データ圧縮率を用いるテキストアート抽出法における テキストの正規化手法

鈴木 徹也
芝浦工業大学 システム理工学部 電子情報システム学科

tetsuya@sic.shibaura-it.ac.jp

1　はじめに

テキストアートは, テキストの表現を豊かにする一方で, テキストの形態素解析や読み上げでは障害となる. **テキストアート**とは, コンピュータで表示される文字を組合せて描かれた絵である. テキストアートは一般にはアスキーアートとも呼ばれる. 近年, Web ページや電子メールで交換されるテキストデータにおいて, このようなテキストアートがよく用いられている.

その問題は, テキスト中のテキストアートの範囲が分かれば解決できる. テキストアートを削除したり, 代替文字列に置き換えたりできるからである.

本論文では, 入力テキスト中のテキストアートの範囲を求める方法を**テキストアート抽出法**と呼ぶ. 同様に, 入力テキストがテキストアートであるか否かを判定する方法を**テキストアート識別法**と呼ぶ.

汎用性を求めるならば, テキストアート抽出法は特定の自然言語に依存すべきではない. なぜなら, 1 つのテキストデータに, 日本語や英語など, 複数の自然言語が含まれる可能性があるからである.

そこで我々は, 特定の自然言語の特徴を利用しないテキストアート抽出法を提案してきた [4, 7, 3]. 我々はテキストデータのテキストアートらしさとして, 同じ文字の連続出現に注目する. そこで我々のその抽出法は, その特徴を計るために, テキストのランレングス符号化 [1] によるデータ圧縮率を利用する.

ところで, これまで我々は, テキストアート中の空白の扱いを検討していなかった. 例えば以下の点である.

> 空白に見える文字には, いわゆる全角空白と半角空白という, 表示上の横幅の違う空白文字がある. 空白に見える文字は統一すべきなのか.
>
> テキストアートを構成する各行の長さが違うことがある. その場合, 短い行の行末より右側に, 空白文字が表示されているように見える. その部分にも空白文字があるとして処理すべきなのか.
>
> テキストアートの左側の余白はテキストアートの一部として扱うべきなのか.

本研究の目的は, 上記 3 点について, 我々のテキストアート抽出法における正規化手法を実験により比較検討することである.

本論文の構成を説明する. まず 2 節で我々のテキストアート抽出法を紹介する. 次に 3 節でテキストの正規化法を三つ提案し, 4 節で実験によりそれらの正規化手法を比較する. 最後の 5 節で結論を述べる.

なお本研究では UTF-8 で符号化されたテキストを扱う.

2　テキストアート抽出法

我々のテキストアート抽出法を説明する. 2.1 節と 2.2 節とで部分的な手続きを説明した後に, 2.3 節でその抽出法を説明する. 2.4 節では, その抽出法で利用するテキストアート識別器の構築法を説明する.

2.1　窓幅 k での走査

窓幅 k での走査とは, 対象テキストの先頭行から最終行まで, k 行毎にある処理を施す手続きである. 注目する k 行を**窓**と呼ぶことにし, k を窓の**幅**と呼ぶことにする. 窓は 1 行ずつずらす.

2.2　テキスト範囲縮小処理

テキスト範囲縮小処理とは, 開始行と終了行との組で表されたテキストの範囲に対して, 次の 2 つの行を取り除いた範囲を求める手続きである.

> 開始行から終了行へ向かって 1 行ずつ注目し, 連続してテキストアートと判定されない行

終了行から開始行へ向かって 1 行ずつ注目し, 連続してテキストアートと判定されない行

注目する 1 行がテキストアートであるかの判定には, テキストアート識別器を用いる.

2.3 テキストアート抽出処理

窓幅 w のテキストアート抽出手続きを示す. この手続きはテキストを入力とし, そのテキストの範囲の集合を出力とする. この手続きが使用するテキストアート識別器 M は予め構築されているものとする.

1. テキストに対して, 窓幅 w の走査を行う. その際, 窓内のテキストを M で識別する.
2. 連続してテキストアートと判定されたテキストの範囲を, 1 つのテキストアート候補範囲とする.
3. 各テキストアート候補範囲に, M を用いてテキスト範囲縮小処理を行う.
4. テキスト範囲縮小処理の各結果を 1 つのテキストアートとして, テキストアートの集合を出力する.

テキストアートの抽出例を示す. 図 1 はテキストアートと非テキストアートとが混在したテキストデータの例である. 図 2 は, 図 1 から求めたテキストアート候補範囲である. この段階ではテキストアートの前後に非テキストアート行が存在する. 図 3 は, 図 2 にテキスト範囲縮小処理を施した結果である. 図 3 では, 図 2 の非テキストアート行が取り除かれている.

