情報検索のための単語分割一貫性の定量的評価

高橋 文彦　　颯々野 学
ヤフー株式会社
{ftakahas, msassano}@yahoo-corp.jp

1 はじめに

日本語を扱う検索システムでは，一般的にインデックスの作成と検索の際に自動単語分割を行う．インデックス作成では，検索の対象となるドキュメントを単語に分割し，単語をそのドキュメントのインデックスタームとして登録する．検索を実行する際には，検索クエリを単語分割し，インデックスと照らし合わせることでドキュメントを探す．現在，単語分割器や単語分割を部分タスクとする形態素解析器は文脈を考慮した手法が一般的であるが，検索のトークナイザ[1]として用いる際に，文脈の有無によって解析結果の揺れが発生し，検索漏れの問題が起こる．例えば，”京都大学に行く”というドキュメントが”京都大 | 学 | に | 行く”と単語分割されインデックスが作られた場合，クエリ”京都大学”に対して”京都 | 大学”と単語に分割するとこのドキュメントを引き当てることができない．一方で，ドキュメントが”京都 | 大学 | に | 行く”と単語分割されてインデキシングされれば，検索で引き当てることができる．したがって検索における単語分割では一貫した解析が重要になる．この検索漏れの問題は，正解の単語単位に分割できるかという指標だけでは評価できない．これは先述した例において，ドキュメントもクエリも”京 | 都大学”と分割しても，検索で引き当てられるが単語分割としては誤っているという現象が起こりうるためである．

そこで本研究では，解析揺れによる検索漏れの問題を定量的に評価するために，情報検索のための単語分割の一貫性を評価する指標を検討する．さらに，その指標を用いていくつかのトークナイザを評価する実験を行い，その結果を議論する．

2 関連研究

日本語の単語分割や形態素解析の評価は，ドキュメントに対して人手で単語境界がアノテーションされた正解コーパスに対して，解析結果がどれほど近いかが議論されてきた [1, 2]．本研究で扱う解析揺れによる検索漏れの問題は，このような評価だけでは議論できない．

[1]本稿では，単語分割器や単語分割を部分タスクとする形態素解析器などのテキスト解析器をトークナイザと呼ぶ．トークナイザで単語に分割することを単語分割と呼び，またトークナイザによって分けられた単位を単語と呼んでいる．

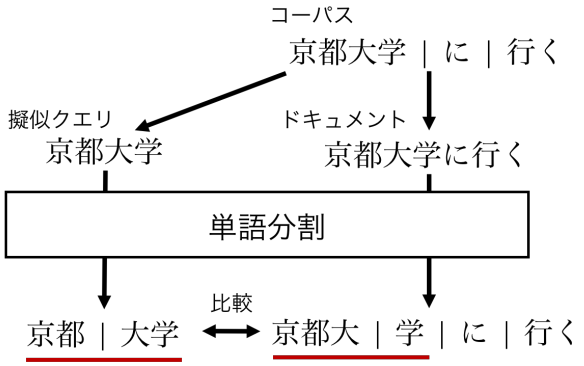

図 1: 評価指標 tokenization disagreement の概略図

本研究では単語境界の揺れをトークナイザの解析揺れで議論するが，そもそも単語境界の定義の曖昧性を指摘する論文もある [3, 4]．文献 [3] では，従来の bag of words の代わりに単語境界の期待値をベクトルとして検索する方法を提案している．

情報検索や機械翻訳といった応用の処理への単語分割の影響はいくつかの研究で議論がある [5, 6, 7]．日本語と同様に単語の区切りが明示されていない言語に中国語があるが，Chang ら [7] は中国語の機械翻訳で，単語分割のバリエーションのエントロピーで一貫性を評価し，機械翻訳精度との関係を調査している．本研究では，日本語の情報検索における単語分割の一貫性を評価し議論する．

3 一貫性の評価指標

単語分割の一貫性を評価するために 2 つの指標を用いる．

3.1 Tokenization disagreement

検索におけるドキュメントとクエリでの単語分割の一貫性を評価するために，擬似的にクエリとクエリを含んだドキュメントの対を作成し，それぞれの単語分割結果を比較したエラー率で評価する（図 1）．擬似クエリとドキュメントの対の数を N，擬似クエリとドキュメントの単語分割の完全一致数を C とした時，下式で計算する．この評価指標を本論文では，tokenization disagreement と呼ぶ．この評価指標は，値が低いほど

