

第16期第1回(通算7回目)「日本語処理解析性能評価」実施結果報告

1 評価結果の客観的評価指標

(1)「日本語処理解析性能評価」実施の目的(趣旨)と評価の実施について

近年、デジタル・フォレンジックや e ディスカバリ用途で、多様な検索機能や解析機能を持つ多くのソフトウェアが開発され、利用されています。しかしながら、海外で開発されたものも多く、日本国内で使用する場合に、どこまで日本語に対応しているかが不明で、ユーザーが使用してみるまでわからないのが現状です。また、実際に性能を評価しようとしても、客観的かつ有効な評価基準や指標も存在しないため、比較自体が困難な状況です。IDF「日本語処理解析性能評価」分科会ではこの状況を改善するために、日本語処理解析性能を評価するための基準となる項目とそれに伴う検索クエリ、さらに実際の評価に使うための評価用データの作成と各種ツールの日本語処理解析性能の評価基準項目を準備しました。また、本評価を実現するために日本語処理解析性能評価委員会が設立され、2017年1月に第1回評価が実施されました。

今期第16期第1回目(通算第7回目)の評価は、2020年3月に(バイシス・テクノロジー(株))の「Autopsy」(※2018年3月に評価受検した Autopsy のバージョンアップ版)の評価を実施致しましたので結果を報告致します。

今回評価受検したバージョンは 2018年3月に評価受検したバージョンと比較して評価用テキストデータの文字コードハンドリングが改善しており評価指標の向上が確認できました。また E メールやアプリケーションデータの評価はテストクエリにブーリアン検索(AND)を用いており、Autopsy がブーリアン検索(AND)に対応して頂ければ大幅な評価指標の向上が期待できます。

尚、これまでの7回の評価実施で得られた知見に基づき日本語処理解析性能評価委員会では、第17期に「評価基準」及び「評価用データ」についての補備・修正検討を行い、より適正な評価を実施できるように致します。第17期の「日本語処理解析性能評価」は、この補備・修正された「評価基準」及び「評価用データ」で実施致します。

(2)評価基準と評価用データ作成について

①評価基準について

単一単語検索や簡単なブーリアン検索を中心とした「基本検索」だけでなく、全角半角の同一視検索、正規表現や近傍検索といった比較的高度な検索機能を含む各種の「応用検索」も評価基準に含めました。尚、単一単語検索の項目にも「数字・漢数字同一視検索」といったものも含まれているので、「基本検索」といっても難易度は高いと想定されます。さらにそれぞれの評価用データとしては日本語特有の文字コードが多数と日本特有のEメールソフト等も対象に含みますので、総じて難易度の低いものから非常に高いと思われるものまで幅広く評価が可能なものとなりました。

②評価用データ作成手順について

評価用データは、IDFコラムより319号、345号、360号のそれぞれのテキスト情報と各種評価基準項目を検証する為に必要な追加修正を319号、345号に加えたものの合計5種類のテキスト情報をベースに、9種類の文字コードのテキストファイル、テキストファイルを添付した 6種類のEメールアプリケーションデータ、Microsoft Office(Word/Excel/PowerPoint)を用いた 各種保存形式の違いによる多数のファイル形式の評価用データを作成しました。

2 評価結果の客観的評価指標

各種評価基準項目ごとに実施しました評価用データに対する検索結果は、評価用データに「ヒットしたか」「ヒットしなかったか」、それはそれぞれ評価用データにおいて「正解としてヒットすべきものか」「正解としてヒットすべきでないか」という2つの軸が存在し、【表1】で表される4象限で表記できます。

【表1】評価結果の4象限

		評価用データにおける正解	
		ヒットすべきもの	ヒットすべきでないもの
検索結果	ヒットした	True Positive (TP)	False Positive (FP)
	ヒットしなかった	False Negative (FN)	True Negative (TN)

統計分析で良く使われる指標として、下記3つの指標がありますので、今回の評価結果を示す客観的数値として公表致します。

(1) Recall : 再現率

正解としてヒットすべきもののうち、検索で正しくヒットしたものの割合

$$\text{Recall : 再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

(2) Precision : 適合率

検索でヒットしたもののうち、正解としてヒットすべきものの割合

$$\text{Precision : 適合率} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

(3) Accuracy : 正解(答)率

検索でヒットしたものとヒットしなかったもの(全体)がそれぞれ正解である割合

$$\text{Accuracy : 正解(答)率} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

今回のバイシス・テクノロジー株式会社 (Autopsy) 評価結果を【表2】に記載します。

【表2】バイシス・テクノロジー株式会社 (Autopsy) 評価結果

テキスト 基本検索		評価用データにおける正解		合計	適合率
		ヒットすべき	ヒットすべきでない		
検索結果	ヒットした	175	0	175	100%
	ヒットしなかった	101	294		
合計		276		正解率	82%
再現率		63%			

テキスト 応用検索		評価用データにおける正解		合計	適合率
		ヒットすべき	ヒットすべきでない		
検索結果	ヒットした	225	0	225	100%
	ヒットしなかった	144	441		
合計		369		正解率	82%
再現率		61%			

Eメール 基本検索		評価用データにおける正解		合計	適合率
		ヒットすべき	ヒットすべきでない		
検索結果	ヒットした	5	0	5	100%
	ヒットしなかった	7	12		
合計		12		正解率	71%
再現率		42%			

アプリケーション 基本検索		評価用データにおける正解		合計	適合率
		ヒットすべき	ヒットすべきでない		
検索結果	ヒットした	0	0	0	-
	ヒットしなかった	64	2		
合計		64		正解率	3%
再現率		0%			

※各テーブル上でカウントされている数は評価用データにおける検索ヒット数ではなく、ヒットファイル数になります。

受検社バイシス・テクノロジー株式会社のコメント、問い合わせ先等

バイシス・テクノロジー株式会社	受検製品:Autopsy	Ver.4.14.1
製品特性	<p>Autopsy はオープンソースのデジタル・フォレンジック ツールです。豊富な機能と使いやすさを兼ね備え、Plugin モジュールによる機能拡張も可能で、オープンソースフォレンジックコミュニティの代表的なプラットフォームとなっています。複数のマシンによるネットワーク構成を組むことにより、複数の調査官による共同作業も可能です。</p>	
受検目的	<p>本製品の日本語処理解析の客観的な性能を知る。</p>	
評価結果へのコメント	<p>Autopsy の現行バージョン(4.14.1)に対する日本語処理解析性能を評価する機会をいただきありがとうございました。今回の評価に用いたバージョンは、日本語処理向けにいくつかの改善を施しています。</p> <ol style="list-style-type: none"> 1. キーワード検索を Unicode テキストで正規化 (4.8.0 で実装) 2. キーワード検索の Solr 検索インデックス化において、日本語について Kuromoji 形態素解析器を採用 3. 組込みファイル抽出モジュールでの ZIP ファイルの日本語ファイル名の問題を解決 <p>今回の性能評価では、Eメールやアプリケーションデータに対する正解率が、製品本来の機能に比して不本意なものとなってしまいましたが、今後のバージョンで改修してまいります。</p>	
問い合わせ先	<p>バイシス・テクノロジー株式会社 〒102-0093 東京都千代田区平河町 2-7-3 PMO 平河町 8F TEL:03-3511-2947 FAX:03-3511-2948 Email: info@basistech.jp</p>	
自社HP公開	<p>WEBサイト(https://www.basistech.jp/blog/)にて、本評価結果に関する自社コメントの公開を予定しております。また、詳細結果等の確認をご希望の場合は、上記問い合わせ先にご連絡下さい。</p>	

以上