2.4 テキストアート識別器の構築

窓幅 w のテキストアート抽出器に用いるテキストアート識別器を, 機械学習アルゴリズム C4.5[6] によって構築する. その結果得られる識別器は, テキストの属性を入力とする決定木である.

学習に利用する訓練データは次のように準備する.

1. テキストアートの集合 (正例) と非テキストアートの集合 (負例) とを用意する.
2. 正例と負例の各データを窓幅 i $(= 1, 2, \ldots, w)$ で走査しながら, 窓毎にテキストの属性を抽出する. 抽出する属性は, ランレングス符号化による圧縮率, 行数, バイト数の三つである.

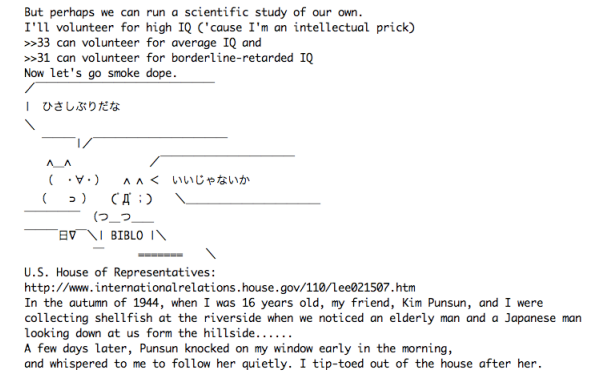

図 1: テキストアート抽出法への入力例

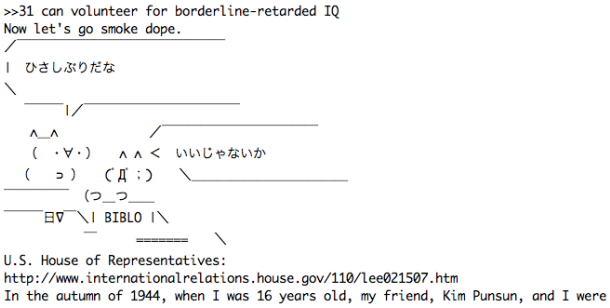

図 2: テキストアート候補範囲の例

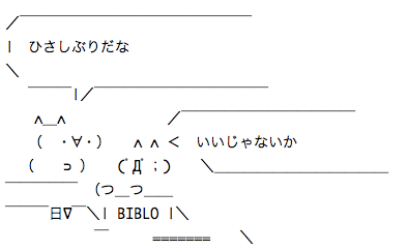

図 3: テキスト範囲縮小処理の結果例

3 テキストの正規化手法

我々のテキストアート抽出法に導入する, テキストの正規化手法を提案する. 正規化はテキストの属性抽出の直前に行う.

その前にまず用語を定義する. Unicode の文字 U+0020 と文字 U+3000 とを, それぞれ**半角空白**と**全角空白**と呼ぶことにする. そしてこれらを合わせて**空白文字**と呼ぶことにする. ある**行の幅**とは, その行末コードを取り除いた部分の文字幅の合計とする. 各**文字の幅**については, 半角文字は幅 1, 全角文字は幅 2 とする. なお [5] による Unicode 文字の分類のうち, 分類 F, W, A を全角文字, それ以外を半角文字とする.

以下が提案する正規化手法である.

正規化手法 1 窓内の全角空白それぞれを半角空白二つに置換する.

正規化手法 2 窓内の行の最大幅を w とするとき, 窓内の全ての行の幅が w となるように各行の末尾に半角空白文字を追加する.

正規化手法 3 正規化手法 1 を施した後, 窓内の一番左端に出現する非空白文字が窓の左端になるように, 各行から等しい文字数だけ空白文字を削除する.

4 正規化手法の比較実験

提案した正規化手法を比較するために, それぞれの手法を用いてテキストアートの抽出を行った.

4.1 実験条件

まずテキストの集合として以下の E と J を用意した.

E 英語テキストの集合 (テキストアート 289, 非テキストアート 290)

J 日本語テキストの集合 (テキストアート 259, 非テキストアート 299)

それぞれのテキストの行末コードは CR LF に統一した.

そしてこれら E と J から, 2 つのテキストの集合 A と B を作成した. A と B の構成は次の通りである.

A 英語のテキストアート 145, 日本語のテキストアート 130, 英語の非テキストアート 145, 日本語の非テキストアート 150

B 英語のテキストアート 144, 日本語のテキストアート 129, 英語の非テキストアート 145, 日本語の非テキストアート 149

A は, テキストアート識別器を構築するための訓練データとした. B からは 800 個のテストデータを作成した. 各テストデータの構成は, 図 1 のように, 1 つのテキストアートを非テキストアートで挟む形式にした. その各部分は B からランダムに選択した.

