自然言語処理 126－9
(1998. 7. 24)

短文分割を利用したテレビ字幕用自動要約

若尾 孝博
通信・放送機構 (TAO)
渋谷上原リサーチセンター
wakao@shibuya.tao.or.jp

江原 暉将
ＮＨＫ／ＴＡＯ

白井 克彦
早稲田大学／ＴＡＯ

あらまし

本研究の題材であるテレビニュース番組の電子化原稿は、１記事中の文数が多くなく、１文当たりの文字数が多いという特徴がある。この為、自動要約として重要文抽出を行うと、情報が文単位で取捨選択され、粗い要約となりがちである。そこで、本研究では、長文を分割出来る条件を設定し、条件に合う場合は、短い文に分割するという処理を行ってから、重要文抽出を行った。そして、短文分割を行う場合と行わない場合での重要文抽出について検討を行った。

キーワード　重要文抽出、短文分割、自動要約、ＴＶニュース原稿

Partitioning long sentences: how useful it is for sentence extraction

Takahiro Wakao
TAO of Japan
wakao@shibuya.tao.or.jp

Terumasa Ehara
NHK / TAO

Katsuhiko Shirai
Waseda University / TAO

Abstract　It is known that there are fewer sentences in a TV news text and the sentences are longer compared with those in a newspaper article. We use such TV news texts and would like to summarize them by selecting important sentences. However, since each sentence is rather long, we end up losing a good amount of information if we omit a whole sentence. Therefore, we adopted a method in which we partition long sentences into shorter sentences, and select important sentences from them. We evaluate the results of the method as well as those of the simple sentence extraction where there is no partitioning of the sentences.

key words　Sentence extraction, Sentence partitioning, Text summarization, TV news text

1. はじめに

近年テキストを自動的に要約する技術に関する研究が国内外で盛んになって来ている（[1]、[2]）。本研究は、通信･放送機構 渋谷上原リサーチセンターで進められている「視聴覚障害者向けの放送ソフト制作技術研究開発プロジェクト」（略して「放送ソフトプロジェクト」）での自動要約技術に関する研究の一環であり、自動的にテレビニュース原稿を要約する手法について、重要文抽出、文字数圧縮などをテーマに研究を進めて来ている（[3]、[4]、[5]）。

本稿では、TV ニュース番組の電子化原稿を題材としている。ニュース番組原稿は、新聞記事と似ているが、両者を比較した場合、ニュース原稿のほうが１記事中の文数が少なく、且つ一文当たりの文字数が多いことが分かっている（[6]）。ここで重要文を自動的に抽出することにより要約を作成すると、文数が少なく、１文が長いため、どうしても粗い要約となってしまう。この欠点を補正するために、本稿では、長文を短文に分割する作業を行い（今後「短文分割」と呼ぶ）、その上で重要文抽出を行うことを試みた。そして、その結果について、人によって判断された各文の重要度と比較することにより評価を行った。この評価では単に重要度の高い文を比較するだけではなく、重要度の低い不要文の特定も、システム、人間双方ともで行い、その一致度を測定し、比較検討した。

2. 原稿

題材とした原稿は、NHK 放送データーベースの１９９２年の記事より選ばれた２００件のテレビニュース番組の電子化原稿である。

記事の大きさは、約５００文字であり、１つの記事当たりの文数は、約５文である。

２００記事平均の１記事の文字数、文数、１文当たりの文字数は以下の通りである。

１記事当たり	
平均 文字数	４９５．０８文字
平均 文数	５．１８ 文
１文当たり	
平均 文字数	９５．５７文字

表 1　対象記事の詳細

3. 短文分割

ある条件を満たす文を記事中から選び、その文を分割する作業を行った。まず、分割を行う条件を記述し、その後に、分割した結果について述べる。

3.1. 短文分割の条件

以下の条件で、文の分割を行った。

- 分割をしない場合
  - 基本的に、動詞と形容動詞と述語名詞と形式名詞の連用文節と終止文節以外は分割しない。
  - 連用文節であっても、連体文節直後の連用文節は分割しない。