一貫した解析を意味する．

$$TA = \frac{N - C}{N}$$

擬似クエリとドキュメントは，単語境界がアノテーションされたコーパスから作成する．名詞の単語を擬似クエリとして抽出し，抽出元のドキュメントを検索対象のドキュメントとして，擬似的なクエリとドキュメントを得る．

この評価値は検索の再現率を担保するための指標であり，適合率を保証する指標でなない．しかし，情報検索では再現率の方が重要な指標であると言われている [8]．

3.2 Tokenization entropy

前節の tokenization disagreement は，文脈の有り無しでの単語分割の一貫性を評価する指標である．文脈の違いによる単語分割の一貫性を評価するために，Chang らの評価指標 [7] を用いて解析結果のエントロピーで評価する．単語 w_i に対する，ドキュメントの単語分割結果の w_i に対応する部分 v_{ij} の条件付きエントロピーとして以下の式で定義される．

$$\begin{aligned} H(V|W) &= -\sum_{w_i} P(w_i) \sum_{v_{ij}} P(v_{ij}|w_i) log P(v_{ij}|w_i) \\ &= -\sum_{w_i} \sum_{v_{ij}} P(v_{ij}, w_i) log P(v_{ij}|w_i) \end{aligned}$$

w_i は，単語境界がアノテーションされたコーパスの名詞を対象とする．この評価指標を本論文では，tokenization entropy と呼ぶ．tokenization entropy は，値が低いほど一貫した解析を意味する．

4 実験

3 章の評価指標を用いて，トークナイザを評価する．

4.1 実験設定

９つのトークナイザを比較する．公開されている一般的なテキスト解析ソフトウェアとして，MeCab[1] 0.996，JUMAN 7.0.1[2]，KyTea[2] 0.4.7，ChaSen 2.3.3[3]，kuromoji[4]，を用いる．kuromoji は，全文検索システム Apache Solr[5] などで使われている．mecab-ipadic[6]，mecab-unidic[7]，mecab-jumandic[8]，mecab-ipadic-neologd[9] をそれぞれを用いて比較する．

[2]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[3]http://chasen-legacy.osdn.jp
[4]https://github.com/atilika/kuromoji
[5]http://lucene.apache.org/solr
[6]http://sourceforge.net/projects/mecab/files/mecab-ipadic/2.7.0-20070801
[7]https://osdn.jp/projects/unidic/releases/58338
[8]http://sourceforge.net/projects/mecab/files/mecab-jumandic/5.1-20070304
[9]https://github.com/neologd/mecab-ipadic-neologd/tree/v0.0.4

表 1: 実験で用いるコーパス

	文字数	単語数	名詞単語数
BCCWJ	2,024,771	1,281,600	106,793
KNB	123,226	58,114	15,398

表 2: 擬似クエリや w_i として用いる名詞の単語の一部

BCCWJ	KNB
一冊	ケイタイ中心
お天気カメラ	昨晩
持ち合い株解消	破裂
組織選挙	努力
事前広報	相手選手
一万七千八百八十五ｋＨｚ	大文字山
調査会社フロストアンドサリバン	ユーザ側
赤単色	下宿生
ＮＨＫ首都圏営業推進センター	漕暦
樹脂製	百人一首

ChaSen の辞書は，ipadic-2.7.0[10] を利用する．また，上記の統計的な手法との比較として最長一致法にルールベースで修正を加えた手法 [9](AsagiDFS) を比較する．AsagiDFS は 64 万文の JUMAN の出力結果からルールの学習をした．

3 章で定義した tokenization disagreement と tokenization entropy はコーパスの品詞体系 [11] に影響した結果になることが予想される．このため，複数の品詞体系で定義されたコーパスを用いて比較する必要がある．従ってコーパスには，UniDic 基準の BCCWJ[10] と JUMAN 基準の KNB コーパス [11] (以下，KNB) を選んだ．コーパスの詳細を表 1 に示す．BCCWJ では長単位を単語単位として扱い，KNB では接尾辞と接頭辞をそれぞれ単語の前後に繋げ名詞連続を一語にする処理を施した．それぞれコーパス中の名詞を，tokenization disagreement の擬似クエリ，tokenization entropy の w_i とした．したがって，表 1 の名詞単語数が tokenization disagreement の N である．表 2 に，実験で擬似クエリや w_i として用いる，名詞の単語をランダムに 10 件表示する．

