広告文におけるスパン単位の誤り推定

上垣外英剛 [1,2]　村上聡一朗 [2]　張培楠 [2]
[1] 奈良先端科学技術大学院大学　[2] 株式会社サイバーエージェント
kamigaito.h@is.naist.jp, {murakami_soichiro,zhang_peinan}@cyberagent.co.jp

概要

広告文はユーザーに対して商品やサービスの価値を簡潔かつ正確に伝える重要な役割を担う。しかし、広告文中に文法的あるいは意味的な誤りが含まれる場合，ユーザーの理解を妨げるだけでなく、誤った情報を伝達してしまう危険性がある。従来の広告文評価では，文全体を単位とした品質推定が主であり、どの部分にどのような誤りが存在するかを詳細に捉えることは困難であった。本研究では，広告文中の誤りをスパン単位で検出し、その種類と深刻度を同時に推定する手法を提案する。さらに既存の広告文データセットを拡張し、スパン単位の誤りに対するアノテーション付きデータセットを構築した。実験の結果、提案手法は複数の事前学習済み言語モデルで有効に機能し、推定された誤りの重大さが人手評価と高い相関を示すことを確認した。

1　はじめに

広告文はユーザーの興味を惹き、商品やサービスの認知および購買行動を促進するための重要な要素である。特に検索キーワード連動型広告においては、限られた文字数の中で宣伝対象の内容を正確かつ魅力的に伝えることが求められる [1]。

このような背景から、広告文の品質を自動的に評価する技術は、広告配信の最適化や品質管理の観点から重要性を増している。実際に、広告文の品質はクリック率やユーザー満足度と強く関連することが報告されている [2, 3]。

一方で、広告文の評価にはいくつかの課題が存在する。従来研究の多くは、広告文全体を一つの単位として品質を推定する手法を採用しており [4]、文中のどの部分にどのような問題が存在するのかを明示的に扱うことは少なかった。そのため、評価結果の解釈性が低く、広告文の具体的な改善につなげにくいという問題がある。

さらに、広告文中の誤りは単なる文法的な不自然さにとどまらず、宣伝対象や内容の誤伝達を引き起こす可能性がある。このような意味的誤りは、読解不能な広告文よりも深刻な影響を及ぼす場合があり、誤情報の拡散という観点からも看過できない問題である [5]。

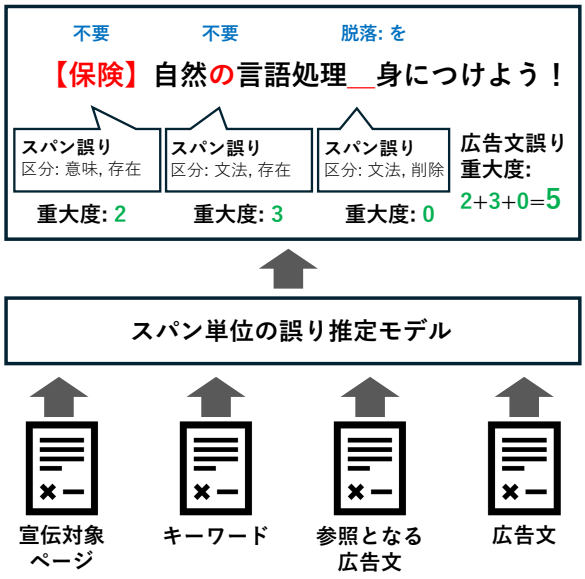

図 1　スパン単位の誤り推定の概要

本研究ではこれらの課題に対し、広告文中の誤りをスパン単位で捉え、誤りの種類と深刻度を同時に推定する手法を提案する。スパン単位での誤り推定は、機械翻訳や文法誤り訂正などの分野では有効性が示されているが [6, 7]、広告文評価への応用はこれまで十分に検討されてこなかった。

本研究の主な貢献は以下の三点である。(1) 広告文中の誤りをスパン単位で定義し、文法的・意味的観点および深刻度を考慮した評価枠組みを提案する点、(2) 既存の広告文データセット CAMERA [8] を拡張し、スパン単位の誤りに対するアノテーション付きデータセットを構築した点、(3) 複数の事前学習済み言語モデルを用いた実験により、提案手法の有効性を実証した点である。

