雑談対話システムにおける対話破綻検出器の評価尺度の検討

Finding appropriate metrics for evaluating dialogue breakdown detectors in chat-oriented dialogue systems

角森唯子[1*]　東中竜一郎[2]　高橋哲朗[3]　稲葉通将[4]
Yuiko Tsunomori[1]　Ryuichiro Higashinaka[2]　Tetsuro Takahashi[3]　Michimasa Inaba[4]

[1] 株式会社 NTT ドコモ　NTT DOCOMO, INC.
[2] NTT メディアインテリジェンス研究所　NTT Media Intelligence Laboratories
[3] 株式会社 富士通研究所　Fujitsu Laboratories Ltd.
[4] 広島市立大学　Hiroshima City University

Abstract: The task of detecting dialogue breakdown, the aim of which is to detect whether a system utterance causes dialogue breakdown in a given dialogue context, has been actively researched in recent years. However, currently, it is not clear which evaluation metrics should be used to evaluate dialogue breakdown detectors, hindering progress in dialogue breakdown detection. In this paper, we propose finding appropriate metrics for evaluating the detectors. In our approach, we first enumerate possible evaluation metrics and then rank them on the basis of system ranking stability and discriminative power. By using the submitted runs (results of dialogue breakdown detection of participants) of the dialogue breakdown detection challenge, we experimentally found that MSE(NB+PB,B) and MSE(NB,PB,B) are appropriate metrics for dialogue breakdown detection.

1　はじめに

近年，雑談対話システムの需要が増加している [1, 2]．しかしながら，雑談が扱う話題は多様かつ複雑であるため，システムがユーザ発話を正しく理解できず，対話破綻（ユーザが対話を継続できなくなる状態）を引き起こす発話を行うことは少なくない [3]．

このような状況を避けるために，対話破綻検出チャレンジ（DBDC）と題し，人間と対話システムとの間で生じる対話破綻を自動検出することを目的とした，評価型ワークショップが開催されている [4]．参加者は，対象のシステム発話に対して，破綻ラベル（B: breakdown, PB: possible breakdown, NB: not a breakdown）[1]と各ラベルの確率を出力する破綻検出器を構築する．対話破綻ラベルの定義は次の通りである：

NB：破綻ではない 当該システム発話のあと対話を問題無く継続できる．

PB：破綻と言い切れないが，違和感を感じる発話 当該システム発話のあと対話をスムーズに継続することが困難．

B：あきらかにおかしいと思う発話．破綻 当該システム発話のあと対話を継続することが困難．

たとえば，以下では S1 から U2 までが対話文脈であり，下線が引かれた S3 が後続するシステム発話である．

S1: 買い物は一人が楽です
U1: 確かに気が楽ですね
S2: 買い物は長いです
U2: まあ見るのも楽しいし
S3: 買い物は一緒が楽しいですね

ここで，S3 は S1 の発話内容と食い違ったことを言っており，対話破綻を引き起こす可能性が高い．すなわち，このような発話について，対話破綻検出器は，B あるいは PB を出力できれば正解となる．

DBDC では，対話破綻検出器の評価に複数の尺度が用いられており，それらは正解ラベルとの一致を評価する尺度（ラベル一致系統）と，正解の分布と比較して評価する尺度（分布距離系統）の 2 つに分けられる．30 人でアノテーションを行い，多数決で正解ラベルを決定するとともに，ラベルの分布を正解の分布としている．しかしながら，どの尺度が適しているかは明確ではないという課題がある．

*連絡先：株式会社 NTT ドコモ
東京都港区赤坂 2 丁目 4 番 5 町
E-mail: yuiko.tsunomori.fc@nttdocomo.com

[1]対話破綻検出チャレンジ 1 と 2 では，NB を○，PB を△，B を×と表していた．

本稿では，対話破綻検出器を評価するための適切な尺度について検討する．まず，これまでに提案されたものと，今回新たに提案するものを併せた全 22 種類の評価尺度を列挙する．次に，順位安定性とシステム弁別性に基づき，評価尺度をランキングすることで，最適な評価尺度を決定する．

