要約長，文長，文数制約付きニュース記事要約

田中 駿 † 笹野 遼平 †† 高村 大也 †† 奥村 学 ††
東京工業大学 総合理工学研究科 † 東京工業大学 精密工学研究所 ††
shun@lr.pi.titech.ac.jp, {sasano, takamura, oku}@pi.titech.ac.jp

1 はじめに

従来から使われてきた新聞やテレビなどを介したニュース配信に加え，インターネットを介したニュース配信が増加しており，スマートデバイス等の普及なども相まって，表示領域に制限のあるデバイスでニュースを閲覧する機会が増えている．図 1 のように，1 文が 1 行におさまる数文から成る要約は，限定された表示領域でも見やすいと考えられ，実際にこのような形式でニュース記事を要約し，掲載しているウェブサイトも存在している [1]．そこで本研究では，日本語のニュース記事を，合計 100 字以内，各文が 20 字から 40 字で構成される 3 文で要約するタスクに取り組む．

一般に広く用いられる要約手法として，文選択と文圧縮を組み合わせたものがある [3, 7, 12] が，表示領域を考慮した要約記事を生成する場合，要約記事全体での文字長だけでなく，要約文 1 文あたりの文長も考慮に入れる必要があるため，文選択と文圧縮だけでは不十分であると考えられる．たとえば，各文の文長を揃えた要約文を生成する場合，図 1 中の記事本文 5 文目，6 文目のように，短文であるが重要な情報を含む文が複数ある場合は，それらを融合した要約文を出力する文融合を行うことが望ましいと考えられる．また，多くの情報を含む長文がある場合は，それを圧縮するだけでなく複数の文に分割し出力する文分割を行うことが望ましいと考えられる．

そこで本研究では，文選択，文圧縮に加え，文融合および文分割手法を取り入れた要約手法を提案する．入力文に応じて適切に文融合や文分割を行うことで，表示領域に強い制限のあるデバイスでの閲覧を考慮したニュース記事要約が可能になると考えられる．

2 関連研究

要約において文融合を行った研究として，Filippova ら [4] や田中ら [10] のものがある．Filippova らは，複数文書要約において係り受け木を拡張することによって文融合を行う手法を提案し，ドイツ語および英語のデータに対して実験を行った．田中らは，要約者がニュース記事の要約を容易に行えるよう，日本語ニュース記事のリード文に対して，記事中の関連した文を融合することを目的に，係り受け木を基にした文融合手法を適用した．これらの研究は，同一の語を含む文節に対する文融合手法であるのに対し，本研究ではそれだけでなく，連続する短文に対する文融合や連体修飾化を伴う文融合も行う．

[1] 人手で要約を行っているサービスとして livedoorNEWS (http://news.livedoor.com/)，自動要約を行っているサービスとして Vingow(https://vingow.com/) などがある．

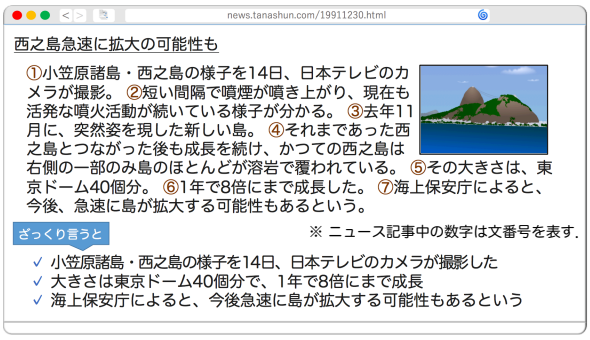

図 1: ニュース記事 (上) と要約記事 (下) の例

日本語を対象に文分割を行った研究としては，武石ら [14]，江原ら [15] の研究がある．武石らは複文を対象とし，南ら [13] の従属節分類を参考に従属節の独立性に着目することによって文分割が可能となる点 (分割点) を決定した．具体的には，入力文 1 文に対して係り受け解析を行った結果を元に，分割点の候補を求めた後，すべての分割点に対して，主語，文末表現および接続詞等の補完を適宜行い，最も適していると判断された分割点において文分割を行う．江原らは，ニュース番組の字幕生成を行う目的での文分割手法を提案した．この手法では，分割後の接続詞の補完や言い換えを含め，多くの規則を用いている．これに対し本研究では，非文法的な文の出力抑制には，後述する容認度 [5] を用いることとし，少量の分割規則のみを用いて文分割を行う．また，要約記事は箇条書きで出力され，要約文はそれぞれ独立した文となるため接続詞の補完や言い換えに関する規則は設けない．

