Augmented Copy: 自然言語処理を重畳するコピー機

荒牧英治[†1]　久保圭[†1]　仲村哲明[†1]　島本裕美子[†1]　宮部真衣[†]

紙媒体には紙独特の利点があり，我々の生活から完全に消えることはないであろう．この紙の利点をそのままに情報を付加すべく，我々は，紙を入出力として扱う処理系，すなわち，紙を入力とし，紙に注釈を加える形で言語処理結果を出力することを提案する．この入力紙に情報を付加または削除する技術は，AR 技術（Augmented Reality；知覚情報に計算機が作り出した情報を重ね合わせる）を，紙媒体に適用したとも考えられ，本稿では，Augmented Copy と呼ぶ．本稿では，Augmented Copy のいくつかの実装例を紹介し，その可能性を議論する．

Augmented Copy: a Copy Machine Augmented by Natural Language Processing

Eiji ARAMAKI [†1]　Kay KUBO[†1]　Tetsuaki NAKAMURA[†1]
Yumiko SHIMAMOTO[†1]　Mai MIYABE[†1]

A paper is traditional media, still appealing its advantage. In order to combine the benefit of a paper and digital media, this study proposes a paper-input and paper-output natural language system. We call our system Augmented Copy (AC), because the proposed system is similar to augmented reality (AR). While AR presents a view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated data, AC presents a paper whose elements are augmented. This study introduces a couple examples of AC, and discussed the feasibility.

1. はじめに

ビッグデータの隆盛を背景に，言語処理技術は急速に進歩しつつあり，匿名化（日本語カルテで 90%[1]，英語電子カルテで 96%[2]）は人間の精度に匹敵し，要約，言い換えなどいくつかの処理においても用途を限定すれば実用化可能，またはその直前の精度まで到達している．しかし，機械翻訳，音声検索など一部の応用志向の技術を除いて（コンピューターの専門家でない）一般ユーザが言語処理技術を直接使用し，その恩恵を感じることはまれである．この原因の一つとして，多くの言語処理システムの扱う対象が計算機上に格納されたテキストファイルであり，実社会で使われる媒体とギャップがあるのが問題であると考える．

一般のユーザにとって，扱うテキストの多くは Microsoft WORD で読み書きする docx 形式のファイルであったり，Microsoft EXCEL のための xlsx ファイルであったり，または，それらが印刷された紙媒体である．さらに，契約書やカルテなど，重要な書類ほど紙媒体が強制される可能性が高い．また，重要でなくとも，新聞や雑誌など日常生活で気軽に利用するメディアも依然として一定割合で紙媒体が好まれる傾向がある．これらは将来的に，徐々に電子化されていくかもしれないが，紙媒体が未だ消滅しないのは，紙媒体ならではの利点があるからであろう．例えば，紙の解像度は 300dpi を超え，ディスプレイの解像度を凌駕して

図 1.　Augmented Copy の例「匿名コピー」．
Figure 1. Example of Augmented Copy “De-identification Copy”

†1 京都大学
Kyoto University

いる．ファイルサイズが 5MB をこえる文章も B0 サイズの紙に印刷し，一覧が可能である．注釈などワープロソフトの行や列の概念に縛られずに書き込みが可能である．自在に折ることができるなど，他にも紙の利点は多数あり，当分，紙が我々の生活から完全に消えることはなさそうである．

そこで，我々は，紙を入出力として扱う処理系，すなわち，紙を入力とし，言語処理結果を紙に注釈を加える形で出力することを提案する．図 1 に Augmented Copy の実例を示す．これは，固有名抽出処理という言語処理を OCR，プリンターと組み合わせたものであり，固有名部分を黒塗りにして出力する．このように入力したテキストを可能な限りそのままに，一部に情報を付加または削除する技術は，AR 技術（Augmented Reality；知覚情報に計算機が作り出した情報を重ね合わせる）を，紙媒体に適用したとも考えられ，本稿では，Augmented Copy と呼ぶ．本稿では，Augmented Copy のいくつかの実装例を紹介し，その可能性を議論する．

