医療テキストからの重要因子抽出とその性能評価

Performance evaluation and extraction of significant factors from medical text

山下貴範 [1,2]　若田好史 [1]　中島直樹 [1]　廣川佐千男 [3]

Takanori Yamasita[1,2], Yoshifumi Wakata[1], Naoki Nakashima[1], and Sachio Hirokawa[3]

[1] 九州大学病院メディカル・インフォメーションセンター
[1] Medical Information Center, Kyushu University Hospital
[2] 九州大学大学院システム情報科学府
[2] Graduate School of Information Science and Electrical Engineering, Kyushu University
[3] 九州大学情報基盤研究開発センター
[3] Research Institute for Information Technology, Kyushu University

Abstract: The digitalization of medical treatment has progressed and huge amounts of medical data are accumulating. Electronic medical data include structured numerical data and unstructured text data. The medical text analysis is expected to improve medical process and the clinical decision support. The present paper analyzes the words that appear in operation records to predict the two peaks and the long hospitalization by Support vector machine, and evaluated them by Feature selection. Three measures were proposed and the prediction performance for importance of the feature word was evaluated. Two measures obtained that less than 10 words resulted in the optimal prediction performance. Moreover, it was confirmed the effect of medical dictionary.

1　はじめに

1.1　背景

病院情報システムの普及に伴い各医療機関においては診療の電子化が進み，膨大な診療データが蓄積されてきている. 診療データには, 患者属性や病名，検査値，処方などのいわゆるコードが付与された構造化数値データと，コードが付与されていないテキストデータや画像データなどの非構造化データがある．診療におけるテキストデータには診療録（カルテ）や退院サマリ，検査レポート，IC，看護記録などがある．これらは，主に医師や看護師により記録される本質的なデータである．

構造化数値データに対する分析は，統計解析や BI ツールなどにより整備されてきている．一方，非構造化テキストデータに対する研究事例は，重要語の抽出や単語関係の可視化[1，2，3]，病名の分類[4]，副作用関係の自動抽出[5]などがあるが，直接医療現場にフィードバックできるような分析法は，まだ多くない．その分析法を確立することは，医療プロセスの意思決定や予防医療の支援に期待されており[6]，医療の質向上に大きく寄与するものである．

1.2　研究目的

我々は，診療の意思決定支援や診療プロセス改善を目的として，非構造化テキストデータに対して，テキストマイニングと機械学習手法を適用し，目的変数に対する重要因子（単語）を抽出し，それを評価する試みを行っている[7，8].

本稿では，九州大学病院においてクリニカルパスを適用した人工股関節置換術施行症例の手術記録から，術後在院日数分布の特徴と長期在院の特徴の 2 つを目的変数として, [9] で提案された SVM(support vector machine)と属性選択の手法を適用した. その推定性能の評価について，全単語と医学辞書を適用した場合について比較を行ったので，それらの結果と今後の課題について考察する．

2　分析データ（人工股関節置換術症例の手術記録）

九州大学病院において, 2008 年 1 月から 2014 年 3 月の期間に実施された人工股関節置換術施行症例の手術記録を 871 件収集した．本症例の標準術後在

院日数は，2012 年 1 月までは 27 日設定であり，2012 年 2 月以降は 25 日に設定されている．術後在院日数の決定要因となる可能性のある単語を，871 件の手術記録から抽出するため，文書単位の検索エンジンを構築した．検索エンジンの構築は NII で公開されている GETA[1]を利用した．