- 分割をする場合
  - 自立語＋（助動詞）＋「ており」＋読点
    例「働きかけており、ｂｂ」→
    「働きかけております。そして、ｂｂ
  - 自立語＋（助動詞）＋「が」＋読点
    例「働きかけますが、ｂｂ」→
    「働きかけます。しかし、ｂｂ
  - 自立語＋（助動詞）＋「もので」＋読点
    例「働きかけるもので、ｂｂ」→
    「働きかけるものです。そして、ｂｂ
  - 自立語＋（助動詞）＋「ものの」＋読点
    例「働きかけるものの、ｂｂ」→
    「働きかけます。しかし、ｂｂ
  - 自立語＋（助動詞）＋「にもかかわらず」＋読点
    例「働きかけるにもかかわらず、ｂｂ」→
    「働きかけます。それにもかかわらず、ｂｂ

このような分割のための規則が１４使われた。

3.2. 短文分割の結果

上記の条件の下で、分割は、対象となる文の長さ、つまり、文節の数により以下の２つの場合に行うこととした。

- 分割１：分割前に文が１２文節以上あり、分割後５文節以上の文に分割される場合。この場合、２００記事の詳細は以下のようになる。

１記事当たり	
平均 文字数	５０４．９２文字
平均 文数	６．４６ 文
１文当たり	
平均 文字数	７８．１６文字

表 2 分割１での詳細

- 分割２： 分割前１文節以上、分割後１文節以上の文になるようにした場合。つまり、これは、分割出来る文は、全て分割しようという場合に相当する。これにより、４７記事が更に分割された。分割された４７記事を見ると以下のようであった。

１記事当たり	
平均 文字数	５１０．２０文字
平均 文数	７．７２ 文
１文当たり	
平均 文字数	６６．０９文字

表 3 分割２での詳細

分割を行うことにより、一記事当たりの文数が増え、１文当たりの文字数も減った。分割前と２つの分割のケースを比較すると図１のようになる。

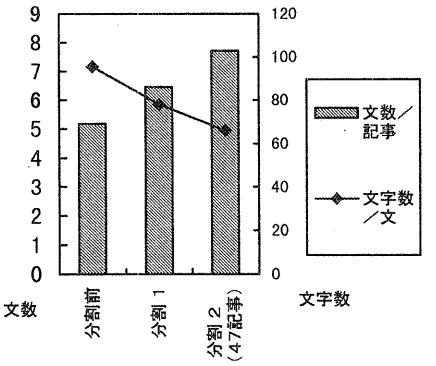

図 1 分割前後での比較

4. 評価

次に、分割を行う前と後の記事を重要文の抽出、不要文の抽出の観点から比較した。

4.1. 分割前の評価

まず、調査したのは、少なくとも４文以上を含む記事５０である。それらの記事に対して、まず、人手で重要度に応じて各文に順位を付けた。それと平行して、２つの自動要約システムを用いて同じ作業を行った。用いたシステムは、１つは市販されているワープロソフトで、その要約機能を利用した（以下システム１と呼ぶ）。もう１つは、現在我々が開発中の自動要約システムである（システム２と呼ぶ）。

調査は、人間による判断と、システムによる判断で順位付けがどれだけ一致しているかを、上位２文、つまり、重要度において第１位と２位の文、と下位２文において行った。一致度は、順位に関係なく、２文とも一致している場合は、1.0 とし、一文だけ一致している場合は、0.5 として計算した。

調査の組み合わせは３通り。

1. 人間とシステム１、
2. 人間とシステム２
3. システム１とシステム２

調査の結果は以下のとおりである。

	一致度	
	上位２文	下位２文
人－システム１	０．７２	０．６８
人－システム２	０．７２	０．５９
システム１－システム２	０．７５	０．６３

表 4　人と２つのシステム間の一致度

上位２文の一致度が下位２文の一致度よりも高くなっているが、全体としてみると、どの組み合わせにおいても、一致度においてそれ程大きな差が認められなかった。

システムと人との判断が異なった例として、資料を論末に付けるサンプル記事をあげる。この記事の場合、記事を読む視点が「女子行員」を中心に読む場合と、「銀行」側からの場合と、複数あり、どちらを重要と思うかにより、各文の重要度の順位が変わって来る。この様な場合、システムと人の判断が違い、結果だけを見ると失敗していることになるが、システムの判断が間違いであると言い切れないと思われる。

また、システムでは、複数の文についてのまとまりを全く考慮されていないが、人が順位付けを行うと、まとまりが発見される場合は、複数の文がまとまって順位付けされ、システムの判断と異なって来る原因となる。