2. 方法

Augmented Copy は言語処理をとりこんだコピーであるが，自然言語処理は，例えば，機械翻訳，文章要約などさまざまである．これらの処理内容や目的は多様であるが，多くの言語処理は，文章を入力として，その一部の文字列に対して，文字列リストを付与し出力するという形で一般化できる． 例えば，機械翻訳（自動翻訳）は，入力が文であり，それに対して翻訳結果の候補を返す．匿名化は，入力が文であり，固有表現を特定した結果を返す．このように言語処理応用の多くは，ある文字からある文字までの間に対して n 個の文字列を付与するというデータ構造で表現できる．本研究で提案する Augmented Copy は，このデータ構造に対応した XML 形式を用いることで， さまざまな言語処理結果を重畳与してプリントする．

これを実現するためには以下の 3 つのモジュールが必要となる．

- **【OCR モジュール】**入力紙を画像として読み込む（IMG-A）．さらに，画像上の日本語文章をテキスト化する（TXT-A）．これは
- **【NLP モジュール】**読み込んだテキスト(TXT-A)に対して，言語処理を行い，結果を得る（TXT-B）．
- **【画像処理モジュール】**言語処理結果（TXT-B）を画像（IMG-A）に埋め込み，画像（IMG-B）を得る．

以下に，各モジュールの仕様を述べる．

2.1 OCR モジュール

画像中のテキスト情報を得る（OCR 処理を行う）．この際，特定の文字列に，重畳表示を行うためには，文字情報に加え，文字の位置情報も取得する必要がある．本システムではメディアドライブ社の「活字文書 OCR ライブラリ v7.0」

図 2. Augmented Copy で付与可能な表現．

Figure 2. Expressions realized by Augmented Copy

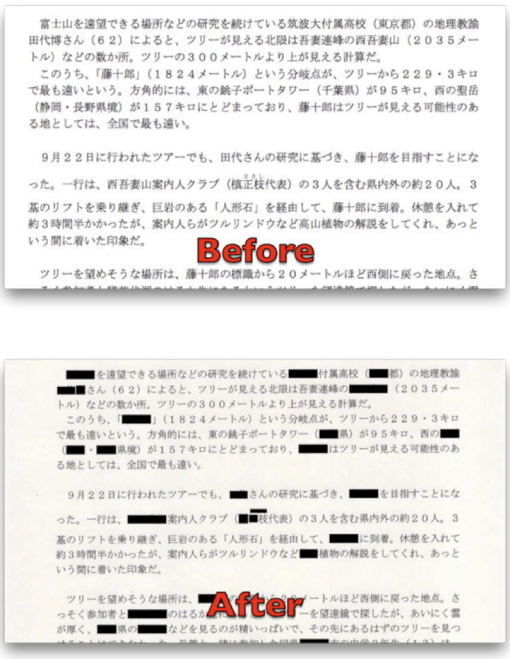

図 3. 「匿名化コピー」の入出力例．

Figure 3. Example of “De-identification Copy”

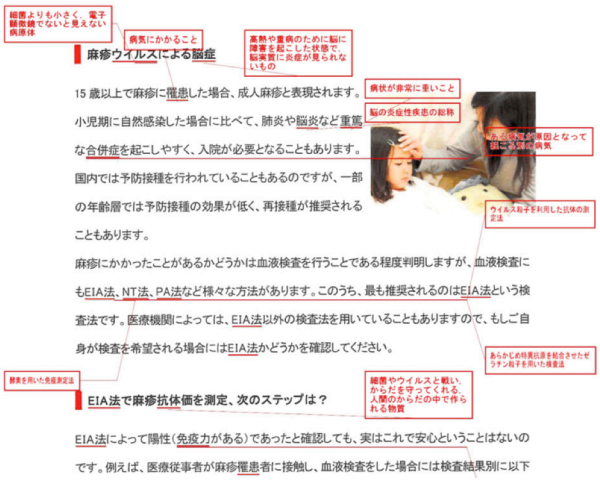

図 4. 「簡単化コピー」の出力例．

Figure 4. Example of “Paraphrase Copy”