4.2 コーパスの品詞体系と評価指標の関係

評価指標 tokenization disagreement と tokenization entropy は，品詞体系や単語単位などの基準に従ってアノテーションされたコーパスを用いる．このため，コーパスのアノテーション基準に影響した結果になる事が予想される．この節では，コーパスの品詞体系と評価指標の関係について実験的に調査した．

図 2 に tokenization disagreement ，図 3 に tokenization entropy の，各トークナイザの BCCWJ と KNB での結果を示す．横軸が BCCWJ，縦軸が KNB での評価指標の値である．tokenization disagreement は KNB と BCCWJ で相関した結果が現れた．一方で，tokenization entropy は，トークナイザの品詞体系ごとに傾向が異なった．JUMAN 基準に準拠

[10]https://osdn.jp/projects/ipadic/downloads/24435/ipadic-2.7.0.tar.gz
[11]本稿では品詞体系に単語の単位や品詞の定義を含む．

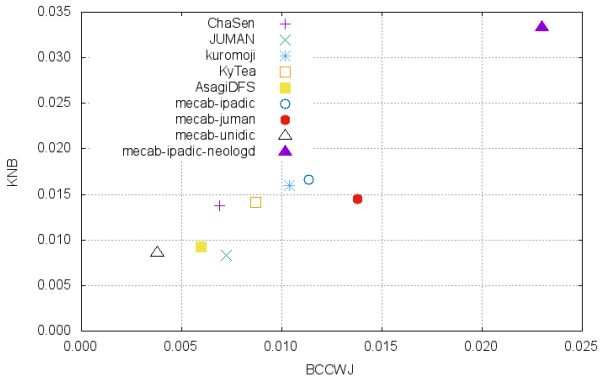

図 2: tokenization disagreement の値

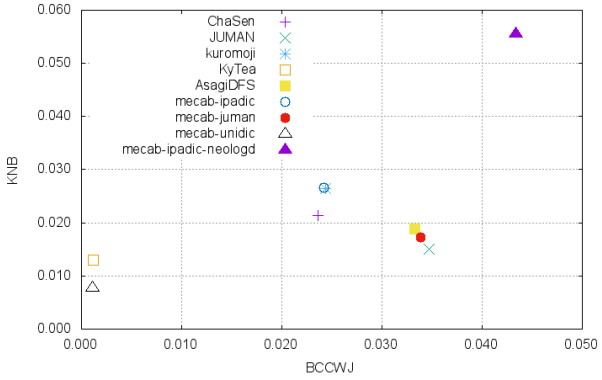

図 3: tokenization entropy の値

する解析器 (JUMAN, mecab-juman, AsagiDFS) は KNB(JUMAN 基準) ではエントロピーが低く，BCCWJ(UniDic 基準) でエントロピーが高い傾向となった（グラフの右下に寄る傾向）．UniDic 基準に準拠する解析器 (mecab-unidic, KyTea) は KNB でエントロピーが高く，BCCWJ でエントロピーが低い傾向（グラフの左上に寄る傾向）となった．したがって，tokenization disagreement よりも tokenization entropy は品詞体系に依存しやすい評価基準といえる．

4.3 トークナイザの比較

表 3 に各トークナイザの評価指標の値を示す．また，tokenization disagreement の実際に一致していないクエリとドキュメントを表 4 に示す．

表 5 に KNB に対する JUMAN 基準に従うトークナイザ（JUMAN, mecab-juman, AsagiDFS）の単語分割精度を示す．単語分割精度は，正解コーパスと単語単位でアライメントを取り，再現率と適合率の調和平均で求めた．単語分割精度では AsagiDFS が最も精度が低いが，tokenization disagreement では mecab-juman よりも一貫した結果になっていることがわかる．これは，AsagiDFS がクエリでもドキュメントでも同じ解析誤りをしているためだと考えられる．実際に AsagiDFS のクエリとドキュメントで一致しているが単語分割を誤っているものを調べると，“ｍｉｘ｜ｉ” や “ニン｜テン｜ドー” など未登録語を多く含んでいた．このことから単語分割精度が高いことが必ずしも一貫性が高いことに対応するわけではないことが確認できた．

