「日本語処理解析性能評価」分科会
活動報告及び活動予定について

平成27年4月15日

主査　絹川　博之（東京電機大学）

「日本語処理解析性能評価」分科会

- 本分科会設立にいたる課題

  デジタル・フォレンジック や **eディスカバリ** を用途とする **検索・解析ツール**に関し、**海外で開発**されたものが多く、日本国内での使用時に、日本語への対応が不明で、下記のような**問題が発生！**

  - **検索モレがあり、重要証拠ファイルの不検出**
  - **文字化けの発生で、内容の確認不可**
  - **どこまで日本語処理対応しているか不明確**
  - **日本語処理の精度の信頼度に課題**
  - **日本語処理対応の定義が不明確**

「日本語処理解析性能評価」分科会

- 本分科会設立の目的
  - **デジタル・フォレンジック** や **eディスカバリ** の対象となる **日本語情報**に対する**処理解析性能を評価**するための有効な**指標**を作成し、**客観的な評価**の実施を可能とする。
  - ツール提供企業の**技術進歩**を促し、デジタル・フォレンジック技術の日本国内でのさらなる発展に寄与する。

「日本語処理解析性能評価」分科会メンバー

（敬称略）

主査：絹川博之（東京電機大）
幹事：野﨑周作（WG1リーダ：㈱UBIC）
幹事：白井喜勝（WG2リーダ：㈱UBIC）

＜検討参加メンバー（団体会員企業は代表者のみ記載）＞
舟橋 信（㈱UBIC）　浦口 康也　（㈱くまなんピーシーネット）
青木 和哉（FTIコンサルティング）　青嶋 信仁　（㈱ディアイティ）
春山 洋（AOSリーガルテック㈱）　緒方 健（おがたコンサルティング）
伊藤 文二　（日本ダイレックス㈱）　岡田 忠（茨城大学大学院）

＜アドバイザーグループ（個人の立場でご参加）＞
森田 陽　石崎 俊　本田 英盟　熊澤 篤　岡野 薫　野本 靖之

2014年度の活動日程

- 「日本語処理解析性能評価」WG:
  2014年2月～2014年7月　6回実施
  「日本語処理解析性能評価」分科会設立準備

- 「日本語処理解析性能評価」分科会:
  2014年8月～2015年3月　5回実施
  - 検討課題の整理とWGの発足
  - 各WGでの検討と分科会での審議

2014年度の活動内容

- 日本語処理解析性能に関する評価基準案の策定
  - **WG1:リーダ　野崎幹事**
    (1) 文字コード
    (2) アプリケーション
  - **WG2:リーダ　白井幹事**
    (3) 基本検索
    (4) アドバンス検索
  - **WG1、WG2共通**
    (5) その他重要事項
- 項目のリストアップ、サンプルデータ作成の難易度を審議検討し、評価基準案の作成
  （詳細は各WGのリーダより報告）

日本語処理解析性能　評価決定項目

(1) 文字コード：８種
UTF-16、UTF-8、UTF-7、Shift-JIS、JIS、ISO-2022-JP、EUC-JP、Windows-31J

(2) アプリケーション：５＋７種
MS-Word/Excel/PowerPoint for WIN.、AdobePDF、InternetExlorer(HTML)
メール：Outlook系4種、LotusNote、Becky!、Thuderbird

(3) 基本検索：３種
AND、OR、NOT

(4) アドバンス検索：５種
正規表現検索、前後に特定文字が来ない検索、グルーピング、近傍検索
全角･半角の片方入力で自動的に両方検索

(5) その他重要事項：３項目
解析速度、評価用データの作成･収集の基本的な考え方、評価対象

「日本語処理解析性能評価」分科会
２０１５年度の活動予定

- 日本語処理解析性能評価委員会の設置準備
  - **評価の実行を目的**
  - **本委員会の役割の明確化**
  - **委員の募集要項の作成**
  - **委員募集の実施、本委員会の立上げ**
- 分科会：5月から、隔月開催
- サンプルデータ素材：IDFコラム原稿の利用
  - **第１１期第6回理事会承認**
  - **著者の使用承認を得たコラム原稿を使用**

WG1検討内容報告

第12期第1回「日本語処理解析性能評価」分科会

2015.04.15

野崎 周作

WG1-1. 文字コード

評価項目とする理由

日本語処理の性能評価を考えるにあたり、複数存在する日本語文字コードを正しく認識し処理できる事が重要となる為評価項目とする。

選定の基準としては日本語の文字コードとして広く使用されているものを対象とする。

選定された文字コード8種

UTF-16、UTF-8、UTF-7、Shift-JIS、Windows-31J、JIS、ISO-2022-JP、EUC-JP

評価データ作成に向けてのポイント

- 復元機能やHDD未使用領域のバイナリ検索機能は評価対象外とする。
  →評価データはファイルとして存在するものを対象とする。
- 純粋な文字コード判別評価の為、評価データとしてヘッダー情報の無いプレーンテキストファイルを作成する。

