E-026

見出し生成における諺の意味とリズムの利用

The use of a meaning and a rhythm of the proverb in the news headline generation

海老澤　弘明‡　　天沼　博‡　　松澤　和光‡
Hiroaki Ebisawa　Hiroshi Amanuma　Kazumitsu Matsuzawa

1．はじめに

近年、インターネットの普及により Web 上で多くのニュースサイトが見られるようになった。日常的に数多くの情報を扱うニュースサイトでは、見出しの優劣がユーザーの行動に大きな影響を与えている。一般的に人が流し読みで認識できるのは 15 文字程度と言われている。見出しはこの 15 文字の中に記事の内容が想像でき、なおかつ興味を引く言葉を詰め込めるかが重要である。そこで、見出しに諺を用いることで記事の内容を容易に短文化することができる。また、諺は独特のリズムを持った口調の良い言葉である。言葉の置き換えの際に、音の相違度を考慮することで読者にインパクトを与えることができると考えた。本研究では、諺のリズムを利用したニュース見出しの生成法を提案する。

2.　ニュース見出しの生成システム

システムは、記事文の形態素解析、諺の候補抽出、音の相違度の比較、単語の置き換え、見出し出力のステップからなる。

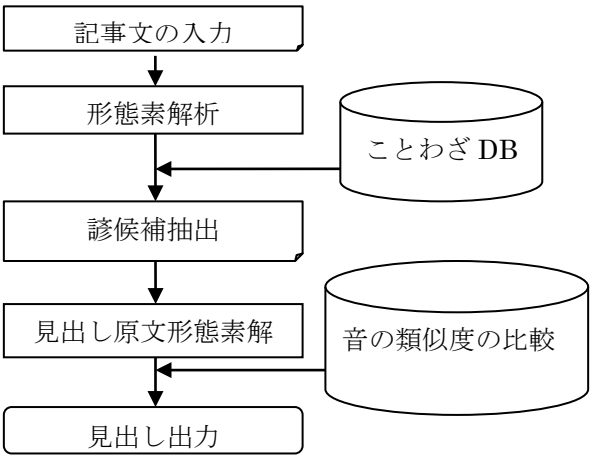

図 1．システムの概要

2.1 ことわざ DB

記事内容に合った諺を選び出すために、365 語の諺をインターネットから収集しことわざ DB を構築した。また、それぞれの諺の解説文を形態素解析して自立語のみ抽出し、諺の意味をあらわすキーワードとしてデータベースに載せた。

キーワードとしてふさわしくないものは選別し、削除、変更した。この際、記事内の単語とデータベースのキーワードが一致しやすいよう、ニュース記事によく使われる特徴的な単語を人手でキーワードとして追加した。

諺毎にキーワードの数に差が出ないように一つの諺につきキーワードは 5，6 個とした。

表 1．　ことわざ DB の例

ことわざ	キーワード
石橋を叩いて渡る	堅固　橋　用心 安全 渡る 確認

2. 2　諺の候補抽出

記事文を ChaSen[1]で形態素解析し単語毎に区切り自立語のみを抽出する。それらをことわざ DB のキーワードと照合し一致した諺を取り出す。

ことわざ	キーワード
攻撃は最大の防御なり	**攻撃**　防御　先制・・・

見出し：崩壊寸前リビアに断末魔…カダフィ無差別攻撃の暴挙

記事：＜ アラブ世界の独裁政権の象徴、リビア・カダフィ政権が崩壊寸前だ。…戦闘機やヘリコプターが反体制デモ隊や市民らを無差別**攻撃**した。＞

文字数の少ない記事、多い記事、それぞれ 10 文に対して諺を抽出した結果、以下のようになった。

表 2　諺の候補抽出結果

ニュース記事	抽出した諺／記事
短文（300～500 字程度）	23.9 個
長文（800～1000 字程度）	45.1 個

‡ 神奈川大学大学院 工学研究科 電気電子情報工学専攻

2. 3 単語置換におけるリズムの利用

新しい見出しは、諺内の単語と記事内の単語の置換を行うことで生成する。この際、置き換える単語は記事内の重要な単語でなくてはならない。そこで、記事の内容を表す単語が多く含まれている元の見出しを利用する。

