自動獲得した言語的パタンを用いた重要文抽出システム

野畑 周† 関根 聡‡ 井佐原 均† Ralph Grishman‡
nova@crl.go.jp sekine@cs.nyu.edu isahara@crl.go.jp grishman@cs.nyu.edu
†独立行政法人 通信総合研究所 けいはんな情報通信融合研究センター
‡Computer Science Department, New York University

1 はじめに

重要文抽出とは、与えられた文書から重要だと考えられる文を文章から抜き出して要約を作成する手法であり、自動要約の主な手法の一つとして用いられてきている ([1], [2], [3], [4])。重要文抽出システムは、文書中のどの文が重要であるかを判断するために、文の長さや文章中での位置、文中の単語の頻度などの評価尺度から文の重要度を見積もる。そして得られた重要度の高い文を与えられた長さになるまで出力し、要約を生成する。

我々は現在、文の位置・文の長さ・文中の語の頻度など複数の評価尺度を組み合わせて重要文抽出を行うシステムを開発している。今回、本システムに固有表現抽出とそれに基づく言語的パタンの自動獲得という、情報抽出で用いられる手法を導入した。情報抽出 (Information Extraction) は、新聞記事などの文書から特定の出来事に関する情報を取り出すタスクである。その下位タスクとして、文書中の人名・地名などの固有名詞、日付・金額などの数値表現を認識する固有表現抽出 (Named Entity Recognition) と呼ばれるタスクが定義されている [5]。我々は日本語新聞記事からの重要文抽出タスク Text Summarization Challenge(TSC)[6] において、固有表現抽出を重要文抽出システムの評価尺度の一つとして導入したが [7]、本研究では本システムを英語新聞記事に適用できるよう拡張し、さらに特定分野に頻出する文パタンを自動獲得し、獲得されたパタンと記事中の文との類似度を求める関数を評価尺度の一つとして導入した。

パタンの自動生成は、情報抽出の分野において研究されている手法の一つである [8], [9], [10]。パタンはここでは、分野ごとに取り出したい特定の出来事を記述する典型的な表現を示すものである。例えば「○×社の△社長が会長に就任」といった表現は人事異動に関するパタンの一つである。パタンに基づいて情報抽出を行うシステムでは、対象とする分野ごとにこのようなパタンを用意してやる必要があるが、それを人手でなく自動的に行う手法が研究されている。我々は、このようなパタンが、自動要約においても対象記事の分野が何らかの形で特定されている場合には、有効であると考えた。

以下では、まず重要文抽出システムの概要と、新たに導入した文パタンに基づく評価尺度について述べる。次いで本システムが参加した英語新聞記事の要約コンテスト Document Understanding Conference (DUC) について述べ、そこでの評価結果を報告する。

2 システムの概要

本節では、重要文抽出に用いたシステムが用いた評価関数と、それらの関数の値に対する重み付けの方法について述べる。

2.1 重要文の評価尺度

本システムでは、文の位置や文長などの各評価尺度について、さらに複数の関数を定義し、トレーニングデータを用いてそのうちの一つを選択する。各評価尺度の用い方を予め複数定義しておくことで、対象とする文書の性質に合うように評価尺度の用い方を変更できることを意図している。システムがもつ評価関数については、[7] で示したものと同一のもの (文の位置、文の長さ、見出し) については説明を簡潔にするために、各評価尺度において選択された関数についてのみ述べる。

2.1.1 文の位置

文の位置情報に基づく関数としては、記事の先頭の文に大きい値を与える関数が選択された。この関数

は、いわゆるLead-basedを示すものであり、n個の文を含む記事について、出力できる文の最大値$T(n > T)$が与えられているときに、文S_iが先頭からT文以内にあればスコア1を与え、そうでなければスコア0を与える。

$$Score_{\mathrm{pst}}(S_i)(1 \leq i \leq n) = 1(i < T \text{ のとき}) = 0(\text{それ以外})$$

システム中には、他に先頭か末尾のどちらかに近い文ほど重要である、という仮定に基づいた関数も定義してあるが、トレーニングデータを用いた実験では選択されなかった。

2.1.2 文の長さ

各文の長さに基づく関数では、各文の長さは単語数によって示されている。ここでは、極端に短い文は重要文として選択されることが非常に稀であるという観測事実に基いて、以下の関数が選択された。

$$Score_{\mathrm{len}}(S_i) = 0 \quad (L_i \geq C \text{ のとき}) = L_i - C \quad (\text{それ以外})$$

この関数は、長さ(L_i)が一定の単語数(C)より短い文にはペナルティとして負の値を与えるものである。他に、長い文にスコア値を与える関数も用意されていたが、選択されなかった。評価の際には、トレーニングデータを用いた実験から単語数の値Cを10に設定した。