WG1-1. 文字コード

選定の経緯及び議論等

- GB2312をはじめとする中国語文字コードも調査対象としてよく目にする。これらの文字コードも判別できると良い。

  →「日本語」処理性能評価に注力する為、対象外とする。（今後検討の可能性有り）

- 人名を正確に表記する為には外字データが必要であり、日本語特有な情報の為評価対象に含んだ方が良い。

  →Windows-31Jは著名な外字集合を含んでいる為、まずはこちらで評価する。

- Base64でエンコードされた情報の処理ができると良い。

  →Base64は電子メールの添付ファイルなどのバイナリ形式のデータ送信に標準的に使用されている為、電子メールアプリケーション評価の中で併せて評価する。

WG1-2. アプリケーション

評価項目とする理由

実際の調査において対象となる情報の多くは、電子メールやオフィスファイルといった電子ファイル群となる。日本語処理の性能評価を考えるにあたり、アプリケーションファイルの処理が重要となる為、評価項目とする。

選定の基準としては、対象とするアプリケーションの種類を絞り込み、更にその中から一般的に広く使用されているアプリケーションを厳選して絞り込み対象とする。

リストアップされたアプリケーション種類

Email、Word Processing、Spreadsheet、Presentation、Archive、Database、Image、Multimedia

選定されたアプリケーション種類

Email、Word Processing、Spreadsheet、Presentation

WG1-2. アプリケーション

選定の経緯及び議論等

- 日本語のテキスト処理をまずは評価する為、Multimediaファイルは対象外とする。
- Databaseファイル内の各レコードは一般的にフォレンジックツールで検索する事は無い為対象外とする。
- Archiveファイルはそれ自体がドキュメントファイルではなく、Archiveファイルに格納されているファイルが調査対象となる。まずはドキュメントファイルレベルでの評価を進める為評価データからは対象外とする。

  →フォレンジックツールがArchiveファイルを展開処理する際にファイル名が文字化けするケースも存在する事から将来的には評価データに含める事を検討する。

評価データ作成に向けてのポイント

検索対象とするテキストは、ドキュメントの本文だけではなく、メタデータも対象とする。

WG1-2. アプリケーション

選定されたアプリケーション12種類

Email

Microsoft Outlook (MSG)
Microsoft Outlook Express / Windows Mail (EML)
Microsoft Outlook (PST)
Microsoft Outlook Express (DBX)
IBM Lotus Notes (NSF)
Becky! Internet Mail
Mozilla Thunderbird

Word Processing、Spreadsheet、Presentation

Microsoft Office for Windows(Word、Excel、Powerpoint)
Adobe (PDF) *テキスト情報を保持したもの

HyperText Markup Language（HTML）

WG1-2. アプリケーション

選定の経緯及び議論等

- Outlook系4種は最も一般的に使用されている為、対象とする。
- Lotus Notesは主に企業内でよく使用されている為、対象とする。
- Becky! Internet Mailは実際の捜査・調査において調査対象となる事が多い為、対象とする。
- Mozilla ThunderbirdはWindows、Mac、Linuxの全てのOSで使用出来ることから、使用者が多いと推測される為、対象とする。
- オフィスファイルはMicrosoft Office for Windowsに絞り込み対象とする。
- Adobe PDFは一般的に使用されている為対象とする。但し、テキスト情報を保持したものを対象とし、OCR性能は評価しない。
- URLエンコード（%エンコード）で記録された情報も検索できると良い。
  →HyperText Markup Language（HTML）を対象に加え評価する。

WG2検討内容報告

第12期第1回「日本語処理解析性能評価」分科会

2015.04.15

白井 喜勝

WG2-1. 基本検索

項目リストアップの考え方

日本語を対象にしたデジタル・フォレンジック調査やeディスカバリにおいて、頻繁に使用される基本的な検索機能をリストアップ。

リストアップされた4項目

AND、OR、NOT、XOR

評価対象選定の考え方

有用性、使用頻度、代替方法の有無等の観点。

選定された基本検索3項目

AND、OR、NOT

選定の経緯等

- AND、OR に関しては、全員一致で即決。
- NOTに関しては、使用頻度に関する議論があった。あきらかに関係ない文書に含まれている共通のキーワードを見つけてNOTで絞る使用法がよく利用されること、文書抽出の際に、不必要なキーワードと必要なキーワードで文書を浮かび上がらせるためにも必要ということで、実務上有用なため選定された。
- XORは、AND、OR、NOTの組み合わせで検索できるため含めない。（第11期第3回「日本語処理解析性能評価」分科会にて合意。）

WG2-2. アドバンス検索１

項目リストアップの考え方

日本語を対象にしたデジタル・フォレンジック調査やeディスカバリにおいて有用な、高度な検索及びそれに類する技術をリストアップ。

リストアップされた9項目

- 正規表現検索
- 前後にアルファベットや数字が来ない、あるいは特定の文字が来ない言葉の検索。
- グルーピング：ブーリアンを（）で括って複雑な論理を組む。
- 近傍検索：２単語以上の検索語が～文字以内の距離にある。
- 全角・半角の片方を入力すると自動的に両方を検索する。
- 類似検索、ファジー検索：類義語を自動的に検索する。
- クラスタリング：複数の文書を類似内容の文書群に自動的に分類する。
- 類似文書：ある文書と類似の文書を探す。
- Predictive Coding：ある観点で欲しい文書を自動的に選別、あるいは、重要度をつける。