3 提案手法

図 2 に提案手法の概要を示す．まず，入力記事が与えられたとき，各文に対して文融合処理または文分割処理を実行する条件に該当するかを判断し，条件に該当した場合，文分割，文融合を行い，それぞれの処理により生成された文と，入力記事中の各文を合わせた圧縮前候補文を列挙する．次に各圧縮前候補文に対し文圧縮処理を行い，要約候補文を列挙する．最後に，列挙された要約候補文から，単語の重要度や文の容認度を手掛かりとし，要約として適切な 3 文を選択し，最終的な要約記事として出力する．この際，ある圧縮前候補文から生成された要約候補文集合からは 1 つしか文を選ばないという制約を課している．

3.1 文融合および文分割処理

文融合では，対象となる 2 文の係り受け木を融合した，拡張された係り受け木を生成する．文融合対象の 2 文のうち，入力記事において先に出現する文を第 1 文，

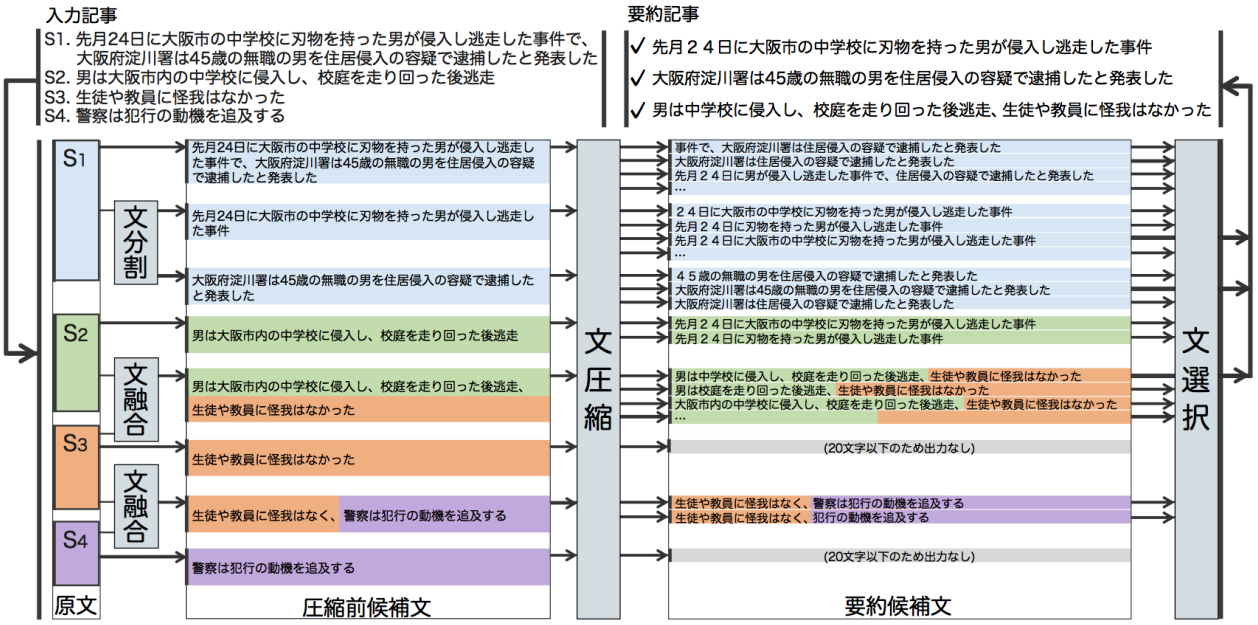

図 2: 要約生成までの流れ

表 1: 文融合・文分割処理を行う条件

処理	条件
文融合 1	記事中で連続している文である 2 文の文字長合計が 40 文字以下である 2 文とも文末が発話で終わっていない
文融合 2	2 文の文末が一致する 文末の述語の主語が 2 文で一致している
文融合 3	記事中で連続している文である 文末の述語の主語が 2 文で一致している
文分割	分割点直前の文節を除く前方の文節が後方の文節に係らない 分割点直前の文節が南ら [13] の従属節分類Aに該当しない 分割点直前の文節が格助詞に相当する句を含んでいない