作成したデータセットを用いた実験の結果、提案

表 1　スパン単位の誤りの深刻度

深刻度	定義
0	広告の意味が理解でき、広告対象に変化はない。
1	広告の意味が理解できない。
2	広告の意味は理解できるが広告対象が変化している。

表 2　スパン単位の誤り区分

種類	操作	定義
文法	削除	必要な機能語が削除されている。
文法	存在	誤った機能語が存在している。
意味	削除	必要な内容語が削除されている。
意味	存在	誤った内容語が存在している。

するスパン単位の誤り推定は高い精度で誤りを検出できること、また推定された誤りの重大さが人手評価と高い相関を持つことが確認された。

2　提案法：スパン単位の誤り評価

2.1　スパン単位の誤りの定義

広告文においては文法の誤りにより読解が難しいこと以外にも、読解そのものは可能ではあるが宣伝対象が誤って伝わることや、異なった宣伝内容が伝わるなどの問題も生じ得る。広告が情報拡散手段の一つであることを踏まえると、読解困難な場合よりも誤情報の拡散が誤りとしてはより深刻である。本研究ではこの点を踏まえ、スパン単位のエラーを表 1 に記載するように 3 種類定義した。なお、これらはあくまでも広告文中のスパンに対して与えられるものであり、これ自体が即ち広告文の評価となるわけではないことに注意が必要である。

また、誤りの区分についても文法上の誤りと意味上の誤りの二種類を定義する。深刻度の定義で説明したように、文法上の誤りが意味上の誤りを誘発する可能性があるため、両者の厳密な切り分けは難しい。そこで、簡易な区分けとして、機能語に関連する誤りと内容語に関連する誤りの 2 種類を考える。さらに、それぞれの誤りについて、誤りを生じる過程に着目し、必要な語や句の脱落に起因する誤りと、不要な内容語の挿入または誤った内容語の存在に起因する誤りの 2 種類を考える。最終的にこれらの直積を考え、表 2 が示すように 4 種類の誤り区分が定義される。

2.2　スパンを用いた広告文の評価

前節で説明したスパンに対する誤りの深刻度と区分に基づく広告文自体の評価を行う。この評価では、与えられた広告文 T と、対応する誤りスパン e の集合 E に対して、T の誤りがどれほど重大であるかを推定する。ここで e の誤りの深刻度を $s(e)$、誤りの区分を $c(e)$ と定義する。この際に、$s(e)$ をそのまま誤りの重大さの推定に使用することもできるが、誤りの区分ごとに重大度が変化することも考慮し、e の誤りの重大度を $w_{<c(e),s(e)>}$ と定義する。これらを用いて、与えられた広告文 T が持つ誤りの重大さは次のように計算される:

$$\mathrm{ErrorScore}(T) = \sum_{e \in E} w_{<c(e),s(e)>}. \tag{1}$$

なお、本研究が初めて広告文の評価においてスパン単位のエラーの重大さを利用する関係上、$w_{<c(e),s(e)>}$ の適切な重み付けが定かではない。そこで、本稿では $c(e)$ に関わらず一様に $s(e)$ を $w_{<c(e),s(e)>}$ として扱う。

3　データセットの作成

スパン単位の誤り推定を行い、その妥当性を検証するためにはデータセットが必要となる。この目的のために、我々は既存の広告文データセットである CAMERA を拡張する形で新たなデータセットを作成した。具体的には CAMERA の各広告文に対し機械的にスパン単位の誤りを挿入し (§3.1)、それらの深刻度を人手でアノテーション (§3.2) した後に、各誤りを組み合わせることでデータ拡張 (§3.3) を行う。各手順の詳細を下記の各小節にて説明する。

3.1　誤りの挿入

誤りの挿入では、§2.1 での定義を遵守した誤りを広告文中で生じさせるため、機械的な規則を用いて行う。文法と意味のそれぞれの区分からの誤り挿入の規則を表 3 に示す。品詞の推定には IPA 辞書を用いた MeCab [9] を適用した。キーワードについては CAMERA の各広告文に対応して含まれている検索キーワードを利用する。ただし、CAMERA のキーワードには助詞のみを含む不適切なものも含まれるため、内容語を含まないキーワードを除外した。また検索キーワード連動型の広告として、検索キーワードを一つも含まない広告は不適切であるため、それらも除外している。キーワードを削除する際は

