n-gramモデルとIDFを利用した統計的日本語文短縮

堀之内寛　山本幹雄
筑波大学

1. はじめに

テキスト自動要約のための文短縮手法について検討を行った。ここでいう文短縮とは、与えられた一文内の一部を削除することによる短縮で、言い換え等は含まない。要約のために短縮された文は、「日本語らしく、かつ意味的に重要箇所を含む」ことが望まれる。ここで、日本語らしさの評価のために統計的言語モデル(n-gramモデル)、意味的な重要箇所の評価として情報検索でキーワード重みとして用いられるidf (tf*idf)を同時に利用するのが本稿の提案である。

統計的言語モデルは、大量の日本語コーパスから「日本語らしさ」を学習する。ある文が与えられ、その一部を削除した短縮文の候補が複数あるとき、統計的言語モデルを使えばどの候補が日本語文として自然であるかを評価できる。しかし、日本語らしさだけで削除部分を決定すると、日本語としては自然だが、意味のない部分のみが残る可能性がある。

tf*idfは情報検索におけるVector-Spaceモデルでよく使われる指標で[Salton83]、idfは少ない文書にしか出現しない単語、すなわち文書を特定する能力の高い(意味的に重要な)単語に大きな値を付与するものである。tf*idfは重要文抽出のための重みとしても使用され[Zechner96]、有効であることが知られている。しかし、一般に機能語はtf*idfが小さい値になるため、tf*idfが小さい部分だけを削除していくと、読みづらい文、最悪の場合は意味不明の文になる可能性がある。

「日本語らしさ」と「意味的な重要度」の２つの指標を適度にミックスすることによって、上記の「日本語らしく、かつ意味を持った」文に短縮できると考えた。

2. 統計的文短縮手法

2.1 削除部分の探索と削除単位

すべての削除候補を同時に評価・決定するのは計算量の観点から困難である。長さｎ文字の入力文に対して、任意の部分を任意の数だけ削除した短縮文候補の数は、各文字を残すか残さないかであるため、2^nと膨大な数になる。ここでは、近似的な方法として、１度の短縮で１つの連続する部分を削除し、これを繰り返すことにより任意の短縮率を得る方法を採用した。（具体的な削除過程の例は3.2～3.4節の実験結果を参照。）

一度に削除する単位としては以下のようなバリエーションを検討した。

- 一つの文節
- 連続する任意の形態素列
- 連続する任意の文字列

文節、形態素列を単位とするためには文節または形態素の解析が必要である。形態素の解析にはJUMAN3.2[松本他97]、文節の認定にはKNP2.0 b3[黒橋97]の解析結果を用いた。連続する形態素列(文字列)を削除単位とする場合は、すべての形態素位置(文字位置)ではじまる最大3個(15個)の形態素(文字)列を削除候補とした。任意の文字列を削除単位とする方法は人間の与える言語知識を全く用いないので精度に限界はあるが、逆に言語知識を用いていないため、保守や頑健性において優れていると言える。

2.2 削除部分の評価基準

前節で述べた短縮文探索アルゴリズムの一回の繰り返しで削除する部分は、以下の重要度の値が最低になる部分とした。引数のsenは削除前の文、partはsen中の削除候補部分である。

重要度(sen, part)＝Perp(sen, part) +a* Weight(part)

Perp(sen, part)はpartを削除した文の日本語らしさ、Weight(part)はpartの意味的重要度、係数aは2項のバランスを取るためのものである。Perp(sen, part)は、senからpartを削除した文に対する、n-gramモデルから算出された文字perplexityである。文字perplexityは文の生起確率を文字あたりに平均し、その逆数を取った値である。すなわち、値が低いほど日本語として出現する確率が高いことになり、対象としている部分を削除した場合の文が日本語らしいことになる。単位を文字とした理由は、形態素に対してトラ