見出し：崩壊 寸前 リビアに断末魔 カダフィ 無差別 攻撃の暴挙

記事：＜アラブ世界の独裁政権の象徴、リビア・カダフィ政権…無差別攻撃した。＞

記事に諺の候補は、一つの記事に対して平均 30 個程度抽出される。これら全てに見出しから抽出した単語を当てはめると、非常に多くの見出しが生成されてしまう。そこで、単語を置き換える指標として諺の持つリズムを利用する。

諺内の単語と見出し原文から抽出した単語が音韻的にどのくらい類似しているかを比較し、類似度が低いものを候補から除外する。

見出しの原文単語	諺内の単語	
暴挙（ボウキョ）	防御（ボウギョ）	候補
崩壊（ホウカイ）	防御	候補
リビア（リビア）	防御	除外

音韻的類似度が高いものから順に候補とし見出しを生成する。

生成された見出し：**攻撃は最大の暴挙**

抽出された諺によっては、音韻的類似度が高い単語の組が出来ない可能性がある。そこで、類似度が上手く取れない場合は単語の文字数を比較し、文字数の近い単語を候補として見出しを生成する。

表 3　システムの生成した見出しの例

生成された見出し	元の見出し
農薬口に苦し （良薬口に苦し）	農薬入りコーヒー殺人未遂で懲役８年を求刑
被災丸儲け （坊主丸儲け）	東日本大震災　被災地でヤミ金が暗躍　生活難につけこむ
もろ刃の正義 （もろ刃の剣）	作戦約４０分間の末ビンラディン容疑者を殺害「正義を達成」

3. 評価実験

システムが生成した見出しについて評価実験を行った。生成された見出しが記事の内容にあっているか、また、元の見出しと比べ記事内容を読みたくなるかの観点から評価した。実験はシステムの生成した見出し 8 個を 5 人に評価してもらった。

3. 1　実験結果

実験項目 1　元の見出しと比べて読みたくなるか。

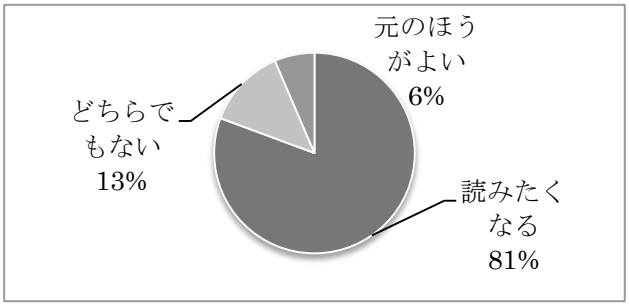

図 2. 実験項目 1 の結果

実験項目 2　生成された見出しが記事内容と合っているか。

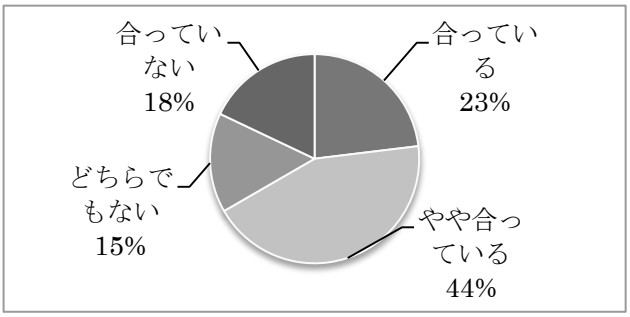

図 3. 実験項目 2 の結果

4. おわりに

本稿では、諺の意味とリズムを利用したニュース見出しの生成法を提案した。提案システムでは、誰でも知っているような諺を使うことで読者に興味を持たせるような見出しを生成する。システムでは諺の持つ独特のリズムを利用するため、諺内の単語と記事内の単語を置き換える際に、単語の文字数、音の類似度を考慮した。実験結果より、システムの生成した見出しは元の見出しよりも読みたくなるという評価が約 8 割あった。しかし、記事内容と見出しの整合性に対する評価が 2 割と低くなってしまった。今後は、記事内容に対する諺抽出方法に改善が必要である。

参考文献

[1] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座松本研究室:
http://chasen.naist.jp/hiki/ChaSen/
[2]海老澤弘明、天沼博、松澤和光　”諺を用いたニュース見出し生成法”人工知能学会第 36 回ことば工学研究会、発表番号 2