術後在院日数の分布について，21 日と 28 日が大きなピークであり（図 1），21～23 日目を peak A，27～29 日目を peak B とした．

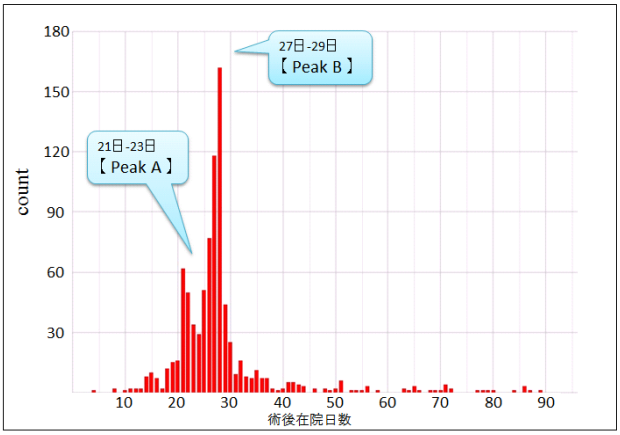

図 1：術後在院日数ヒストグラム

3　SVM による推定

3.1　術後在院日数ピークの推定

全文書に出現する単語の数は 2,144 個である．また，全文書に対して医学辞書単語と適合した単語は 406 個であった．全単語と医学辞書単語を比較するために，それぞれの検索エンジンを構築した．

そして SVM を適用し，Peak A ならびに Peak B に該当する文書群の特徴付けを試みた．具体的には，まず全ての単語を属性としてベクトル化し，線形カーネルでモデルを構築した．そして各単語 wi について，その単語だけを含む仮想的な文を考え，このモデルを適用した際，Peak A または Peak B に該当する場合は Positive，そうでなければ Negative とした．

得られた推定値をその単語のスコア *score(wi)* とした．スコア順の Positive，Negative 単語を表 1，表 2 に示す．[10]ではこのスコアを単語の重要度としているが，今回はさらに単語 wi を含む文書頻度 df(wi) とスコアの積である *score(wi)*df(wi)* と df(wi)の対数とスコアの積である *score(wi)*log(df(wi))* の 3 つの指標を設定した．

[1] http://geta.ex.nii.ac.jp/geta.html

表 1：Peak A の特徴語(score)

	Positive word	Negative word
1	ワッシャ	人工関節
2	中枢	特発性
3	肉眼的	筋腹
4	後壁	骨幹
5	刺入	亜脱臼
6	一横指	術創
7	骨髄	外反
8	内反	全周
9	横指	菲薄化
10	肉芽	サポート

表 2：Peak B の特徴語(score)

	Positive word	Negative word
1	外反	伸展位
2	2 週	肉眼的
3	3 週	状態
4	4 週	切除後
5	1 週	関節面
6	交叉	前方脱臼
7	滑液	中枢
8	内転筋	人工関節
9	バイクリル	近位端
10	鋭的	硬膜外麻酔

3.2　長期在院の推定

検索エンジンは 3.1 と同様に構築し，長期在院（標準術後在院日数 25 日以上）のモデルで，術後在院日数が標準日数以上であれば Positive，標準日数より短ければ Negative とした．スコア順の Positive，Negative 単語を表 3 に示す．さらに，その推定指標についても 3.1 と同様に 3 つの指標で評価を行った．

表 3：長期在院の特徴語(score)

	Positive word	Negative word
1	外反	頭側
2	巨大	ワッシャ
3	緊張性	軽度
4	内反	CT
5	骨折線	前下方
6	癒痕	関節腔
7	鋭的	TM
8	人工関節	MM
9	骨幹	血腫
10	肢位	中枢

4 属性選択による推定性能

4.1 術後在院日数ピークの属性選択

Peak A，Peak B を特徴付ける各指標の上位 N 個，Negative 上位 N 個のモデルから 5 分割交差検定を利用し，その平均値の Accuracy，F-measure による推定性能の評価を行った．

Peak A の Accuracy の属性選択について図 2 に示す．辞書単語では N=2，3 の部分で高い性能に達しているが，全単語では N=200 である．score と score*log(df)の最大推定値は 0.8 以上に達した．全単語と比較しても score と score*log(df)は少ない N で性能が上がっている．score*df については全単語の方が少ない N で性能が上がっているが，最大推定値は全単語，辞書単語共に 0.8 程度である．Peak A の F-measure においては，両単語の 3 指標において 0.3 程度であり，性能が上がらない結果となった（図 3）．