表 3　誤りを挿入するための規則

区分	操作	概要
文法	削除	助詞を一箇所削除
	存在	助詞を入れ替え
意味	削除	キーワードを削除
	存在	キーワードが複数存在する場合に入れ替え
		キーワードを同じ業種の別の広告のものと入れ替え
		キーワードを異なる業種の別の広告のものと入れ替え
		キーワードが複数存在する場合に他のキーワードを文頭に追加
		同じ業種の別の広告のキーワードを文頭に追加
		異なる業種の別の広告のキーワードを文頭に追加

表 4　データセットの統計 (広告文の数)

合計	訓練	開発	テスト
2,146	999	143	1,004

可読性を担保するため、後続する格助詞も同時に削除している。なお、CAMERA に含まれる広告文を参考に、キーワードを文頭に追加する際には【】で囲んで追加している。既に【】または「」で囲まれた単語が存在する場合にはキーワードの強調が行われていると判断し、キーワードの追加は行わない。

3.2　アノテーション

表 3 に基づいて擬似的に作成された各誤り箇所に対し、表 1 で定義される深刻度を付与するためのアノテーションを実施する。アノテーションは広告文の評価に関する専門家 3 名によって行う。アノテート時には誤りを挿入する前と後のそれぞれの広告文が提示され、誤り箇所が表 1 におけるいずれの深刻度に該当するかを選択しなければならない。人的資源の制限により、1000 件のスパン単位の誤りを対象とした。アノテーションの結果、3 人の一致率は Fleiss's kappa [10] で 0.61 となり、強い一致が見られた [11]。なお、各事例のラベルの決定はアノテータの判断による多数決で行い、3 人の決定が互いに異なる事例は除外した。この過程を経て、最終的に 989 件の事例が残った。アノテーション結果の詳細については付録 A に記載する。

表 5　文法と意味の誤りの結合。下線は削除の対象を、__ は削除された箇所を表す。

原文	保険と共済の違いを動画で紹介
内容: キーワードを削除	__共済の違いを動画で紹介
文法: 助詞を削除	保険と共済の違いを動画__紹介
拡張	共済の違いを動画紹介

3.3　データ拡張

スパン単位の誤りは区分が異なる場合に相互に影響を与えることがないため、表 5 に示すように、同じ広告文に付与された文法と意味それぞれの誤りを組み合わせることが可能である。この手順により、最終的に 989 件の事例を 2146 件まで拡張することに成功した。これらの事例を同一分割内に同一の広告文とキーワードの組が存在しないように訓練、開発、テストデータに分割する。表 4 にそれぞれの事例数を示す。データセットの詳細な統計については付録 B に記載する。

4　スパン単位の誤り推定

スパン単位の誤りの推定では、広告文 $T = \{t_1, \cdots, t_n\}$ 中の各トークン t_i に対してスパンの開始と種類に対応するラベルの系列 $L = \{l_1, \cdots, l_n\}$ を予測する系列ラベリング問題として定式化される。各ラベル l_i は BIO タグ に基づいて定義され，B は誤りスパンの開始，I は同一スパン内の継続，O は誤りに該当しないトークンを表す。さらに、B および I には誤り区分に対応するラベル $c \in C$ を付与し、B-c，I-c の形式で表現する。ここで C は表 2 で定義された誤り区分の集合である。その上で B-c，I-c には誤りの深刻度に対応するラベル $s \in S$ を付与し、B-c-s，I-c-s の形式で表現する。なお、S は表 1 で定義された誤りの深刻度の集合である。

モデルは各トークン t_i に対してラベル l_i を予測し，隣接する B-c-s と I-c-s の列により誤りスパンを同定する。この定式化により、広告文中に存在する複数かつ可変長の誤りスパンを同時検出し、各誤りの種類と深刻度を推定することが可能となる。

なお、実際の広告文は広告対象の Web ページと広告キーワードに紐づいてユーザーに提示されるため、本研究ではこの状況を踏まえ、広告キーワード K と宣伝対象のページ P を広告文 T に結合してモデルに入力する。この際、モデルは T に対応するトークン箇所に対してスパン単位の誤りの推定を行う。

表 6　作成したデータセットにおける誤り推定の結果。P は Precision を、R は Recall を、r はピアソンの積率相関係数を、ρ はスピアマンの順位相関をそれぞれ表す。**太字のスコア**は各評価尺度で最高であったことを示す。