2 関連研究

雑談対話システムにおいて，問題があるシステム発話の検出について研究が行われている．例えば，Xiang ら [5] は，ユーザの発話意図やセンチメントを特徴量として使用し，問題があるシステム発話の検出を行っている．最近では，DBDC [6] が 3 回開催されており，問題を引き起こす発話の検出への関心はさらに高まっている．その一方で，対話破綻検出器においてどんな評価尺度が妥当かについては，ほとんど研究されていない．過去の DBDC では，9 つの評価尺度が使用されていたが，最適な尺度が明確でなかったため，最良の検出器を決めることや，参加者のチューニングが困難であるという課題が残されていた．

本稿では，情報検索（IR）研究 [7] で使用されている手法を参考にし，順位安定性とシステム弁別性の基準に基づいて最適な評価尺度を実験的に検討する．なぜなら，IR 研究では，複数の評価者の評価結果と比較してシステム出力を評価しており，IR 研究における評価尺度を選定する枠組みは，対話破綻検出にも応用できると考えられるからである．なお，対話研究においては，最適な評価尺度を見つけるために相関が使用されている [8] が，対象がスカラー値の場合にしか直接適用できない．我々は，対話破綻検出のラベルの分布を正解の分布としていることから，相関ベースの方法を適用することは困難であると考えられる．

3 提案手法

我々は，対話破綻検出器の評価に適切な尺度を実験的に求める．まず，DBDC で使用されていた尺度に加え，本稿で新たに提案する尺度を加えた，全 22 種類の評価尺度を列挙する．次に，IR 研究で使用されている順位安定性とシステム弁別性に基づき，尺度のランキングを行う．ランキングを行う際に使用するデータとして，DBDC に提出された run（参加者の破綻検出器の結果）を使用する．

3.1 評価尺度の候補

DBDC において，ラベル一致系統と分布距離系統の 2 系統の評価尺度が使用されている [4]．

ラベル一致系統 検出器にシステム発話が破綻かどうか NB/PB/B のラベルのいずれかで出力させ，アノテーション結果の多数決で決めた正解ラベルとの一致を評価する．

分布距離系統 検出器にシステム発話が破綻かどうかを確率分布で出力させ，評価データ（各発話毎の NB/PB/B の頻度分布）との近さを評価する．

DBDC では全部で 9 種類の尺度が使用されていた．本稿では有効と考えられれるいくつかの尺度を追加し，全 22 種類の評価尺度について検討する．表 1 に，本稿で使用する尺度すべてを示す．(2)–(6)，(9)–(10)，(14)–(16) と (20)–(22) が新たに追加した尺度である．

アノテーションの一致率が高い発話は，一致率が低い発話よりも重要度が高いと考えていることから，重み付きの尺度 (4)–(6)，(9)–(10)，(14)–(16) および (20)–(22) を提案する．本稿では，重み付けにシンプソン係数を使用し，各発話の重み w は次の式で求める．

$$w = \sum_{l \in \{NB,PB,B\}} p_l^2, \tag{1}$$

p_l は各ラベル l の確率を表している．例えば，分布が $(p_{NB}, p_{PB}, p_B) = (0.33, 0.33, 0.33)$ の場合は $w = 0.33$，$(p_{NB}, p_{PB}, p_B) = (0.0, 0.0, 1.0)$ の場合は $w = 1.0$ と計算される．このように，アノテーションの一致率が高い発話ほど，重み w の値も大きくなる．表 1 において，重み付きの評価尺度は “+w“で表している．このような重みの使用は，“unanimity-aware gain“においても提案されており [9]，よりシステム弁別性が向上することが示唆されている．

3.2 評価尺度の基準

最適な評価尺度を選ぶために，IR 研究で利用されている，順位安定性とシステム弁別性 [7] の 2 つの基準を使用する．これらの指標の算出に使用するデータとして，複数の対話破綻検出器の結果（DBDC では run と呼ばれている）を用いる．