2.1.3 tf*idf

この関数は、記事中の単語の頻度(tf)と、その単語がある文書集合の中で現れた文書の数(df)を用いてtf*idf値を計算し、文のスコア付けを行う。この関数を用いる意図は、「記事に特有な単語をより多く含む文は、その記事においてより重要だ」とみなす仮定に基づく。tf*idf値の定義としては、以下の3種類の定義を与え、トレーニングデータからシステムに選択させた:

$$\text{T1. tf*idf}(w) = tf(w)\log\frac{DN}{df(w)}$$

$$\text{T2. tf*idf}(w) = \frac{tf(w)\text{-}1}{tf(w)}\log\frac{DN}{df(w)}$$

$$\text{T3. tf*idf}(w) = \frac{tf(w)}{tf(w)+1}\log\frac{DN}{df(w)}$$

DNは与えられた文書集合中の文書数である。ここでは、Wall Street Journalの2年分の記事(1994, 1995)を文書集合として用いた。また、文のスコアは、以下の式のようにそれらの和によって与えられる:

$$Score_{\mathrm{tf*idf}}(S_i) = \sum_{w \in S_i} \text{tf*idf}(w)$$

トレーニングデータを用いた実験の結果、式T1が選択された。

2.1.4 見出し

この関数は、「見出しと類似している文は重要である」という仮定に基づき、各記事の見出しとの類似度を計算し、評価尺度として与える。この関数では、見出しに含まれる単語に対するtf*idf値を用いた類似度の定義と、見出しに含まれる固有表現に対するtf値を用いた類似度の定義2種類を用意した。トレーニングデータを用いた実験の結果、固有表現を用いた類似度が選択された。これは、文(S_i)中の全固有表現について、その固有表現が見出し(H)に含まれていれば、そのtf*idf値を文のスコアに加算するものである。文のスコアを与える式を以下に示す:

$$Score_{\mathrm{hl}}(S_i) = \frac{\sum_{e \in H \cap S_i} \frac{tf(e)}{tf(e)+1}}{\sum_{e \in H} \frac{tf(e)}{tf(e)+1}}$$

2.2 言語的パタン

本節では、言語的パタンの獲得方法とそれを用いた評価尺度について述べる。情報抽出におけるパタンの自動獲得は、例えば、地震の発生を報道する記事には、「○月×日x時y分ごろ、△□で地震があった」といった表現がよく現われるように、「分野(domain)を特定したときに文書によく現れる表現は情報抽出において重要だ」という仮定に基づいている。

DUCにおいては、約10記事ずつを1セットとして30記事セットのデータが配布された。この各記事セットを情報抽出における一つの分野とみなし、各セットごとにパタンの自動獲得を行った。ここで用いているパタンの獲得手法は、日本語情報抽出において提案された手法に基づいている[10]。パタンの獲得方法は以下の過程に従って行われる:

1. 文の解析:
   与えられた記事セット中の記事全文について品詞・固有表現のタグづけ、係り受け解析を行う。

2. 部分木の抽出：
   係り受け木中の部分木を全て取り出す。

3. 固有表現による抽象化：
   部分木中に固有表現があった場合には、その固有表現を対応するクラスに置き換えたものと、元の表現のままの二通りの部分木をパタンとして用意する。複数の固有表現がある場合は、各置換の組み合わせだけの部分木を生成する。

4. 部分木のスコア付け：
   木全体の頻度と、部分木中の各単語の idf 値を掛け合わせたものを部分木のスコアとして求める。このスコアの定義は、その記事セットに特有な部分木を取り出すという意図に基づいており、スコアが高い部分木ほど重要なパタンであると仮定することになる。

パタンは重要文抽出を行う前に取り出され、スコアとともにシステムに格納される。実際に重要文抽出を行うときには、システムは各文 S_i について品詞・固有表現のタグづけ、係り受け解析を行って係り受け木を作成し、次いで格納されたパタンとの比較を行う。あるパタン P_j が文 S_i の係り受け木の一部と一致した場合には、そのパタンのスコアが文の評価尺度として加算される。一致した全パタンのスコアを加算し、その値の対数をとったものを最終的な文の評価尺度としている。パタンのスコアと文の評価尺度をそれぞれ式に表わすと以下のようになる：

$$\begin{aligned} PatScore(S_i) &= \sum_j F_{P_j} \frac{\sum_{w \in P_j} \log \frac{DN}{df(w)}}{|P_j|} \\ &\quad (P_j \text{ が } S_i \text{の一部に一致}) \\ &= 0 \text{ (それ以外)} \\ Score_{\text{pat}}(S_i) &= \log(PatScore(S_i) + 1) \end{aligned}$$

ここで、F_{P_j} はパタン P_j の記事セット中の頻度、$|P_j|$ は P_j 中の単語数を示す。

2.3 重み付け

本システムでは、先に述べた各評価関数の値の和を文の重要度とする。各評価関数（$Score_j$）の値には重み付け（α_j）を与え、それらの和が各文（S_i）の重要度となる。

$$\text{Total-Score}(S_i) = \sum_j \alpha_j Score_j(S_i)$$

重みの値は、トレーニングデータから求めた。各重みの値域を予め定めておき、その値域内でトレーニングデータに対する重要文抽出の評価を繰り返し行い、最適な結果を与える重みの値を記録した。