[3]を用いた．当ライブラリは文字の開始位置（左上座標）と終了位置（右下座標）を取得可能であり，実装の要件を満たす．

2.2 言語処理モジュール

提案する Augmented Copy は特定の言語処理技術に依存しない．本研究では，匿名化（個人情報削除）[4]，難解語句検出と解説付与[5]，スペルミス検出[5]，病名の自動分類[5,6]という 4 つのサービスの実装を行った．ここでは，辞書に基づく言い換えを例に説明する．テキスト部について，一定文字毎（現状では 800 文字）に言語処理サーバーにテキストを送信する ．言語処理サーバーは受け取った解析を行い（この場合は，辞書にマッチする語があるかどうか調べる），該当箇所があればその開始位置と終了位置を XML 形式で返す．解析結果は，単なる位置情報でなく，どのように可視化を行うかを含む．現在，下線，枠線，ルビなどに加え，校正メモ領域の追加など 8 つの重畳表示に対応している（図 2）．

2.3 画像処理モジュール

言語処理結果に従い画像を編集し，プリントを行う．これは ImageMagick-6.8 を用いて行った．この際，校正メモ領域注釈を加える場合には，どの場所に校正メモ領域を作成するか自由度がある．本実装では，校正箇所からもっとも近傍にある空白をさがし，その場所に注釈を加える．十分な空白がない場合は，テキストにオーバラップして描画するという実装をとっている．

3. 実装例

実装した言語処理技術は，匿名化，難解語句検出と解説付与，スペルミス検出，病名の自動分類であり，いずれも同一のデータ構造で動作可能である．アプリケーションの概観を図 3 に示す．以下にそれぞれの概要を示す．

- **匿名化**
  「匿名化コピー」：日本語固有表現認識（Named Entity Recognition）技術との組み合わせ．人名，施設名など固有名詞を黒塗りに変換してプリントする．
- **難解語句検出と解説付与**
  「簡単化コピー」：日本語校正技術との組み合わせ．医療ドメインでの漢字変換ミスが 300 例収載されており，マッチする箇所を強調し，プリントする．
- **スペルミス検出**
  「校正コピー」：日本語の難解な医療用語約 400 語とその解説のデータベースを用い，難解語に対して，解説を注釈として付与し，プリントする．
- **病名の自動分類**
  「コーディング・コピー」：医療事務の現場では，病名に対して，病名コード（ICD-10）を付与する必要が生じる．この処理時間を軽減させるために，文章中に含まれる病名に病名コードを付与しプリントする．

これらの処理結果の例を図 3（匿名化）と図 4（簡単化）に示す．上記の他にも，年号の削除による穴埋め問題自動作成といった教育目的での利用など，多くの応用先が考えられる．

4. おわりに

本研究では，紙の利点を活かした言語処理技術として，紙を入出力として扱う処理系，すなわち，紙を入力とし，言語処理結果を注釈として加える形で紙を出力（プリント）する Augmented Copy を提案した．

Augmented Copy は，媒体自体は変化していないので，従来デジタルの欠点であった閲覧性の低さ，十年単位の長期の見読性への不安などを損なうことなく，言語処理結果を重畳できる．将来的に，入出力が紙でないといけないという適切なニーズがあれば，キラーアプリケーションとなる可能性を含んでいる．今後，この枠組を利用したアプリケーションを開発する予定である．

謝辞　本研究の一部は，JST さきがけ「自然言語処理による診断支援技術の開発」プロジェクトおよび，若手研究（A）「表記ゆれ及びそれに類する現象の包括的言語処理に関する研究」の助成を受けた．

参考文献

1) Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, Eiji Aramaki: Overview of the NTCIR-10 MedNLP task, In Proceedings of NTCIR-10, 2013.
2) Eiji Aramaki, Takeshi Imai, Kengo Miyo, Kazuhiko Ohe: Automatic Deidentification by using Sentence Features and Label Consistency, Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
3) http://mediadrive.jp/topics/2010/20100205katsujilib07.html
4) 荒牧英治, 増川佐知子, 宮部真衣, 森田瑞樹: テキストの k 匿名化, 第 155 回データベースシステム研究発表会, 2012.
5) 宮部真衣, 森田瑞樹, 荒牧英治: 医療テキストを対象とした言語処理実装システムとそのデータ構造 第 33 回医療情報学連合大会 (第 14 回日本医療情報学会学術大会).
6) Hiroto Imachi, Mizuki Morita, Eiji Aramaki: NTCIR10 MedNLP Baseline System, In Proceedings of NTCIR-10, 2013.