後に出現する文を第 2 文と呼ぶこととする．また，文融合処理によって生成された文を融合文と呼ぶこととする．本研究で扱う文融合には後述する 3 つのタイプがあり，それぞれ，表 1 に示す条件に該当した場合に処理が行われる．また，文分割では，入力記事中の 1 文を複数文に分割することにより，入力記事中の 1 文から複数の圧縮前候補文を生成する．分割後の文のペアにおいて，文中の語が入力文の前半に出現するものを前方，後半に出現するものを後方と呼ぶこととする．各処理の詳細を以下にまとめる．

文融合 1　記事中で連続する 2 文であり，合計文字長が 40 文字以下の場合，2 文の係り受け木を結合し，融合を行う．この際，自然な融合文を生成するため，第 1 文の文末が用言である場合には，文末を連用形に変換し読点を付加，文末が体言である場合には文末に読点を付加した後，第 2 文の文末に係るように係り受け木を生成する．図 2 で行っている文融合はいずれもこのタイプに該当する．

文融合 2　同一の文末を持つ 2 文の係り受け木を結合し融合する．文末から係り受け木をたどり，文末から枝が繋がっている状態でどこまで 2 文が一致するかを調べる．そして 2 文で一致する文節を融合し，拡張された係り受け木を得る．この処理により，たとえば，「海中からナイフが見つかった」と「犯行に使ったナイフが見つかった」という 2 文から「海中から犯行に使ったナイフが見つかった」という融合文が生成される．

文融合 3　連体修飾化を行い文を生成することが可能な 2 文に対して融合を行う．第 1 文の主格を連体形に変換し，主格に係る文節を含め，第 1 文の文末と文末に係る文節を入れ替える．第 1 文の主格を第 2 文の文末に係るように係り受け木を生成し，融合文を得る．この処理により，たとえば，「ナイフが見つかった」と「ナイフは犯行に使ったとみられる」という 2 文から「見つかったナイフは犯行に使ったとみられる」という融合文が生成される．

文分割　入力文中のある連用節が，表 1 に示す文分割を行う条件全てに該当した場合に，その連用節の直後を分割点とし，分割点の位置で文を分割する．この際，前半の文末は終止形に変換し出力する．また，人手で要約文を作成する場合，要約文の末尾を「事件」，「問題」などといった体言で終わらせる場合がある．このような形式は要約において非常に有効であると考えられることから，本研究では，分割点として連用節の直後に加え，「事件」，「問題」という句が含まれる文節の直後も対象とする．図 2 で行っている文分割はこのタイプに該当する．

3.2　文圧縮

各圧縮前候補文に対し，係り受け木に基づいた文圧縮を行う．具体的には，係り受け木の根を含む部分木のうち，制約を満たす文を全て要約候補文として列挙する．これらの制約は，高い文法性の確保と，文選択時の計算コストの低減のために導入している．圧縮後の文に対する制約を表 2 に示す．

3.3　文選択

文圧縮によって出力された要約候補文中から，要約として最も適切であると考えられる 3 文を選び出力する．本研究では，要約に含める単語の重要度と文の容認度を考慮した，最大被覆問題 [8] の拡張として以下のように定式化する．ここで，文の容認度とは，言語モデルの生起確率から求める文法性のスコアである．

表 2: 圧縮後の文に対する制約

1. “「” または “」” のどちらか一方だけが圧縮文中に存在しない
2. 圧縮文の文長が 20 文字以上 40 文字以下である
3. 元文に “A から B” という内容が含まれる場合, A，B のどちらか一方だけが圧縮文中に存在しない
4. 細かい範囲の時間を表す語 {“朝”,“未明”,“夕”,“深夜”,“下旬”,“上旬”, “中旬”,“分”,“午前”,“午後”} が圧縮文中にあり，元文に広い範囲の時間を表す語 {“日”,“時”,“月”,“年”,“今朝”,“同日”,“同月”,“同年”} が含まれている場合，広い範囲の時間を表す語も圧縮文に含まれていなければならない
5. 用言とその直前格は同時に圧縮文に含める
6. 体言とその直前格は同時に圧縮文に含める