順位安定性 適切な評価尺度は，データセットが異なっていたとしても，run のランキング結果が同じものになるべきだと考えられる．順位安定性は，複数の異なるデータセットにおいて，run の順位が安定しているかどうかを評価する．まず，ランダムにサンプリングして作成した複数のデータセットを用意し，各データセットに対して各尺度で run のランキングを行う．そして，各尺度におけるランキングペアについて順位相関係数を計算し，この値を用いて順位安定性の値とする．

表 1: 評価尺度．“+w“は重み付き尺度であることを示している．重みは式 (1) で計算される．

	尺度	説明
		ラベル一致系統
(1)	Accuracy(NB,PB,B)	対話内のそれぞれのシステム発話について，対話破綻検出器が出力したラベル（予測ラベル）とその正解ラベルを比較し，正しく分類されたラベルの数をラベルの総数で割ることによって計算する．
(2)	Accuracy(NB,PB+B)	PB と B を単一ラベルとみなした場合の (1)
(3)	Accuracy(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (1)
(4)	Accuracy+w(NB,PB,B)	$c_n = \begin{cases} 1, & \text{予測ラベルと正解ラベルが一致する場合;} \\ 0, & \text{それ以外;} \end{cases}$ $\text{Accuracy} = \frac{\sum_{n=1}^{N} c_n w_n}{\sum_{n=1}^{N} w_n}$ n は発話インデックス，N は発話の総数，w は重みを表す．
(5)	Accuracy+w(NB,PB+B)	PB と B を単一ラベルとみなした場合の (4)
(6)	Accuracy+w(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (4)
(7)	F1(B)	対話内のそれぞれのシステム発話について，予測ラベルとその正解ラベルを比較し，B ラベルの Precision と Recall の調和平均を算出する．Precision と Recall の定義は（9）を参照．
(8)	F1(PB+B)	PB と B を単一ラベルとみなした場合の (7)
(9)	F1+w(B)	$pred_n(labels) = \begin{cases} 1, & labels \text{ に予測ラベルが含まれる場合;} \\ 0, & \text{それ以外;} \end{cases}$ $gold_n(labels) = \begin{cases} 1, & labels \text{ に正解ラベルが含まれる場合;} \\ 0, & \text{それ以外;} \end{cases}$ $\text{TP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{B}) w_n$ $\text{FP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{NB, PB}) w_n$ $\text{TN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{NB, PB}) w_n$ $\text{FN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{B}) w_n$ $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$ $\text{F1} = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
(10)	F1+w(PB+B)	PB と B を単一ラベルとみなした場合の (9)
		分布距離系統
(11)	JSD(NB,PB,B)	各システム発話について，対話破綻検出器が出力した分布（予測分布）と正解の分布から，Jensen-Shannon divergence を計算する．
(12)	JSD(NB,PB+B)	PB と B を単一ラベルとみなした場合の (11)
(13)	JSD(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (11)
(14)	JSD+w(NB,PB,B)	重み付けした (11)．重み w は式 (1) で計算される．
(15)	JSD+w(NB,PB+B)	PB と B を単一ラベルとみなした場合の (14)
(16)	JSD+w(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (14)
(17)	MSE(NB,PB,B)	各システム発話について，予測分布と正解の分布から，平均二乗誤差を計算する．
(18)	MSE(NB,PB+B)	PB と B を単一ラベルとみなした場合の (17)
(19)	MSE(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (17)
(20)	MSE+w(NB,PB,B)	重み付けした (17)．重み w は式 (1) で計算される．
(21)	MSE+w(NB,PB+B)	PB と B を単一ラベルとみなした場合の (20)
(22)	MSE+w(NB+PB,B)	NB と PB を単一ラベルとみなした場合の (20)

システム弁別性 適切な評価尺度は，run の差異に敏感であるべきだと考えられる．そこで，各尺度において，評価したすべての run のペアについて，その有意差を検定し，有意に異なると判別できるペアの数を求める．最も多くの有意差を認めることができた尺度を，システム弁別性が高い評価尺度とみなす．