4.2. 分割後の評価

次に、短文分割をした結果について、同じように評価を行った。

評価の対象とした記事数は２０記事であり、ここでも、３文以下となる記事は含まれてない。組み合わせとしては、人手の結果とシステム２の結果のみを評価した。

まず、分割１の場合、つまり、分割の対象となる文が１２文節以上の長さで、かつ分割後の文が５文節以上となるという条件での分割をおこなった結果について、上位２文、下位２文の一致度を調査した。

結果は表５の通りである。

	一致度	
	上位２文	下位２文
人－システム２	０．７０	０．６０

表 5　分割１での一致度

また、分割２の場合、つまり、分割の対象となる文を１文節以上で、かつ分割後の文が１文節以上となる場合の一致度は表６の通りである。

	一致度	
	上位２文	下位２文
人－システム２	０．６７５	０．６２５

表 6　分割２での一致度

これらの結果を見ると、短文分割を行った結果は、重要度に応じた各文の順位付けにおいては、分割をしないものと比較しても、上位２文、下位２文において、余り差のない一致度を示していると言える（図２参照）。

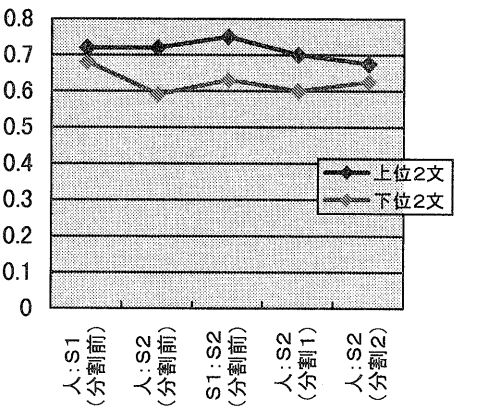

図 2　分割前後の一致度の変化

4.3. 不要文抽出に観点から

短文分割を行っても、行わなくても重要文順位付けの一致度において、あまり差の無いことが判明したが、重要度の低い不要文の削除を実際に行ってみると、分割を行わない場合と、行う場合とで、差があると思われる。

例えば、論末のサンプル記事であるが、分割前の記事に対して、システム２の不要と判断する文（下位２文）は、第４、第７文である。これに対して、分割２の場合では、不要文は、第８、第１０文と判断される。

分割前の記事から、第４文または第７文を削除すると、情報がまとまって削除されることになるが、分割２の場合だと、第８，１０文両方を削除しても、文の流れがとぎれることなく、しかも、削除される情報も少ないと言える。

つまり、短文分割を行うことによって、文の持つ情報がより細かな単位となり、不要文の削除を考えた場合、より木目の細かい削除が可能となると言える。

5. まとめ

ＴＶニュース番組の電子化原稿を題材にして、長文を短く分割して、その後に重要文の抽出を行うことについての検討を行った。

これから、判明したことは、まず、短文分割を行っても、システムによる重要度の判定には、あまり、影響がないということである。これは、重要度において上位２文、下位２文の両方に言えることである。

次に、不要文の削除に注目してみると、短文分割を行う場合と行わない場合によって差があるように思われる。分割を行う各文の持つ情報量が減ることになり、削除をおこなっても失われる情報が少なくなりより良い要約が出来るようになると考えられる。

今後は、文より小さな単位である、文節を単位として取り上げて、自動要約を、削除可能な文節を特定することにより行っていく手法を考えていく予定である。

参考文献

［1］ 言語処理学会第４回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集 １９９８年３月

［2］ The proceedings of ACL 1997 Workshop, *'Intelligent Scalable Text Summarization'*, 1997 Madrid, Spain.

［3］ Wakao, T., Ehara, E., Sawamura, E., Abe, Y., Shirai, K. *Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired* in the proceedings of ACL 97 workshop, Natural Language Processing for Communication Aids, pp 55-58.

［4］ 若尾、江原、白井「テレビニュース番組の字幕に見られる要約の手法」情報処理学会、自然言語処理研究会、NL-122-13.