$$max. \quad \alpha \sum_j w_j z_j + (1-\alpha) \sum_i g_i x_i \tag{1}$$

$$s.t. \quad \sum_i x_i = 3\ ;\ \forall i, \tag{2}$$

$$\sum_i l_i x_i \leq 100\ ;\ \forall i, \tag{3}$$

$$\sum_i a_{ij} x_i \geq z_j;\ \forall j. \tag{4}$$

x_i は要約候補文集合の i 番目の文を要約に採用する場合に 1 となる二値変数，z_j は単語 j が要約に採用される場合に 1 となる二値変数である．目的関数中の w_j は単語 j の重み定数であり，g_i は候補文 i に対する容認度の値である．目的関数の第一項は，要約に含める単語の重みの総和であり，第二項は，要約に含める候補文に対する容認度の総和である．また，α はトレードオフパラメータである．式 (2) は要約候補文集合から選ぶ文数を 3 文とする制約，式 (3) は要約記事の文字長を 100 文字以内とする制約である．a_{ij} は候補文 i 中に単語 j が含まれる場合に 1 となる定数であり，式 (4) は候補文と単語の整合性を保つための制約である．l_i は候補文 i の文字数である．また，より精度の高い要約記事を生成するため，式 (2) から (4) の制約式に加え，1 つの圧縮前候補文からは 1 文しか要約に採用できないという制約を課している．

4 実験

4.1 実験設定

実験に用いるデータとして，ニュース要約サイト livedoorNEWS より収集した 10,607 記事をデータセットとして用いた．このデータセットのうち，8,487 記事を訓練データ，1,060 記事を開発データ，1,060 記事を評価データとして使用した．

また，形態素解析器として JUMAN7.0 ，係り受け解析器として KNP 4.15 [11] を用いた．形態素解析の際に用いる辞書には，ニュース記事に含まれる多くの新語を解析できるようにする目的で，Wikipedia の見出し語を追加した．本実験では，要約生成のための整数計画ソルバとして，ILOG CPLEX を用いた．

単語の重み w_i は，Yih ら [9] の手法の中で単一文書要約に用いることのできる素性に加え，ニュース記事要約に特化した素性として，入力記事第 1 文に出現する単語かどうかの二値素性と，ニュースタイトルに出現する単語かどうかを示す二値素性を用いて，対数線形モデルより算出した．対数線形モデルの学習には LIBLINEAR バージョン 1.96 [2] を用い，L2 正則化項を付加，ハイパーパラメータは $C = 1$ とした．また，容認度は Lau らの研究 [5] に基づき，式 (5) で求めた．

$$g_i = \frac{\log P_m(s_i) - \log P_u(s_i)}{|s_i|} \tag{5}$$

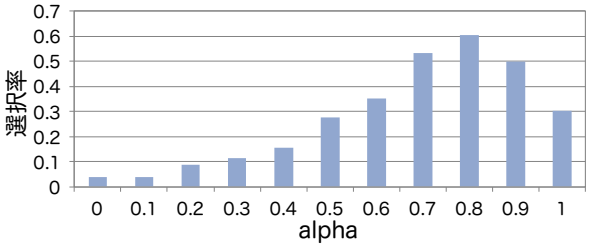

図 3: パラメータチューニングの結果

式 (5) 中の s_i は，i 番目の候補文である．$P_m(s_i)$ は，言語モデルによって出力される文 s_i の生起確率で，$P_u(s_i)$ は文 s_i に含まれる単語の生起確率の総和であり，g_i は，文の生起確率から全単語の生起確率の和を引いたものを単語数で割り正規化したものとなっている．本研究では Lau らに倣い，言語モデルとして Recurrent Neural Network Language Model [6][2] を用いた．言語モデルの学習にはインターネット上から無作為に収集した 500 万文を使用した．なお，収集したデータにおいて出現頻度が 1 回の単語に関しては未知語とすることによってスムージングを行った．