4 実験

順位安定性とシステム弁別性に優れる適切な尺度を実験的に求める．なるべく言語に依存しない評価尺度をもとめるために，英語と日本語それぞれの run に対して評価尺度のランキングを行う．DBDC3 では，英語と日本語の二つのトラックがあるため，それらを用いることで実現できる．その平均順位を計算することで，両言語で上位にランキングされる尺度を選択する．本章では，使用したデータセットと，それぞれ基準の値を計算する手順について述べる．

4.1 データセット

本稿では，DBDC3 の英語と日本語のデータセット[2]と，DBDC3 の参加者が提出した run を使用する．

DBDC3 データセット 英語の 4 システムと，日本語の 3 システムから対話データが収集されている．システムごとに 50 対話あり，合計で 350 対話が収録されている．すべての対話は，20 か 21 発話からなり（ユーザが 10 発話，システムが 10 発話か 11 発話），すべてのシステム発話は 30 人のアノテータによって対話破綻ラベルがアノテーションされている．

使用する run DBDC3 では，1 つの参加チームについて，各言語 3 つまで run を提出できる．英語と日本語それぞれ 12 ずつの run の提出があった．各チームの手法としては，ニューラルネットワークやサポートベクターマシン（SVM），ランダムフォレストを使用した様々な手法が見られた．

4.2 手順

順位安定性については，異なるデータセットの run における順位相関係数を使用する．本稿では，Kendall's τ [10] を使用して，順位相関係数を求める．英語と日本語それぞれのデータセットについて，最初にすべてのデータをマージする．次に，マージされたデータからランダムに 20 %ずつサンプリングすることによって，2 つのサブセットのデータを作成する．各サブセットの run をランキングを行い，これらの順位間で Kendall の τ を計算する．安定した結果を得るために，このプロセスをすべての尺度について 500 回繰り返し，Kendall の τ の平均値を求める．

システム弁別性については，まず，英語と日本語の各データセットそれぞれにおいて，すべての run ペアに対して有意差がある割合を計算し，その割合によって尺度をランキングする．次に，英語と日本語の平均順位を計算する．これらの手順を，すべての尺度に対して実施する．システム弁別性の計算には `Discpower`[3] [11] を使用する．

4.3 結果

表 2 に，順位安定性のランキング結果を示す．全体として，分布距離系統の尺度（MSE，JSD）の方が，ラベル一致系統の尺度よりも優れていた．分布距離系統の尺度の中でも，英語と日本語の平均順位が最も高いことから，MSE(NB+PB,B) が最も安定性が高いということが示された．また，重み付けされた尺度は、重み付けされていない尺度と比較して，あまり機能していないことがわかった．

表 3 に，システム弁別性の結果を示す（$\alpha = .05$）．すべての run ペアについて，有意差があった run の割合を示す．日英ともに run の数は 14 であるため，すべての run ペアの数は $\binom{14}{2} = 91$ である．分布距離系統の尺度（MSE，JSD）の方が，ラベル一致系統の尺度よりも優れていた．また，重み付けされた尺度の順位が低いことから，順位安定性の結果と同様に，重み付けはシステム弁別性にもあまり寄与していないことがわかった．

4.4 最適な評価尺度の選定

表 4 に，順位安定性とシステム弁別性の平均順位における上位 5 つの評価尺度を示す．MSE(NB+PB,B) および MSE(NB,PB,B) は，同じ平均順位であり，どちらも最も良い評価尺度であることがわかった．

MSE と JSD が上位にあることから，分布距離系統の方がラベル一致系統の尺度よりも適切であることが確認できる．これは，分布距離系統の尺度は，ラベル一致系統においてラベルに変換するときに失われる情報も考慮できているためだと考えられる．また，NB と B が単一のラベル (NB+PB,B) とみなされる場合と，すべてのラベルを区別する場合 (NB,PB,B) との間に差異

