第２回「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１４期第１回目（通算第２回目）の評価は、１０月より受検製品（企業）の募集を開始し、１２月に1社（(株）くまなんピーシーネット）の「Intella」の評価を実施致しましたので結果を報告致します。

（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

Recall　：再現率　＝　$\frac{TP}{TP+FN}$

（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

Precision ：適合率　＝　$\frac{TP}{TP+FP}$

（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

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

今回の株式会社くまなんピーシーネット（Intella Professional）評価結果を【表２】に記載します。

【表２】株式会社くまなんピーシーネット（Intella Professional）評価結果

テキスト　基本検索		評価用データにおける正解			
		ヒットすべき	ヒットすべきでない	合計	**適合率**
検索結果	ヒットした	220	0	220	**100%**
	ヒットしなかった	56	294		
	合計	276		**正解率**	**90%**
	再現率	**80%**			

テキスト　応用検索		評価用データにおける正解			
		ヒットすべき	ヒットすべきでない	合計	**適合率**
検索結果	ヒットした	288	0	288	**100%**
	ヒットしなかった	81	441		
	合計	369		**正解率**	**90%**
	再現率	**78%**			

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

アプリケーション　基本検索		評価用データにおける正解			
		ヒットすべき	ヒットすべきでない	合計	**適合率**
検索結果	ヒットした	56	0	56	**100%**
	ヒットしなかった	8	2		
	合計	64		**正解率**	**88%**
	再現率	**88%**			

※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（株）くまなんピーシーネットのコメント、問い合わせ先等

株式会社 くまなんピーシーネット	受検製品：Intella Professional	Ver.2.0.1
製品特性	Intellaは、簡単に言えば「検索ソフトウェア」です。２０１０年当時に今と殆ど変わらない日本語処理能力の高さに惚れ込み取り扱いを開始しました。膨大なデジタルデータが扱われる現在では、削除されたデータを探すことより、膨大なデータの中身を素早く簡単に知ることが重要と当社では考えています。多くのデータは誰かに見せるために作成するものであり、その情報を作成者が所有していなくても、他の第三者が所有していたりするからです。Intellaは、ユーザーが思いつく検索方法で結果を出し、ユーザーが見つけ易い形で結果を表示できます。これはIntellaが発表されて以降、世界中に広まった視覚的な表現方法であり、同一の結果でも視点を変えて見せることで、ユーザーが直感的に感じ取れた情報を探し出すことができる特徴的な機能です。Intellaを使えば、１００台を超えるパソコンでも、数百人分に及ぶメールボックスやスマートフォンの証拠ファイルでも、思いつくキーワードだけで重要な手掛かりを見つけ出すことができるようになります。	
受検目的	本製品の日本語処理解析の客観的な性能を知る。	
評価結果へのコメント	本製品の高い日本語解析能力を認めていただき、評価委員の皆様には感謝しています。ノーヒットの原因は些細なものばかりだったため、早急に改善してユーザーの利便性を今以上に向上させ、次回の受検に挑みたいと思います。	
問い合わせ先	株式会社くまなんピーシーネット 〒860-0834　熊本県熊本市南区江越２－１－８ ＴＥＬ：０９６－３７３－２２１３　ＦＡＸ：０９６－３７３－２２１４	
自社ＨＰ公開	ＷＥＢサイト（ https://www.kumanan-pcnet.co.jp ）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。	

以上