モデル	参照広告文あり					参照広告文なし				
	P	R	F_1	r	ρ	P	R	F_1	r	ρ
bert-base-multilingual-cased	.311	.326	.296	.610	.517	.377	.369	.362	.359	.376
xlm-roberta-base	.272	.169	.153	.363	.267	.111	.023	.038	.070	.058
mmBERT-base	.659	.585	.598	.714	.678	.235	.141	.171	.233	.193
llm-jp-modernbert-base	.698	.593	.612	.692	.667	.358	.324	.332	.368	.311
modernbert-ja-130m	**.755**	**.723**	**.719**	**.799**	**.783**	**.524**	**.422**	**.440**	**.431**	**.402**

さらに、広告文の品質を推定する上で、人手の参照広告文 R が利用できる場合とそうではない場合が存在する。両者に対応するため、人手の参照を利用する際は K, P, R, T の順序で結合を行い入力とする。なお、この際もモデルは T に対応するトークン箇所に対してスパン単位の誤り推定を行う。

5　実験

§3 にて作成したデータセットを用いて、§4 で説明したスパン単位の誤り推定に対する評価を行う。

5.1　実験設定

データセット　学習及びテストには表 4 に記載されているデータセットの分割を使用した。なお、入力形式として参照を使用する場合は、広告キーワード K, 宣伝対象のページ P, 参照 R, 広告文 T をこの順序で事前学習済みモデル固有の専用トークンを用いて結合した。参照を使用しない場合は K, P, T の順序で同様に結合した。

モデル　モデルには HuggingFace Transformers [12] から利用できる事前学習済みの Transformer エンコーダとして bert-base-multilingual-cased [13], xlm-roberta-base [14], mmBERT-base [15], llm-jp-modernbert-base [16], modernbert-ja-130m [17] を使用した。系列ラベリングには HuggingFace Transformers の固有表現抽出用スクリプト[1)]を転用し、ハイパーパラメータとして初期値を使用して学習を行った。詳細な設定については付録 C に記載した。

評価尺度　スパン単位の誤り推定を評価するために、スパン単位の Precision, Recall, F_1 を使用した。F_1 を報告する際には各クラスの事例数で F_1 を重み付き平均したものを報告している。さらに、式 1 に従って計算される広告文が持つ誤りの重大さが、人手によるものと相関しているかを確認するために、両者の間でのピアソン相関 r とスピアマンの順位相関 ρ を計算する。

1）https://github.com/huggingface/transformers/blob/main/examples/pytorch/token-classification/run_ner.py

5.2　実験結果

表 6 に実験結果を示す。参照が存在する場合、提案法による広告文の評価はピアソンの積率相関及びスピアマンの順位相関が xlm-roberta を除くモデルにおいて高い値を達成した、実用的な方法であることが分かる。さらに、誤りスパンの特定に関しても ModernBERT [18] に基づくモデルでは高い値を達成しており、こちらも実用的であると言える。

その一方で、参照が存在しない場合、これらの値は大きく低下してしまうものの、モデルによっては広告文の評価において中程度の人手評価との相関が存在する。さらに誤りスパンの特定に関しても、一部の参照あり設定の結果を超える結果も存在しており、一定の成果を上げていると考えられる。

6　まとめ

本研究では、広告文評価における解釈性と実用性の向上を目的に、スパン単位で誤りの種類と深刻度を推定する手法を提案した。提案法の貢献は文法および意味的な誤りを区別し、それぞれの誤りが広告文全体の品質に与える影響を定量化する枠組みを示した点にある。また、既存データセットを拡張することで、訓練及び評価に使用可能な、スパン単位の誤り推定に適した新たなデータセットを構築した。

実験の結果、提案手法は複数の事前学習済み言語モデルにおいて有効であり、推定された誤りの重大さが人手評価と整合することが示された。今後の課題としては，誤り区分ごとの重み付けの最適化や，実際の広告配信システムへの応用，さらに生成モデルによる広告文生成との統合などが挙げられる。

参考文献

[1] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In **Proceedings of the 16th International Conference on World Wide Web**, WWW '07, p. 521–530, New York, NY, USA, 2007. Association for Computing Machinery.

[2] Shaunak Mishra, Changwei Hu, Manisha Verma, Kevin Yen, Yifan Hu, and Maxim Sviridenko. Tsi: An ad text strength indicator using text-to-ctr and semantic-ad-similarity. In **Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management**, CIKM '21, p. 4036–4045. ACM, October 2021.