4.2 内容語と容認度の重みの決定

まず，文選択の際に用いるトレードオフパラメータ α を決定するために，クラウドソーシングサービスを用いてパラメータチューニングを行った．各記事に対して，ニュース記事本文と，ステップ幅を 0.1 とした 11 通りのパラメータから生成された要約記事を 3 つ提示した．被験者はニュース本文に目を通した上で，文法性や内容の充実度を総合的に判断し，3 つの要約記事の中でどれが最も良いかを選択した．5 人の被験者に同一の要約記事を提示し，3 人以上が良いと答えたパラメータを正しいとし，パラメータの値が選ばれる割合を計算し，評価に使用するパラメータを決定した．図 3 に異なる α による選択率の変化を示す．選択率は $\alpha = 0.8$ がピークとなっている．これは，文圧縮など原文に対する改変操作により生じうる文法性の低下に対し，容認度が有効に働くことを示している．パラメータチューニングの結果から，次節以降の実験では，$\alpha = 0.8$ を使用した [3]．

4.3 評価実験

最終的な要約記事の評価はクラウドソーシングを用いて行った．文融合および文分割，文圧縮を全て用いた提案手法を ALL と呼び，表 3 に示す 5 つの手法と比較することにより，提案手法の有効性を確認した．評価データ 1,060 記事の中からランダムに選択した 200 記事を対象に，比較手法それぞれにおける要約記事をアノテータに示し，DUC [1] における QualityQuestion を参考にした 5 つの観点から評価を行った．評価項目を表 4 に示す．各評価観点に対して，それぞれ 良い (5 点)・やや良い (4 点)・ふつう (3 点)・やや悪い (2 点)・悪い (1 点) で評価した．また，有意差検定として並べ替え検定を行った．

[2] http://rnnlm.org/, オプションとして-hidden 40 -rand-seed 1 -bptt 3 -class 200 で実行した．

[3] ただし，比較手法のうち，文圧縮制約を除いた手法においては，要約文の候補数が他の手法に比べ非常に増加するため別途パラメータチューニングを行い，$\alpha = 0.7$ で実験を行った．

表 3: 比較手法とその概要

比較手法	概要
-Fus.(-F)	提案手法から文融合を除いた手法
-Div.(-D)	提案手法から文分割を除いた手法
-CmpCnst.(-C)	提案手法から 20 文字以上 40 文字以下以外の文圧縮制約を除いた手法
-F-D-C	20 文字以上 40 文字以下の文圧縮制約のみを用いて文圧縮だけで要約生成を行う手法
HUMAN	livedoorNEWS の人手による要約記事

表 4: 評価実験における評価項目

評価項目	概要
内容の良さ	ニュース記事の要点をつかんだ要約記事になっているか
文法性	日本語として正しい要約記事か
冗長性の少なさ	同じような内容を何度も繰り返していないか
指示内容の明確さ	要約文中の名詞および代名詞が何を指しているのかはっきりと分かるか
総合	総合的な要約記事の良さ

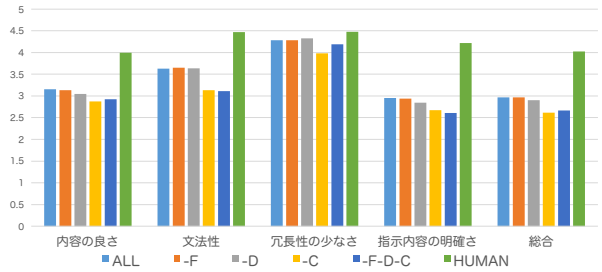

図 4: 評価項目毎の結果

評価実験の結果を図 4，表 5 に示す．また，参考のため ROUGE-2 のスコアを R2 として示す．実験の結果，HUMAN は全ての評価項目において高いスコアを出していることがわかった．HUMAN 以外の結果では，内容の良さの観点に関しては ALL が他の手法を有意に上回った．文法性に関しては-F, -D, 冗長性の少なさに関しては-Dと比べ，ALL のスコアが下回る結果となったが，統計的に有意な差ではなかった．総合の観点においては，ALL が最も高いスコアであったものの，有意な差は出なかった．

4.4 比較実験

提案手法における文融合，文分割，文圧縮制約の有用性の差をより詳しく調査するため，-F, -D, -C の 3 つの比較手法に対し追加の比較実験を行った．具体的には，比較手法それぞれに対し，提案手法 (ALL) の出力と比較手法の出力のペアを被験者に提示し，文法性および内容の充実度の 2 つの観点を総合的に考え，どちらの手法による要約記事が良いかを選択してもらった．有意差検定としてマクネマー検定を行った．比較実験の結果を表 6 に示す．