Peak B の Accuracy の属性選択について図 4 に示す．辞書単語では Peak A と同様に，N=1~2 で 3 指標の推定値が約 0.7 となり属性選択の効果が認められる．比較して全単語の場合，score*df は N=200 で 0.664，score*log(df)は N=100 で 0.687 である．F-measure については Peak A と同様全単語，辞書単語共に 0.5 程であり属性選択の効果は見られない（図 5）．

4.2 長期在院の属性選択

術後在院日数を 25 日以上とする Positive 上位 N 個，Negative 上位 N 個のモデルから，4.1 と同様に 5 分割交差検定を利用し，Accuracy，F-measure による推定性能の評価を行った．Accuracy を図 6 に，F-measure を図 7 に示す．

全単語と辞書単語を比較すると，score が辞書単語の方が小さい N=30（全単語では N=200）でピークに達しており，score*log(df)は安定した傾向である．そして，score による推定性能では Accuracy，F-measure 共に N=20～30 がピークとなっているが，score*df と score*log(df)では N=10 未満で最適な性能とほぼ同じレベルの推定性能が得られている．

辞書単語の Accuracy について，all word で 0.477 であり，score は N=20 で 0.689 に達した．一方，score*df は N=5 で 0.658，score*log(df)は N=4 で 0.621 の値であり，少ない単語でピークに達している．後半のピークでは score*df は N=60 で 0.657，score* log(df)は N=50 で 0.700 の値を示す．このことから score*df と score* log(df) が score に比べて属性選択の効果があることが示される．

辞書単語の F-measure について,all word で 0.487，score は N=30 で 0.780，score*df は N=5 で 0.768，N=60 で 0.765，score* log(df)は N=4 で 0.703，N=50 で 0.802 を示し，Accuracy と同様に，前半と後半のピークが認められ，score*df と score* log(df)の属性選択の効果が認められる．

5 考察

人工股関節置換術症例の手術記録を対象に術後在院日数の 2 つのピークと長期在院の推定を目的として，SVM による特徴語の抽出と，属性選択で推定性能の評価を行った．

術後在院日数を目的として特徴語抽出の仕組みは構築できたが，単語のみでの扱いは困難であり，解釈について課題が残った．

属性選択では，Peak A，Peak B において F-measure では性能の効果が認められなかったが，Accuracy では全単語より辞書単語の方が良く，更に全単語と辞書単語の 3 指標については，Peak A の場合，辞書単語の score と score*log(df)の方が少ない単語で性能が良く，Peak B では，3 指標全て辞書単語の方が良かった．長期在院の場合では，全単語より辞書単語の方が，3 指標では score より score*df と score*log(df)の性能が良い顕著な結果となった．単語を絞ることと，単語スコアのみではなく，頻度も考慮することで属性選択の効果を上げることができた．

6 まとめと今後の課題

医療テキストからテキストマイニングと機械学習手法を利用して特徴語抽出と属性選択する仕組みを構築した．今回の結果から，テキストマイニング時の辞書の有意性と属性選択時の文書スコアと文書頻度を利用することで推定性能が上がることについては評価ができた．

しかし，抽出した単語のみでは解釈が難しい結果となっているため，充実した辞書構築が課題となっている．更に同義語や類似語処理も必要であり，臨床的な評価と解釈が可能な手法を目指す．

謝辞

本研究は，学術振興会科学研究費基盤(Ｂ)15H02778 及び国立研究開発法人日本医療研究開発機構（AMED）の【MID-NET を用いた医薬品等のベネフィット・リスク評価のための薬剤疫学研究等の実践的な分析手法及び教育に関する研究】の助成による．

参考文献

[１] J.A. Goldman., W.W. Chu., D.S. Parker., and R.M. Goldman.: Term Domain Distribution Analysis: a Data Mining Tool for Text Databases, Methods of information in medicine, vol.38 (2), pp.96-101, 1999.

[２] 竹村匡正，松井弘子，窪田英明，祐延良治，芦田信之：放射線読影レポートからの自然言語知識抽出による自動分類の試み, 医療情報学, vol.23, no.1, pp.95, Apr.2003.