イグラム以上のモデルを求めることが困難であることと、削除単位にかかわらず同じモデルを使用したかったためである。

Weight(part)はpartの意味的な重要度であり、削除する単位の違いによって以下のように使い分けた。

文節：Part中の自立語のtf*idf値の平均
形態素列：Part中の形態素のtf*idf値の合計
文字列：Part中の任意の部分のridf値の最大値

ridfはコーパス全体の出現頻度(tf)からPoissonモデルで推定されるidfと実際のidfの差である[Church & Gale 95]。文節と形態素列の場合のtf*idf値のtfは要約対象の文が含まれる記事から算出するため、記事全体の情報を使用しているが、ridfの計算では使用しない。Weight関数の包括的な検討は今後の課題である。

3. 実験

3.1 モデルとシステム

統計的言語モデルは、毎日新聞CD-ROM版91年版[毎日新聞]を訓練データとし、CMU-Cambrigde SM/tk [Rosenfeld95]を用いてBack-off スムージングされた5-gramモデルを推定した。このモデルを削除単位に関係なく共通に使用した。

Weight値としてのtf*idfを計算するために、形態素あるいは文字列に対するdf(document frequency)と記事内tf(term frequency)、ridfを計算するためにはコーパス全体におけるtfがさらに必要である。記事内tfは短縮対象の文を含む記事から実行時に毎回計算を行った。dfとコーパス全体に対するtfに関しては、コーパスのSuffix Arrayからすべての部分文字列クラスのtf とdfを求め[Yamamoto&Church98]、これを表として計算を行った。コーパスとしては、毎日新聞CD-ROM版91～93年の3年分のデータを用いた。

3.2 Perplexityのみの基準による文短縮結果

まず、統計的言語モデルの能力を確認するため、perplexityのみの基準による文短縮を示す。図1 (a)は、重要度のweight係数aを0とし(すなわちperplexiyのみの基準)、削除単位を任意の部分文字列とした場合の各繰り返しにおける短縮結果である。下線部分が次の繰り返しにおいて削除された部分である。

削除単位が任意の文字列であり、辞書等の言語知識を使用していないにもかかわらず、削除結果は日本語としておおむね妥当である。しかし、削除する部分は意味的な重要度には関係ないため、「メダル」というこの文で最も重要なキーワードを早い段階（3回目）で削除していることが分かる。「メダル」を削除した際にこれに付属する助詞「を」を残したため、「を」が「委員会」に付属してしまい意味が変ってしまっている。意味が変わると致命的なアプリケーションでは文節を削除単位とすれば意味が大きく変わることは生じにくい。

3.3 Weightのみの基準による文短縮結果

次にweightのみの基準による文短縮結果を図1 (b)に示す。削除単位は形態素である。これは、重要度のWeight項の係数aを大きくしたものに等しい。この場合、日本語らしさの知識を使っていないので、削除単位を任意の部分文字列にすると、意味不明の文になりやすい。

例を見ると、機能語のtf*idf値は低いため、徹底して機能語を削除している。このため、重要部分は確実に残っており、機能語をすべて削除した段階では記事のタイトルのようである。しかし、機能語の長さは一般に短いので短縮率は低い。さらに削除を進めると「公開」というこの文にとって重要な形態素が早い段階で削除されている。この点、前節のperplexityのみを使った基準が述部を削除するのは、短縮が進んでからであるのと対照的である。また、途中で「八日」の「日」のみを削除して残された「八」が日本語として意味不明となっている。

3.4 重要度(Perp+Weight)基準による文短縮結果

前節2つの基準は、それぞれ一長一短があることが分かったが、これらはお互いに補いあうことができる性質である。perplexityだけの場合、早い段階で削除された「メダル」をtf*idf項が削除されるのを防ぎ、tf*idfだけの場合に意味不明の部分が生じるのをperplexityが防ぐ。図1 (c)に2つの基準を合わせた重要度で削除部分を決めた文短縮例を示す。削除単位は任意の形態素列である。6回の削除で生成された「ノルウェー・オリンピック委員会メダルを公開。」は図1の中で最もよい要約ではなかろうか？

4. 評価

人間が行った文短縮を正解データとして、文字レベルでの一致度による評価を行った。正解データとして、毎日新聞CD-ROM版1994年版から選んだ14の記事に対して5人の被験者が行った要約を用いた。被験者には、まず記事から重要と思われる文を全体の約3割選び、それらの各文に対して不必要と思われる部分を削除してもらった。