[2] `https://dbd-challenge.github.io/dbdc3/data/`

[3] `http://research.nii.ac.jp/ntcir/tools/discpower-en.html`

表 2: 順位安定性の結果（各言語における 1–3 位は太字にしている）

	英語		日本語		
尺度	Kendall's τ	順位	Kendall's τ	順位	平均順位
MSE(NB+PB,B)	0.81	**3**	0.85	**2**	**2.5**
MSE(NB,PB,B)	0.79	6	0.86	**1**	**3.5**
MSE+w(NB+PB,B)	0.82	**2**	0.83	5	**3.5**
JSD(NB+PB,B)	0.81	4	0.83	4	4.0
JSD+w(NB+PB,B)	0.82	**1**	0.77	9	5.0
JSD(NB,PB,B)	0.77	12	0.85	**3**	7.5
JSD+w(NB,PB+B)	0.79	5	0.63	13	9.0
MSE(NB,PB+B)	0.78	11	0.77	8	9.5
JSD(NB,PB+B)	0.78	10	0.74	10	10.0
MSE+w(NB,PB+B)	0.78	8	0.68	12	10.0
MSE+w(NB,PB,B)	0.73	14	0.82	6	10.0
JSD+w(NB,PB,B)	0.75	13	0.78	7	10.0
Accuracy(NB+PB,B)	0.79	7	0.58	16	11.5
Accuracy+w(NB+PB,B)	0.78	9	0.61	15	12.0
Accuracy+w(NB,PB,B)	0.3	21	0.68	11	16.0
F1+w(B)	0.66	16	0.5	17	16.5
F1(B)	0.66	15	0.48	18	16.5
Accuracy(NB,PB,B)	0.26	22	0.63	14	18.0
F1(PB+B)	0.65	17	0.21	20	18.5
Accuracy(NB,PB+B)	0.62	18	0.18	21	19.5
Accuracy+w(NB,PB+B)	0.56	20	0.26	19	19.5
F1+w(PB+B)	0.61	19	0.14	22	20.5

表 3: システム弁別性の結果（各言語における 1–3 位は太字にしている）

	英語		日本語		
尺度	有意差がある run ペアの割合	順位	有意差がある run ペアの割合	順位	平均順位
MSE(NB,PB,B)	67.0	6	76.9	**2**	**4.0**
MSE(NB+PB,B)	70.3	**2**	70.3	8	**5.0**
JSD(NB+PB,B)	67.0	6	74.7	4	**5.0**
MSE+w(NB+PB,B)	68.1	4	72.5	7	5.5
MSE(NB,PB+B)	68.1	4	64.8	9	6.5
MSE+w(NB,PB,B)	61.5	12	76.9	**2**	7.0
Accuracy(NB+PB,B)	71.4	**1**	52.7	14	7.5
JSD+w(NB+PB,B)	62.6	9	73.6	6	7.5
JSD(NB,PB,B)	60.4	14	81.3	**1**	7.5
JSD(NB,PB+B)	63.7	8	64.8	9	8.5
Accuracy+w(NB+PB,B)	70.3	**2**	50.5	16	9.0
JSD+w(NB,PB,B)	60.4	14	74.7	4	9.0
MSE+w(NB,PB+B)	61.5	12	60.4	11	11.5
F1+w(B)	62.6	9	50.5	16	12.5
F1(B)	62.6	9	48.4	18	13.5
JSD+w(NB,PB+B)	59.3	16	59.3	12	14.0
Accuracy+w(NB,PB,B)	19.8	21	58.2	13	17.0
Accuracy(NB,PB,B)	14.3	22	52.7	14	18.0
F1(PB+B)	56.0	17	15.4	22	19.5
Accuracy(NB,PB+B)	52.7	18	16.5	21	19.5
F1+w(PB+B)	50.5	19	17.6	20	19.5
Accuracy+w(NB,PB+B)	37.4	20	20.9	19	19.5