WG2-2. アドバンス検索２

評価対象選定の考え方

有用性、使用頻度、代替方法の有無等の観点、評価データを準備できるか等。

選定されたアドバンス検索5項目

- 正規表現検索：カード番号等ではなく、京都と東京都の区別等日本語に関連するもの。
- 前後にアルファベットや数字が来ない、あるいは特定の文字が来ない言葉の検索。
- グルーピング：ブーリアンを（）で括って複雑な論理を組む。
- 近傍検索：２単語以上の検索語が～文字以内の距離にある。
- 全角・半角の片方を入力すると自動的に両方を検索する。区別しない場合と区別する場合の切り替えの有無も区別する。

WG2-2. アドバンス検索３

選定の経緯等１：正規表現検索

- 全員一致で必要だとの議論になったが、カード番号等ではなく、京都と東京都の区別等日本語に関連するもの。
- 英語では単語と単語との間にスペースを含むので「Tokyoto」と「Kyoto」の区別はできるが、日本語の場合、「京都」と検索したときに「東京都」を含むデータが検索にかかってしまう。
- ＊(アスタリスク)で、英字1文字も日本語1文字も検索出来る等、日本語に対して自由に正規表現検索ができるかどうか、の評価が欲しい。

選定の経緯等２：前後にアルファベットや数字が来ない、あるいは特定の文字が来ない言葉の検索。

- 全員一致で必要。

WG2-2. アドバンス検索４

選定の経緯等３：グルーピング：ブーリアンを()で括って複雑な論理を組む。

- 田中 not （太郎 or Taro.tanaka@abcd.com or 経理 。。。） とより精度の高い絞り込みによってレビューの時間を減らすことができ、非常に多用する。
- 複雑になりすぎて、わからなくなってしまったり、検索に負荷をかけてしまいがちで、英語でも日本語でも同様。
- 関連するドキュメントを抽出する際には必須。逆も然りで、何が関連しないかということも重要。関連しないドキュメントとキーワードが分かるだけでも論理式を活用する必要が有る。

選定の経緯等４：近傍検索　２単語以上の検索語が～文字以内の距離にある。

- 実務上は良く使う機能で便利。
- 具体的な使用例として、カルテルの調査で、「価格　w/5調整」(5文字以内)とすることによって、「価格の調整」、「価格を調整」、「価格が調整」、「価格の一部を調整」 などといった表現をヒットさせる。これをせずに「価格調整」をキーワードにすると前述のものはヒットしない。また、「価格 AND 調整」で検索してしまうと、「価格」と「調整」という一般的な単語が入っている文書はかなり多くヒットしてしまう。

WG2-2. アドバンス検索５

選定の経緯等５：全角・半角の片方を入力すると自動的に両方を検索する

- 区別しない場合と区別する場合の切り替えの有無も明示する。
- 評価用データには、全角／半角や、英語／日本語が混在していることを考慮する必要がある。
- 全角アルファベットは、日本語しか無いので評価用データにはいれるべきだが、全角英数をどの程度いれるのかも、検討する必要がある。

WG2-2. アドバンス検索６

選定の経緯等６：次回以降の検討課題として今回見送られた4項目

- 類似検索、ファジー検索：類義語を自動的に検索する。
- クラスタリング：複数の文書を類似内容の文書群に自動的に分類する。
- 類似文書：ある文書と類似の文書を探す。
- Predictive Coding：ある観点で欲しい文書を自動的に選別、あるいは、重要度をつける。

いずれも、評価基準と評価データの準備の両面で難しいということで今回の評価項目からは見送られ、次回以降の検討課題となった。

その他重要事項

第１２期第１回「日本語処理解析性能評価」分科会

2015.04.15
白井 喜勝

その他重要事項-１

１．解析速度

評価用データの用意に関連して、適切な評価方法を模索し、評価可能であれば評価する。以下、議論内容。

- 同一サンプル・同一環境の原則、測定環境の併記。
- サンプルデータを複数回コピーして測定可能な量にする。
- インデキシング等処理速度、検索自体の速度を区別する。

その他重要事項-2

2. 評価用データの作成・収集の基本的な考え方

- ソースとなるテキストを収集し、全角／半角の表記の揺らぎを作るなど、検索性能の評価が出来るようなテキストに調整する。
- IDFのメルマガコラムをソーステキストとする。
- 評価対象の各文字コードをプレーンテキストとし、評価を実施する。
- テキストを評価対象の各アプリケーションに保存し、評価を実施する。
- 各アプリケーションのファイルを添付したメールに対し、検索の評価を実施する。

3. 評価対象

- 評価に応募してきたメーカーの持ち込み形態に応じて評価を行う。
- ツールに対しても、検索エンジン自体に対しても評価を行う。
- 基本的に製品（ツール）を評価対象とする。