2人以上が重要文であると選択した文を対象に、2人以上が削除した部分を正解削除部分と認定した。このデータを観察すると、文の前半部分を削除している場合が非常に多いことが分かった。日本語

（ａ）perplexityのみの基準による文短縮結果例（削除単位は任意の文字列）
ノルウェー・オリンピック委員会は八日、リレハンメル五輪の各メダルを公開した。
ノルウェー・オリンピック委員会は八日、各メダルを公開した。
ノルウェー・オリンピック委員会メダルを公開した。
ノルウェー・オリンピック委員会を公開した。
オリンピック委員会を公開した。
← 意味が変わってしまっている

（ｂ）weightのみの基準による文短縮結果例（削除単位は形態素）
ノルウェー・オリンピック委員会は八日、リレハンメル五輪の各メダルを公開した。
ノルウェー・オリンピック委員会は八日リレハンメル五輪の各メダルを公開した。
⋮
ノルウェー・オリンピック委員会八リレハンメル五輪各メダル公開した
← 意味不明
⋮
ノルウェーオリンピック委員会リレハンメル五輪メダル公開
ノルウェーオリンピックリレハンメル五輪メダル公開
ノルウェーオリンピックリレハンメル五輪メダル
← 意味的にはよいが、短縮率が低い

（ｃ）重要度(perplexity+weight)基準による文短縮結果例（削除単位は任意の形態素列）
ノルウェー・オリンピック委員会は八日、リレハンメル五輪の各メダルを公開した。
ノルウェー・オリンピック委員会は八日、リレハンメル五輪メダルを公開した。
ノルウェー・オリンピック委員会は、リレハンメル五輪メダルを公開した。
ノルウェー・オリンピック委員会はリレハンメル五輪メダルを公開した。
ノルウェー・オリンピック委員会は五輪メダルを公開した。
ノルウェー・オリンピック委員会メダルを公開した。
ノルウェー・オリンピック委員会メダルを公開。
← 図１の中で最適？
ノルウェー・オリンピック委員会メダルを。

図１　文短縮結果例

の新聞記事の場合、主節が最後でその前に複数の長めの従属節がくることが多く、これらの従属節をまとめて削除している傾向が強かったためである。このヒューリスティックスはかなり有効であるが、節全体の削除は重要文選択に近い手法が必要と考えられ（[福島他99]の方法が近い）、本手法の対象とは少しずれているため、節全体を削除している部分は評価対象から除いた。結果として評価用として用いた文は、合計40文（述べ129文・人）で、元の文の長さに対して平均で65.6%に短縮されている。

評価指標としては、適合率と再現率を用いた[三上他98]。ただし、削除単位が異なっても同じ評価を行いたいため、文字を単位として計算した。

$$\text{適合率} = \frac{\text{システムの要約中の正解文字数}}{\text{システムの要約の文字数}}$$

$$\text{再現率} = \frac{\text{システムの要約中の正解文字数}}{\text{正解要約の文字数}}$$

短縮率を正解データと一致させた場合の各手法の適合率と再現率を表１に示す。perplexityとtf*idfの組み合わせは、それぞれ単独で使用するよりも改善されていることが分かる。また、Weightのみの基準以外は削除単位を文字列よりも形態素列、形態素列よりも文節とした方がよい結果となっている。Weightのみの基準の場合は逆の優劣となっているが、これは削除単位によってWeightの計算方法が異なるためそれが直接影響したためと思われる。「左から削除」のヒューリスティックスは、非常によい結果となっている。これは、単文内でも日本語の場合、文末が重要であるということである。

さらに詳しく比較するために、適合率と再現率の関係をプロットしたグラフを図2に示す(削除単位を文節にした場合)。１つの文節を削除するごとに適合率と再現率を計算し、一定の範囲の再現率と組みになっている適合率の値を集めて平均を取ったものである。図2より、perplexityとtf*idfの組み合わせは、再現率を変化させた場合でも、それぞれ単独で使用するよりも一貫して改善されていることが分かる。「左から削除」のヒューリスティックスはここでも

