「日本語処理解析性能評価」分科会
検討内容詳細

平成28年4月22日
幹事　野﨑　周作（株式会社UBIC）
幹事　白井　喜勝（株式会社UBIC）

1．評価項目（1）

文字コード：9 種（プレーンテキストにて評価）

	文字コード種類	Codepage ID
1	UTF-8	65001
2	UTF-7	65000
3	UTF-16 LE	1200
4	UTF-16 BE	1201
5	Windows-31J	932
6	JIS	50220
7	JIS-Allow 1 byte Kana	50221
8	JIS-Allow 1 byte Kana - SO/SI	50222
9	EUC-JP	51932

- Windows では、Shift-JIS をShift-JIS（拡張部分含まない）とWindows-31J（拡張部分含む）で区別しておらず、Windows-31J を標準で使用しており、拡張文字を含むテキストをShift-JIS(拡張部分含まない)で変換してもWindows-31J の拡張部分が自動的に適用されるため、 Shift-JISはWindows-31J で代表して評価用データを作成する。
- プレーンテキストを使用した文字コード判別の評価においては、ヘッダーに頼らずに文字コードを正しく判別する事ができるかを評価する為、UnicodeのBOM（Byte Order Mark）無しで評価用データを作成する。

1. 評価項目（2）

アプリケーション：Eメール6種＋ドキュメント4種

	Eメール種類
1	Microsoft Outlook(MSG)
2	Microsoft Outlook Express / Windows Mail(EML)
3	Microsoft Outlook(PST)
4	Microsoft Outlook Express(DBX)
5	Becky! Internet Mail
6	Mozilla Thunderbird

- Lotus Notes（NSF）は評価用データ作成が困難な為、初回の評価項目からは除外

	ドキュメント種類
1	プレーンテキスト
2	Microsoft Word 2013（17ファイル形式）
3	Microsoft Excel 2013（24ファイル形式）
4	Microsoft PowerPoint 2013（28ファイル形式）

- Microsoft Officeを使用して選択可能な各種ファイル形式で保存することにより、各種ファイル形式の評価用データを作成。（例：pdf、htmなど）

1. 評価項目（3）

基本検索：2種

	評価項目	備考	検索クエリ
1	単一単語検索	単一単語	電子
		長文字列/区切	疑わしきは被告人の利
		短文字列	疑
		漢数字揺らぎ	21（半角）
			21（全角）
			二十一
			二一
		Windows-31Jコード評価判別（NEC特殊文字）	㈱①
		Windows-31Jコード評価判別（IBM拡張文字）	任ⅷ
		JISコード評価判別（半角カナ）	ｱｲｳｴｵ
		アスタリスク	*
		サロゲートペア	鋏齬朧頻剞腑
		改行の取り扱い	影響
2	2単語のブーリアン検索	ブーリアン検索（AND)	電子 AND 改ざん
		ブーリアン検索（OR)	電子 OR 問題
		ブーリアン検索（NOT)	電子 NOT 出版

1. 評価項目(4)

アドバンス検索:6種

	評価項目	備考	検索クエリ
1	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	証拠が後ろに付かない「電子」
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		前後にアルファベットが来ないctの検索
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。		(電子 OR 問題) AND (出版 OR 裁判)
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。	近傍検索の距離判定が文字数かバイト数かを判定する為に2条件で検索する	「電子」「改ざん」が15以内
			「電子」「改ざん」が30以内
			「改ざん」「電子」が15以内
			「改ざん」「電子」が30以内
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	全角・半角を区別する／区別しない、の切り替えができるかも確認する	アメリカ(全角)
			ｱﾒﾘｶ(半角)
			action(全角)
			action(半角)
			319(全角)
			319(半角)
			＆(全角)
			&(半角)
6	表記のゆらぎ	自動的に表記の揺らぎを検索する機能	ディジタル
			デジタル
			アメリカ(カを"漢字のちから")

2. 評価用データの作成

作成手順概要

- 3種のコラム(319号、345号、360号)原文に対して、(1)-(4)の評価に使用するキーワード及びクエリを検討。
  (評価項目(1)-(4)に適宜フィードバックし、評価項目を再設定。)
- 評価項目に必要な編集を319号及び345号に実施。360号は原文のみ使用し、合計5種類の評価テキストを作成。
- メモ帳に5種類の評価テキストをコピーし、テキストコンバーターツールにて評価項目(1)で設定した文字コードのテキストファイルに変換。
  それぞれ9種の文字コードのテキストファイルを評価用データとして作成。
- 3種のコラム(319号, 345号, 360号)原文テキストを使用し、Microsoft Officeを用いてアプリケーションの評価用データを作成。
- Eメールの評価用データは2つのWEBメールアカウント間で送受信したEメールを各種Eメールアプリケーションをインストールしたpcにダウンロードすることにより評価用データを作成。
  (メール本文にはコラム319号の原文テキストをコピーし、添付ファイルにはコラム345号の原文を用いたプレーンテキストファイル(UTF-8)を使用。)

２．評価用データの作成

評価用データ　ファイル名称のルール

nnn-mm-xx-vv-yy-z- NNNNNNNNNN.拡張子

- nnn: 対象コラム番号3桁（001-364）
- mm: メールアカウント分類
- xx: Microsoft分類
- vv: OfficeFile保存形式コード
- yy: 文字コード
- z: 変更の有無（0：原文、1：編集有り）
- NNNNNNNNNN: コラムタイトル先頭10文字

３．評価委員会募集要項内容

評価方法詳細は、これから設立されるＩＤＦ日本語処理解析性能評価委員会にて策定予定。

現時点での方針

①評価対象ソフトウェア

評価委員会にエントリーする組織が提供するソフトウェア。

②評価実施方法

- 評価に使用する評価用データ及び評価条件は、ＩＤＦ「日本語処理解析性能評価」分科会にて予め準備されたものを使用する。
- 評価に使用するハードウェア及びソフトウェアはエントリーする組織が準備を行う。
- ソフトウェアの最適な条件にて評価を行う為、エントリーする組織の技術者がソフトウェアの設定やオペレーションを行う。
- 第三者検証のため委員会メンバーによる立ち合いのもと評価を行う。

３．評価委員会募集要項内容

③応募資格

- デジタル・フォレンジック関連の日本語を処理するソフトウェアや、ソフトウェアを使用したサービスを提供しており、ＩＤＦ日本語処理解析性能評価委員会設立提案に賛同する企業等の組織。
- 評価対象となるソフトウェアを保持し、使用するハードウェアと共に委員会に貸し出し可能であること。
- 組織外にハードウェアを持ち出せない場合は、委員会メンバー立ち合いのもと評価が可能であること。
- 委員会で定める評価活動に技術者を派遣可能であること。

④評価スケジュール：未定（評価を受ける企業の意見を反映する。）

３．評価委員会募集要項内容

⑤評価結果

- 評価結果は現場で使用するユーザーに役立てて頂く為に公表する。
- 評価に使用したハードウェアのスペックや使用したソフトウェアのバージョンは参考情報として公表する。
- データ処理及び検索等の解析にかかる時間は公表しない。
- 最終的な公表方法に関しては評価委員会にて決定する。
- 認定、推奨等の言葉は使わず、結果のみを公表する。

⑥評価委員会にエントリーするメリット

- 評価用データは評価委員会による検証の後、ソフトェアの性能向上に向けた検証等に継続的に利用できる。
- 評価の結果ある一定の性能を示したソフトウェアには委員会認定ソフトウェアとすることを検討する。