[3] Penghui Wei, Xuanhua Yang, Shaoguo Liu, Liang Wang, and Bo Zheng. Creater: Ctr-driven advertising text generation with controlled pre-training and contrastive fine-tuning, 2022.

[4] Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. AdTEC: A unified benchmark for evaluating text quality in search engine advertising. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 7672–7691, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[5] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. **Science**, Vol. 359, No. 6380, pp. 1146–1151, 2018.

[6] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In **Proceedings of Translating and the Computer 35**, London, UK, November 28-29 2013. Aslib.

[7] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, editors, **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 163–170, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics.

[8] Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. Striking gold in advertising: Standardization and exploration of ad text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 955–972, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[9] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] Joseph L Fleiss. Measuring nominal scale agreement among many raters. **Psychological bulletin**, Vol. 76, No. 5, p. 378, 1971.

[11] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. **biometrics**, pp. 159–174, 1977.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[15] Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. mmbert: A modern multilingual encoder with annealed language learning, 2025.

[16] Issa Sugiura, Kouta Nakayama, and Yusuke Oda. llm-jp-modernbert: A modernbert model trained on a large-scale japanese corpus with long context length, 2025.

[17] Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. ModernBERT-Ja. https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a, 2025.

[18] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics.

付録

表 7　各アノテータが選択したラベルとその頻度

アノテータ 1	
1: 広告の意味が理解できない	459
0: 広告の意味が理解でき、広告対象に変化はない	283
2: 広告の意味は理解できるが広告対象が変化している	258
アノテータ 2	
1: 広告の意味が理解できない	398
2: 広告の意味は理解できるが広告対象が変化している	309
0: 広告の意味が理解でき、広告対象に変化はない	293
アノテータ 3	
1: 広告の意味が理解できない	469
2: 広告の意味は理解できるが広告対象が変化している	331
0: 広告の意味が理解でき、広告対象に変化はない	200

A　アノテーション

表 7 に各アノテータが選択したラベルの頻度を示す。理解が困難な広告文に対してはアノテータ間で傾向は同様であるものの、理解可能な誤りに対する判断の違いには内容の解釈に対する個人差の存在が分かる。

B　データセット統計

表 8 に各データ分割におけるラベルの頻度を示す。B タグと I タグの頻度の比率から複数のトークンに跨るような長大なスパンはあまり多くはないことが分かる。また、CAMERA データセットに含まれる広告文が広告タイトルである関係上、助詞があまり含まれておらず、結果として意味ラベルの方が文法ラベルよりも多い傾向にある。

C　ハイパーパラメータ

学習時に使用したハイパーパラメータについて説明する。事前学習済み言語モデルの重みを引き継いだ後に訓練エポック数に 3 を、学習器には AdamW を使用し、初期学習率には 5e-05、バッチサイズは 8 に設定し fine-tuning を実施した。学習にあたっては 1 枚の NVIDIA RTX2080Ti を使用した。

表 8　データセットの各分割における各ラベルの頻度

訓練	
B-意味: 存在-深刻度: 1	355
B-文法: 存在-深刻度: 1	308
B-意味: 削除-深刻度: 2	227
B-意味: 存在-深刻度: 2	170
I-意味: 存在-深刻度: 1	166
B-文法: 削除-深刻度: 1	45
I-意味: 存在-深刻度: 2	31
B-意味: 削除-深刻度: 1	22
I-文法: 存在-深刻度: 1	8
B-文法: 削除-深刻度: 2	5
開発	
B-意味: 存在-深刻度: 1	70
B-意味: 削除-深刻度: 2	29
B-意味: 存在-深刻度: 2	17
I-意味: 存在-深刻度: 1	12
B-文法: 削除-深刻度: 1	9
B-意味: 削除-深刻度: 1	6
B-文法: 存在-深刻度: 1	6
テスト	
B-文法: 存在-深刻度: 1	479
B-意味: 存在-深刻度: 2	312
B-意味: 存在-深刻度: 1	283
B-意味: 削除-深刻度: 2	210
I-意味: 存在-深刻度: 1	98
I-意味: 存在-深刻度: 2	72
B-文法: 削除-深刻度: 1	31
B-文法: 削除-深刻度: 2	19
I-文法: 存在-深刻度: 1	13
B-意味: 削除-深刻度: 1	1