表1 各手法の適合率と再現率(%)
上段が適合率、括弧の中が再現率
短縮率は正解データに合わせた場合

手法＼削除単位	文節	形態素列	文字列
Perplexityのみ	72.2 (66.1)	70.7 (64.7)	68.2 (62.6)
Weightのみ	64.4 (59.2)	66.1 (64.1)	66.7 (65.5)
Perplexity+ Weight	74.9 (68.4)	74.4 (71.4)	69.6 (66.4)
左から削除	74.5 (67.0)	---	---
ランダム	61.8 (56.5)	---	---

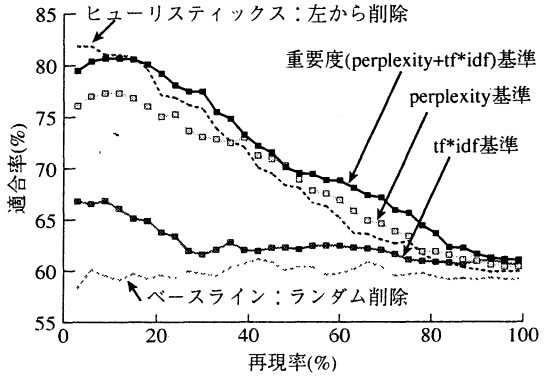

図2 適合率=再現率グラフ
削除単位：文節

強力であるが、現実的な意味で有用な再現率の高い部分で提案手法に対して2～5%程度適合率が低いことが分かる。

一般的に、要約システムの数量的な評価は困難であり、本節のように一つの指標だけの評価から一般的な結論は得られない。人間による要約文の適切さの評価[三上他99]等を今後行う必要がある。

5. おわりに

本稿では、統計的言語モデルを「日本語らしさ」、tf*idfを「部分の意味的な重要度」の基準として用いた文短縮手法の提案と評価を行った。本手法の特徴は、短縮のためのほとんどの知識をコーパスから自動的に得ることができるため、保守性と頑健性の点で期待できることである。信頼性と正確さにおいてまだ問題もあるが、indicativeな要約システムの一部として広い範囲で適用可能であると考えている。

謝辞

言語データとして毎日新聞CD-ROM版、形態素解析・文節解析には京都大学のJUMAN3.2, KNP2.0を使用させていただきました。関係者各位に感謝いたします。

参考文献

[黒橋97] 黒橋禎夫：「日本語構文解析システムKNP version 2.0 b3 使用説明書」, 1997.

[松本他97] 松本, 黒橋, 山地, 妙木, 長尾：「日本語形態素解析システム JUMAN version 3.2」, 1997.

[福島他99] 福島, 江原, 白井：「短文分割の自動要約への効果」, 自然言語処理, Vol. 6, No. 6, pp.131-147, 1999.

[毎日新聞] 毎日新聞社：「毎日新聞CD-ROM版1991年～1994年版」, 日外アソシエーツ.

[三上他98] 三上, 山崎, 増山, 中川：「文中の重要部抽出と言い換えを併用した聴覚障害者用字幕生成のためのニュース文要約」, 言語処理学会第4回年次大会ワークショップ論文集, pp.14-21, 1998.

[三上他99] 三上, 増山, 中川：「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.

[Church & Gale 95] Church, K. and W. Gale. Poisson mixtures, Natural Language Engineering, 1(2), pp.163-190, 1995.

[Rosenfeld95] R. Rosenfeld. The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation, Proc. ARPA Spoken Language Technlogy Workshop, pp.47-50, 1995.

[Salton83] G. Salton and M.J. McGill. The SMART and SIRE Experimental Retrieval Systems, pp.118-155, New York: McGraw-Hill, 1983.

[Yamamoto&Church 98] M. Yamamoto and K.Church. Using suffix arrays to compute term frequency and docuement frequency for all substrings in a corpus, In proceedings of 6th Workshop on Very Large Corpora, Ed. Eugene Charniak, Montreal, pp.28-37, 1998.

[Zechner96] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In Proc. of COLING96, pp.986-989, 1996.