ALL vs -F では，統計的な差は確認できなかった．特に融合 2 では，拡張された係り受け木から文を生成する場合に非文が生成されてしまう場合が多いなどの問題があった．ALL vs -D では，ALL が選ばれた回数が 122 回で-D が選ばれた回数が 78 回，p 値が 0.0023 となり，ALL の有用性が示された．この理由としては，「事件」や「問題」で分割を行うことなどで，箇条書きとしての要約記事の読みやすさが向上したことなどが考えられる．ALL vs -C では，ALL が有意に良い要約を出力していることがわかった．

表 5: 評価実験の結果

手法	内容の良さ	文法性	冗長性の少なさ	指示内容の明確さ	総合	R2
ALL	3.153	3.631	4.286	2.950	2.970	0.319
-F	3.133*	3.647	4.284	2.941	2.964	0.314
-D	3.048*	3.638	4.330	2.844*	2.904*	0.308
-C	2.873*	3.135*	3.983*	2.671*	2.617*	0.286
-F-D-C	2.923*	3.113*	4.190	2.610*	2.665*	0.281
HUMAN	3.995	4.468	4.476	4.218	4.023	-

表中の*は ALL と，HUMAN を除く比較手法のスコアにおいて，検定の結果 p 値が 0.05 以下である項目を示している．

表 6: 比較実験において提案手法，比較手法が選ばれた回数

比較ペア	提示記事数	提案手法	比較手法	p 値
ALL vs -F	88	47	41	0.5943
ALL vs -D	200	122	78	0.0023
ALL vs -C	200	156	44	0.0000

5 まとめ

本研究では，表示領域に制限のあるデバイスでのニュース記事の要約表示を想定し，合計 100 字以内，各文が 20 字から 40 字で構成される 3 文で要約することを目的とした，文融合および文分割処理を取り入れた要約手法を提案した．実験の結果，文分割や文圧縮制約を用いることによって，限られた表示領域の中で要点をつかんだ要約生成を行うことが可能であることを確認できた．今後の課題としては，文融合の精度向上を図ることや，係り受け木を用いるだけでない，より柔軟な要約生成手法を検討することなどが挙げられる．

謝辞 本研究は JSPS 科研費 26280080 の助成を受けた．

参考文献

[1] *DUC : Document Understanding Conference, in HLT/NAACL Workshop on Text Summarization*, 2007.
[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.
[3] K. Filippova and M. Strube. Dependency tree based sentence compression. In *INLG*, pp. 25–32, 2008.
[4] K. Filippova and M. Strube. Sentence fusion via dependency graph compression. In *EMNLP*, pp. 177–185, 2008.
[5] J.-H. Lau, A. Clark, and S. Lappin. Unsupervised prediction of acceptability judgements. In *ACL*, pp. 1618–1628, 2015.
[6] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky. RNNLM-recurrent neural network language modeling toolkit. In *ASRU*, pp. 196–201, 2011.
[7] H. Morita, R. Sasano, H. Takamura, and M. Okumura. Subtree extractive summarization via submodular maximization. In *ACL*, pp. 1023–1032, 2013.
[8] H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *EACL*, pp. 781–789, 2009.
[9] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *IJCAI*, pp. 1776–1782, 2007.
[10] 田中英輝, 美野秀弥, 宮崎勝, 小早川健, 熊野正, 後藤淳, 加藤直人. 文融合法に基づいた放送ニュースリード文の具体化. 情報処理学会研究報告 NL-194, pp. 1–6, 2009.
[11] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学. 構文・述語項構造解析システム KNP の解析の流れと特徴. 言語処理学会第 19 回年次大会, pp. 110–113, 2013.
[12] 三上真, 増山繁, 中川聖一. ニュース番組における字幕生成のための文内短縮による要約. 自然言語処理, Vol. 6, No. 6, pp. 65–81, 1999.
[13] 南不二男. 現代日本語の構造. 大修館書店, 1974.
[14] 武石英二, 林良彦. 接続構造解析に基づく日本語複文の分割. 情報処理学会論文誌, Vol. 33, No. 5, pp. 652–663, 1992.
[15] 江原暉将, 福島孝博, 和田裕二, 白井克彦. 聴覚障害者向け字幕放送のためのニュース文自動短文分割. 情報処理学会研究報告 NL-138, pp. 17–22, 2000.