トークナイザの品詞体系による差を議論するために，mecab-ipadic, mecab-unidic, mecab-juman を比較する．それぞれのトークナイザは IPADic 基準，UniDic 基準 (短単位)，JUMAN 基準の辞書を用いているので，辞書の品詞体系が異なる．実験の結果は tokenization disagreement でも tokenization entropy でも，mecab-unidic が最も値が低く mecab-ipadic と mecab-juman は同程度だった．UniDic 基準は，他の品詞体系に比べて短い単位を単語として扱うため，重複した部分文字列の辞書エントリが少なく，解析が一貫していたと考えられる．実際に辞書エントリの部分文字列の一致率を計算すると，mecab-unidic が最も低かった．

トークナイザの語彙サイズによる差を議論するために，mecab-ipadic, mecab-ipadic-neologd を比較する．mecab-ipadic-neologd は，mecab-ipadic の辞書を拡張したものであり，それぞれの語彙サイズは 200 万語程度，40 万語程度である．準拠する品詞体系やモデルは同じで，これらの大きな違いは語彙である．結果は tokenization disagreement でも tokenization entropy でも，語彙サイズの大きい mecab-ipadic-neologd の方が一貫性が低かった．mecab-ipadic-neologd は長い単位を単語として定義するため検索のクエリと合わない問題が発生する．例えば，表 4 のように，ドキュメントで”ハリーポッターとアズカバンの囚人”を一語として扱うためにクエリ”ハリーポッター”とマッチしない．解析結果をみると同様のケースが多く見られた．辞書の未登録語の部分で解析揺れが多く発生すると考えられるため，未登録語を減らす手段として辞書に単語を追加する方法が考えられるが，単に辞書に単語を追加すれば検索漏れを軽減できるわけではないことが確認できた．

また，mecab-ipadic, ChaSen を比較すると，tokenization disagreement でも tokenization entropy でも ChaSen の方が mecab より低い値となっており，ChaSen の方が一貫性が高いことがわかる．どちらも IPA 基準に準拠したトークナイザであるが，パラメータ推定のモデルや前後のアドホックな処理が異なる．HMM モデルは長い単語を出力しやすい最長一致法に似た傾向があるが，この傾向により ChaSen が CRF を使っている mecab-ipadic よりも一貫性が高かったと考えられる．実際に CRF と HMM のモデルの比較をするためには，前後のアドホックな処理の他に辞書や学習コースを揃えて比較する必要がある．

4.4 ドメインごとの比較

図 4 に，BCCWJ に含まれるドメインごとに各トークナイザで平均を取った評価指標の値を示した．tokenization disagreement は確率値であり，tokenization entropy はエントロピーである．どちらの評価指標で

表 3: トークナイザの比較

	tokenization disagreement		tokenization entropy	
	BCCWJ	KNB	BCCWJ	KNB
ChaSen	0.69%	1.38%	0.0236	0.0214
JUMAN	0.72%	**0.83%**	0.0347	0.0150
kuromoji	1.04%	1.60%	0.0244	0.0265
KyTea	0.87%	1.41%	0.0012	0.0130
AsagiDFS	0.60%	0.92%	0.0333	0.0189
mecab-ipadic	1.14%	1.66%	0.0242	0.0266
mecab-juman	1.38%	1.45%	0.0339	0.0172
mecab-unidic	**0.38%**	0.86%	**0.0011**	**0.0077**
mecab-ipadic-neologd	2.30%	3.33%	0.0434	0.0556

表 4: 単語境界が一致しないクエリとドキュメント．”|”は単語境界を表す．

トークナイザ	クエリの解析結果	ドキュメントの解析結果
mecab-ipadic-neologd	ハリーポッター	ハリーポッターとアズカバンの囚人 | 以外 | です | か |?
mecab-ipadic	マドンナ | リリー	オーガニック | な | マドンナリリー | の | もつ | 植物 | 性 | 作用 | 物質
JUMAN	行方 | 不明	どっか | 行方 | 不明に | なった | まま