表 4: 順位安定性とシステム弁別性の平均順位

尺度	順位安定性の順位	システム弁別性の順位	平均順位
MSE(NB+PB,B)	2.5	5.0	3.8
MSE(NB,PB,B)	3.5	4.0	3.8
JSD(NB+PB,B)	4.0	5.0	4.5
MSE+w(NB+PB,B)	3.5	5.5	4.5
JSD+w(NB+PB,B)	5.0	7.5	6.3

がないこともわかった．この理由として，NB+PB と B を区別することは，3 つのラベルを区別することと同等に難しいためと考えられる．これを検証するために，対話破綻アノテーションのアノテータ間の一致率

（Fleiss' κ）を計算した．英語のデータセットにおいて，すべてのラベルを区別した場合，一致率は $\kappa = 0.065$ であった．NB と PB を単一のラベルとみなした場合は $\kappa = 0.077$ であり，PB と B を単一のラベルとみなした場合は $\kappa = 0.095$ であった．日本語においても同様の傾向が見られた．これは，NB と PB+B の区別に比べて，NB+PB と B の区別がすべてのラベルを区別することと同様に困難であることを示しており，我々の推測を裏付けている．

順位安定性とシステム弁別性の結果によると，重み付けされた尺度は有効性は確認できなかった．この理由として，重みは単に推測しやすい問題を際立たせるだけで，推測が難しい問題に寄与しないため，run を区別することには寄与しないのではないかと考えられる．

5　まとめと今後の課題

本稿では，対話破綻検出器の評価に適した尺度を検討する手法について提案を行った．まず，考えられる評価尺度を列挙し，順位安定性とシステム弁別性に基づいてランキングを行った．DBDC に提出された run を使用することで，実験的に MSE(NB+PB,B) と MSE(NB,PB,B) が最も適した尺度であることがわかった．最後に，もし最適な評価尺度を 1 つに決めるとすると，アノテーションコスト削減の観点から，MSE(NB+PB,B) を使用することが良いのではないかと考えられる．

今後は，今回列挙した尺度では不十分な可能性もあることから，さらに有効と考えられる尺度についても追加検討を行う予定である．また．複数の尺度を組み合わせて使用することで，より最適な評価尺度について検討する．本稿では重み付き尺度は有効ではないという結果であったが，直感的にはアノテーションの一致率が高い発話の重みが大きいことは重要だと考えられる．他の重み付けの方法についても検討し，今回有効ではなかった理由についてさらに調査する予定である．最後に，我々の提案した評価尺度で対話破綻検出器を評価していくことで，対話破綻検出器の精度を向上に貢献し，さらには雑談対話システム全体の改善にもつなげたい．

参考文献

[1] Richard S Wallace. The anatomy of alice. In *Parsing the Turing Test*, pp. 181–210. Springer, 2009.

[2] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pp. 928–939, 2014.

[3] Bilyana Martinovski and David Traum. Breakdown in human-machine interaction: the error is the clue. In *Proc. ISCA tutorial and research workshop on Error handling in dialogue systems*, pp. 11–16, 2003.

[4] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proc. LREC*, pp. 3146–3150, 2016.

[5] Yang Xiang, Yaoyun Zhang, Xiaoqiang Zhou, Xiaolong Wang, and Yang Qin. Problematic situation analysis and automatic recognition for chinese online conversational system. In *Proc. CLP*, pp. 43–51, 2014.

[6] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. Overview of dialogue breakdown detection challenge 3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*, 2017.

[7] William Webber, Alistair Moffat, and Justin Zobel. The effect of pooling and evaluation depth on metric stability. In *Proc. EVIA*, pp. 7–15.

[8] Marilyn Walker, Candace Kamm, and Diane Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, Vol. 6, No. 3-4, pp. 363–377, 2000.

[9] Tetsuya Sakai. The effect of inter-assessor disagreement on IR system evaluation: A case study with lancers and students. *Proc. EVIA*, pp. 31–38.

[10] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, Vol. 30, No. 1/2, pp. 81–93, 1938.

[11] Tetsuya Sakai. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Digital Courier*, Vol. 3, pp. 625–642, 2007.