表 1: 各評価尺度の貢献度

評価尺度	重み× S.D.
文の位置	277
文の長さ	8
tf*idf	96
見出し	18
パタン	2

各評価尺度がどの程度結果に寄与しているかをみるために、表1に各評価尺度の標準偏差と、それに対する重みを掛け合わせたものを示した。最も値の大きい評価尺度は「文の位置」であり、次いで「tf*idf 値」であった。見出しや文パタンに基く評価尺度は、それらに比較して結果に寄与する割合が小さかった。

3 実験と結果

本節では、本システムが参加した英語新聞記事の要約コンテスト Document Understanding Conference (DUC) の課題と評価方法について説明し、本システムの評価結果を報告する。

3.1 DUC

Document Understanding Conference(DUC) は、米国 DARPA(ARDA) の支援の下に National Institute of Standards and Technology(NIST) によって実施されている、自動要約の評価コンテストである [11]。DUC2001 では、単一文書の要約と複数文書の要約の 2 種類の課題が出された。対象とするデータは両課題において共通であり、トレーニングデータとして 30 記事セット、テストデータとして新たに 30 記事セットが主催者から配布された。各記事セットには約 10 記事ずつ含まれており、AP 通信や Financial Times, Los Angels Times、 Wall Streat Jounal などの新聞から取られている。各記事セットには、コンテスト実施時には明示されていなかったが、例えば「最高裁判事に任命されたトーマス氏についての記事」「ピナツボ火山の噴火についての記事」などの主題ごとに集められた記事から成っている。単一文書の要約では、各記事を 100 語以内に要約して出力することが課された。複数文書の要約では、各記事セットごとに 50 語、100 語、200 語、400 語の 4 種類の要約が課された。

表 2: 要約結果の主観評価 (被験者の評価の平均)

	本システム (順位)	ベースライン (Lead-based)	全システムの平均
文法性	3.711 (5)	3.236	3.580
結束性	3.054 (1)	2.926	2.676
一貫性	3.215 (1)	3.081	2.870
全体	9.980 (1)	9.243	9.126

我々はDUCの単一文書要約課題にのみ参加したので、以下では、DUC単一文書要約課題の評価方法と、我々のシステムの評価結果について述べる。

3.2 DUCでの評価結果

DUCにおける要約結果の評価は、被験者がシステムによって生成された要約を人間が作成した要約と比較して判定する、主観評価によって行なわれた。主観評価は、Grammaticality(文法性)、Cohesion(結束性)、Organization/coherence(一貫性)の3つの基準について行われた。10人の被験者が各基準について5段階評価(4が最も高く、0が最も低い)を行った。各被験者の評価結果の平均を表2に示す。全システムの結果は、参加した11システムとベースラインの結果の平均値である。ベースラインの要約は、全ての記事について、先頭から100語出力したものである。

本システムの結果は、どの評価基準においてもベースライン、全システムの平均を上回っている。また、システム全体での順位も、文法性では5位であったが、それ以外の評価では1位であり、全体でも1位であった。

4 おわりに

我々は、重要文抽出システムに、記事セットから自動獲得される言語的パタンを用いた新たな重要文の評価尺度を導入した。我々のシステムは、英語新聞記事の要約を行うコンテストDUCに参加し、単一文書の要約課題において一位の成績をおさめた。今後の課題としては、言語的パタンをより有効性の高いものとするために、情報検索の手法を用いて、要約対象として与えられた記事群より大規模な記事群を用意し、その記事群から言語的パタンを獲得することを行いたいと考えている。また複数文書の要約においても、このような言語的パタンを用いた手法の有効性を確かめたいと考えている。

参考文献

[1] H. Edmundson. New methods in automatic abstracting. *Journal of ACM*, Vol. 16, No. 2, pp. 264–285, 1969.

[2] 野本忠司, 松本祐治. 人間の重要文判定に基づいた自動要約の試み. In *IPSJ-NL 120-11*, pp. 71–76, July 1997.

[3] C. Aone, M. E. Okurowski, and J. Gorlinsky. Trainable, Scalable Summarization Using Robust NLP and Machine Learning. In *Proc. of COLING-ACL'98*, pp. 62–66, 1998.

[4] Chin-Yew Lin. Training a selection function for extraction. In *Proc. of the CIKM'99*, 1999.

[5] DARPA. *Proceedings of the Sixth Message Understanding Conference(MUC-6)*, Columbia, MD, USA, November 1995. Morgan Kaufmann.

[6] TSC. http://oku-gw.pi.titech.ac.jp/tsc/, 2001. Text Summarization Challenge.

[7] 野畑周, 関根聡, 村田真樹, 内元清貴, 内山将夫, 井佐原均. 複数の評価尺度を統合的に用いた重要文抽出システム. 言語処理学会 第7回年次大会, March 2000.

[8] Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044–1049, 1996.

[9] Roman Yangarber, Ralgh Grishman, and Pasi Tapanainen. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP2000)*, 2000.

[10] Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. Automatic pattern acquisition for japanese information extraction. In *Proceedings of Human Language Technology Conference*, San Diego, California, USA, 2001.

[11] DUC. http://www-nlpir.nist.gov/projects/duc/, 2001. Document Understanding Conference.