なお機械学習アルゴリズム C4.5 は, データマイニングツール Weka[2, 6] の実装を用いた.

テキストアートの抽出は, 窓幅を 1 から 5 まで変えながら, 正規化手法なし, 正規化手法 1, 正規化手法 2, 正規化手法 3 の各場合について行った.

集合 A と B とを入れ替えて同様の抽出実験を行い, 適合率の平均, 再現率の平均, F 値 (適合率と再現率の調和平均) の平均とを求めた.

表 1: 正規化手法を利用しない抽出結果

窓幅	適合率	再現率	F 値
1	0.939	0.879	0.908
2	0.924	0.917	0.920
3	0.891	0.893	0.892
4	0.864	0.867	0.865
5	0.836	0.849	0.842

表 2: 正規化手法 1 の抽出結果

窓幅	適合率	再現率	F 値
1	0.955	0.930	0.942
2	0.949	0.951	0.950
3	0.927	0.929	0.928
4	0.927	0.920	0.923
5	0.895	0.901	0.898

表 3: 正規化手法 2 の抽出結果

窓幅	適合率	再現率	F 値
1	0.939	0.879	0.908
2	0.744	0.742	0.743
3	0.543	0.532	0.537
4	0.523	0.513	0.518
5	0.726	0.749	0.737

表 4: 正規化手法 3 の抽出結果

窓幅	適合率	再現率	F 値
1	0.896	0.925	0.910
2	0.911	0.943	0.927
3	0.882	0.915	0.898
4	0.855	0.900	0.877
5	0.833	0.894	0.862

4.2 結果

表 1 から表 4 に実験結果を示す. 各場合の結果を簡単にまとめると次のようになる.

正規化手法なし 窓幅 2 のとき F 値は最大値 0.920 をとる.

正規化手法 1 窓幅 2 のとき F 値は最大値 0.950 をとる．その値は正規化手法無しの場合と比べて 3.26%向上した．

正規化手法 2 窓幅 1 のときに F 値は最大値 0.908 をとる．しかし, 実際にこの正規化手法の効果があるのは窓幅 2 以上の時である．したがって, 今回の実験では, この手法には F 値の向上は認められない．

正規化手法 3 正規化手法 3 は正規化手法 1 を含むので, 正規化手法 3 は正規化手法 1 と比較する．正規化手法 3 の場合, 窓幅 2 のとき F 値は最大値 0.927 をとる．その値は正規化手法 1 のそれの 97.6%である．

4.3 考察

正規化手法 2 と 3 とで, F 値の最大値が向上しなかった理由を考える．我々のテキストアート抽出法は, テキストアートらしさとして, 同じ文字の連続出現に注目している．そういったテキストアートらしさを, 正規化手法 2 は非テキストアートに与え、正規化手法 3 はテキストアートから奪う．そのため, テキストアートと非テキストアートとの識別が難しくなったと考えられる．

5 おわりに

本研究では, 我々が提案したテキストアート抽出法における, テキストデータの三つの正規化手法を実験により比較した．我々のその抽出法は, テキストのテキストアートらしさとして, 同じ文字の連続出現に注目する．比較した正規化手法はどれも空白文字に関する手法である．実験により, テキストアートの左右の余白は調整せず, 全角空白それぞれを二つの半角空白に置換する正規化手法が有効であると確認した．その正規化手法では, 正規化をしない場合と比べて, 適合率と再現率との調和平均が 3%以上向上した．

参考文献

[1] M. ネルソン (著), 荻原 剛志 (訳), 山口 英 (訳). データ圧縮ハンドブック-C プログラマのための圧縮技法紹介. トッパン, 1994.

[2] The University of Waikato. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. http://www.cs.waikato.ac.nz/ml/weka/ (Retrieved on Jan. 24, 2011).

[3] Tetsuya SUZUKI. A Decision Tree-based Text Art Extraction Method without any Language-Dependent Text Attribute. *International Journal of Computational Linguistics Research*, Vol. 1, No. 1, pp. 12–22, 2010.

[4] Tetsuya SUZUKI and Kazuyuki HAYASHI. Text data compression ratio as a text attribute for a language-independent text art extraction method. In *Proceedings of the Third International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2010)*, 2010.

[5] Unicode, Inc. East Asian Width property data file. http://www.unicode.org/Public/UNIDATA/EastAsianWidth.txt (Retrieved on Jan. 16, 2011).

[6] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.

[7] 林和幸, 鈴木徹也. テキスト圧縮を用いた言語に依存しないテキストアート抽出法. 情報処理学会研究報告. DD, [デジタル・ドキュメント], Vol. 2009, No. 3, pp. 1–6, 2009-09-18.