表 5: KNB に対する単語分割精度

	JUMAN	mecab-juman	AsagiDFS
F-mesure	0.9631	0.9524	0.9494

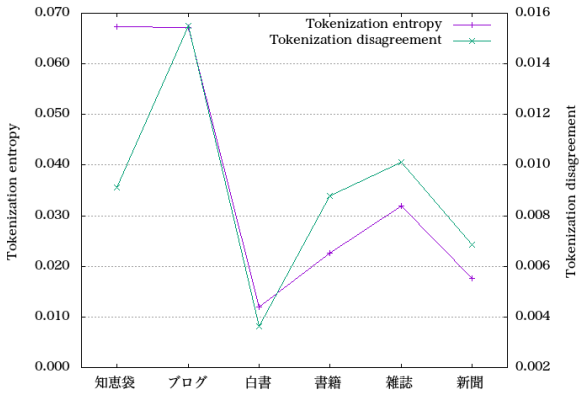

図 4: ドメインごとの比較

も低い値が一貫していることを示す．

どちらの評価指標でも同様の傾向が見られ，ブログや知恵袋などで値が高く，白書や新聞で値が低い．ブログや知恵袋は，多様な表現が含まれるドキュメントであり，未登録語の頻出や語の省略が多用されるため [12]，解析誤りにより一貫した解析が困難であったと考えられる．一方で白書や新聞は，一定の語で書かれ学習コーパスと同じドメインまたは近い書かれ方をしているため，解析誤りも少なく，解析結果の揺れが起こりにくいと考えられる．

5 おわりに

本研究では，単語分割の解析揺れによる検索漏れの問題を指摘し，評価方法を提案すると共に単語分割の一貫性の観点からトークナイザの評価を行った．実験により，単語分割精度が高いことが一貫性の高さに対応するわけではなく，UniDic 辞書が一貫性に効果的であり，単に辞書に単語を追加すればよいわけではないことまた，CRF よりも HMM の方が一貫した解析をする可能性があることを確認した．モデルによる解析揺れの影響の比較は，さらに調査が必要である．

本研究で提案した tokenization disagreement は，検索の再現率を保証する指標である．これは例えばすべての文字単位で区切っても高い評価になる．情報検索に対する単語分割の影響を総合的に評価するためには適合率も評価する指標が必要である．今後は検索の適合率も考慮した指標を提案すると共に，より一貫性の高いモデルや一貫性を保った辞書の拡張方法を提案したい．

参考文献

[1] Kudo, Yamamoto, and Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *EMNLP*, pp. 230–237, 2004.

[2] Neubig, Nakata, and Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *ACL-HLT*, pp. 529–533, 2011.

[3] 工藤. 形態素周辺確率を用いた分かち書きの一般化とその応用. 第 11 回言語処理学会年次大会, 2005.

[4] 萩原, 関根. 半教師あり学習に基づく大規模語彙に対応した日本語単語分割. 第 18 回言語処理学会年次大会, 2012.

[5] Sudoh, Nagata, Mori, and Kawahara. Japanese-to-English patent translation system based on domain-adapted word segmentation and post-ordering. In *AMTM*, pp. 234–248, 2014.

[6] Peng, Huang, Schuurmans, and Cercone. Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR. In *COLING*, pp. 1–7, 2002.

[7] Chang, Galley, and Manning. Optimizing Chinese word segmentation for machine translation performance. In *StatMT*, pp. 224–232, 2008.

[8] Halstead, 奥村. ロバストな日本語形態素解析-辞書依存性の低いハイブリットアルゴリズムの提案-. 情報処理学会第 54 回全国大会, 1997.

[9] Sassano. Deterministic word segmentation using maximum matching with fully lexicalized rules. In *EACL*, pp. 79–83, 2014.

[10] Maekawa. Balanced corpus of contemporary written Japanese. In *Workshop on Asian Language Resources*, pp. 101–102, 2008.

[11] 橋本, 黒橋, 河原, 新里, 永田. 構文・照応・評価情報つきブログコーパスの構築. 自然言語処理, Vol. 18, No. 2, pp. 175–201, 2011.

[12] Takahasi and Mori. Keyboard logs as natural annotations for word segmentation. In *EMNLP*, pp. 1186–1196, 2015.