない場合がある。そのため、複数の単語を同時に出力する要約手法も考えている。具体的には、$|RelatedWord(c) \cap InputWord|$ から洩れた入力の名詞、つまり出力候補から連想できない入力の名詞群に対して、同様の手法により再度要約を行う。このとき得られた単語も出力に加える、といった処理を想定している。これにより、出力候補から連想できない入力の意味を補完することができると考える。

8　おわりに

本稿では、連想知識として文内共起を用いることにより端的な要約を生成する手法を提案した。本手法では出力の候補が複数存在するため、その候補の中から要約にふさわしいものを選定する条件を用意し、計 5 通りの手法による出力を評価した。その結果、最も良い手法では lenient において MRR 値が 0.447 となった。また同様の手法により、10 位以内に正解といえる要約を 75%の精度で取得することができた。

今後の課題としては、並び替えの結果、正解となる語がより上位になるように条件を追加・変更する必要がある。また、1 単語のみに要約を行うのではなく、複数の単語を要約として表示し、より入力の意味を保持した要約を生成する必要がある。

使用した言語資源及びツール

(1) 黒橋, 河原. 日本語形態素解析システム JUMAN version 5.1, 京都大学, 2005.

参考文献

[1] 奥村学, 難波英嗣. テキスト自動要約. オーム社, 2005.

[2] 近藤恵子, 奥村学. 言い替えを使用した要約の手法. 情報処理学会研究報告 NL-116-20, pp137-142, 1996.

[3] M. Banko, V.Mittal, M. Witbrock. Headline Generation Based on Statistical Translation. In Pro-ceedings of 38th Meeting of Association for Computation Linguistics, Hong Kong, pp.218-325, 2000.

[4] 岡崎直観, 松尾豊, 石塚満. 関連する複数新聞記事からの重要文抽出法. 第 3 回 AI 若手の集い MYCOM2002, pp80-86, 2002.

[5] 谷川信弘, 砂山渡. テキストの結論重視型要約の生成. 第 23 回人工知能学会全国大会, 1B4-1, 2009.