［5］ 若尾、江原、白井「テレビニュース字幕のための自動要約」言語処理学会併設ワークショップ論文集

［6］ 江原 暉将、沢村 英治、若尾 孝博、 阿部 芳春、白井 克彦 「聴覚障害者のための字幕つきテレビ放送制作への自然言語処理の応用」言語処理学会 第３回年次大会 １９９７年

資料
サンプル記事
オリジナル記事（７文を含む）

[1] 千葉市に本店がある京葉（ケイヨウ）銀行の成田西（ナリタニシ）支店の女子行員が、他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まり、京葉銀行ではきょう、この女子行員を懲戒解雇するとともに、千葉県警察本部に被害を届け出ました。
[2] 京葉銀行によりますと、この女子行員は京葉銀行成田西支店の出納係をしていた二十五歳の行員で、平成二年の九月ごろから今年四月にかけて、京葉銀行が個人向けに発行しているカードローンの七十人あまりの口座から、二十回にわたっておよそ三億円を引き出し、着服していた疑いが持たれています。
[3] このカードローンは京葉銀行が平成元年二月から個人を対象に設けた主力商品で、解雇された女子行員は支店の帳簿を操作して、客が開設したカードローンの口座から勝手に現金を引き出したうえ、ローンを利用した客に郵送される通知についてもコンピューターを操作して客に通知が届かないようにしていたとみられています。
[4] 銀行では、ことし四月になって帳簿に不審な点があることがわかったため、監査を行った結果、不正が明るみに出たもので、銀行の調べに対してこの女子行員は事実関係をほぼ認めているものの、金の使い道については言葉を濁しているということです。
[5] この女子行員は昭和六十年四月に京葉銀行に入社し、平成元年の四月から成田西支店の貸し付け係、平成三年十月から成田西支店の出納係を担当しており、ふだんの勤務態度は真面目で上司の信頼もあつかったということです。
[6] 京葉銀行では、きょう付けでこの女子行員を懲戒解雇するとともに、近く警察に告訴する方針で、成田西支店の支店長についても監督不行き届きで降格処分にしました。
[7] また、記者会見した京葉銀行の綿貫弘一（ワタヌキコウイチ）専務は「各銀行ともカードローンは個人向け融資の中で最もポピュラーな商品で京葉銀行でも契約数は四万四千件余りにのぼっている。二度とこうした事件が起きないよう対策を講じていきたい」と話しています。

短文分割結果（分割２の場合、１３文）

[1] 千葉市に本店がある京葉（ケイヨウ）銀行の成田西（ナリタニシ）支店の女子行員が、他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まりました。
[2] そして、京葉銀行ではきょう、この女子行員を懲戒解雇しました。
[3] それとともに、千葉県警察本部に被害を届け出ました。
[4] 京葉銀行によりますと、この女子行員は京葉銀行成田西支店の出納係をしていた二十五歳の行員で、平成二年の九月ごろから今年四月にかけて、京葉銀行が個人向けに発行しているカードローンの七十人あまりの口座から、二十回にわたっておよそ三億円を引き出し、着服していた疑いが持たれています。
[5] このカードローンは京葉銀行が平成元年二月から個人を対象に設けた主力商品で、解雇された女子行員は支店の帳簿を操作して、客が開設したカードローンの口座から勝手に現金を引き出したうえ、ローンを利用した客に郵送される通知についてもコンピューターを操作して客に通知が届かないようにしていたとみられています。
[6] 銀行では、ことし四月になって帳簿に不審な点があることがわかったため、監査を行った結果、不正が明るみに出たものです。
[7] そして、銀行の調べに対してこの女子行員は事実関係をほぼ認めています。
[8] しかし、金の使い道については言葉を濁しているということです。
[9] この女子行員は昭和六十年四月に京葉銀行に入社し、平成元年の四月から成田西支店の貸し付け係、平成三年十月から成田西支店の出納係を担当していました。
[10] そして、ふだんの勤務態度は真面目で上司の信頼もあつかったということです。
[11] 京葉銀行では、きょう付けでこの女子行員を懲戒解雇しました。
[12] それとともに、近く警察に告訴する方針で、成田西支店の支店長についても監督不行き届きで降格処分にしました。
[13] また、記者会見した京葉銀行の綿貫弘一（ワタヌキコウイチ）専務は「各銀行ともカードローンは個人向け融資の中で最もポピュラーな商品で京葉銀行でも契約数は四万四千件余りにのぼっている。二度とこうした事件が起きないよう対策を講じていきたい」と話しています。