[３] Y.C. Chang., H.J. Dai., J.C. Wu., J.M. Chen., R.T. Tsai., and W.L. Hsu.: TEMPTING system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries, Journal of Biomedical Informatics, vol.46, Supplement, pp.S54-S62, Dec.2013.

[４] T. Suzuki., H. Yokoi., S. Fujita., and K. Takabayashi., Automatic DPC code selection from electronic medical records: Text mining trial of discharge summary, Methods of Information in Medicine, vol.47 (6), pp.541-548, 2008.

[５] 三浦康秀, 荒牧英治, 大熊智子, 外池昌嗣, 杉原大悟, 増市博, 大江和彦:電子カルテからの副作用関係の自動抽出，言語処理学会第 16 回年次大会，pp.78-81, 2010.

[６] D.D. Fushman., W.W. Chapman., and C.J. McDonald,: What can Natural Language Processing do for Clinical Decision Support?, Journal of Biomedical Informatics, vol.42(5), pp.760-772, 2009.

[７] T. Yamashita., Y. Wakata., S. Hamai., Y. Nakashima., Y. Iwamoto., B. Flanagan., N. Nakashima., and S. Hirokawa.: Extraction of Key Factors from Operation Records by Support Vector Machine and Feature Selection, Indian Journal of Medical Informatics, vol.8, pp.70-71, No 2(2014) APAMI 2014 Special Issue.

[８] T. Yamashita., Y. Wakata., S. Hamai., Y. Nakashima., Y. Iwamoto., B. Flanagan., N. Nakashima., and S. Hirokawa.: Presumption Model for Postoperative Hospital Days from Operation Records, International Journal of Computer & Information Science, vol.16, pp.50-59, 2015.

[９] T. Sakai., and S. Hirokawa.: Feature Words that Classify Problem Sentence in Scientific Article,: Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, pp.360-367, 2012.

[１０] A. Donabedian., Evaluating the Quality of Medical Care,: The Milbank memorial fund quarterly, vol.44, pp.166-206, 1966.

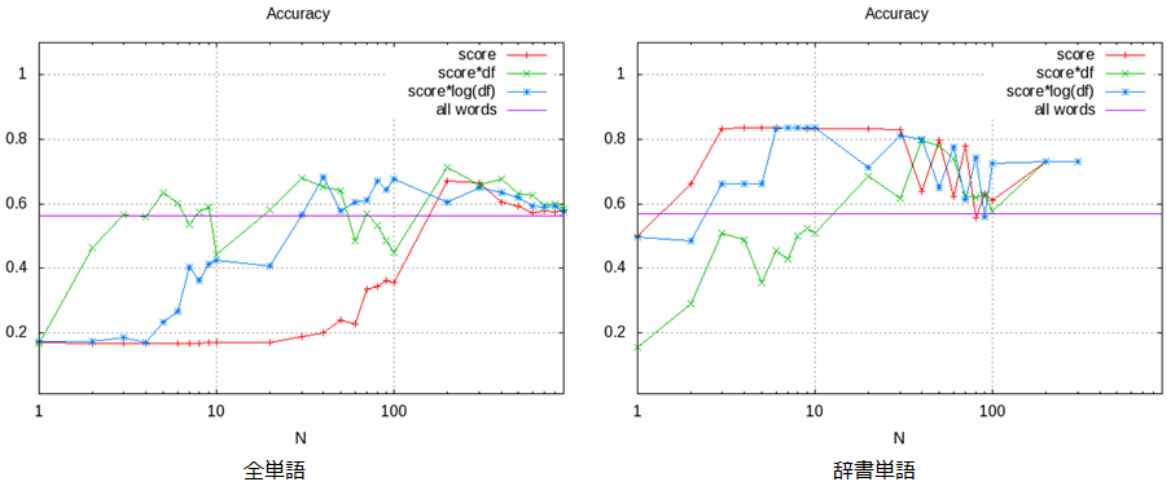

図 2：Peak A　Accuracy

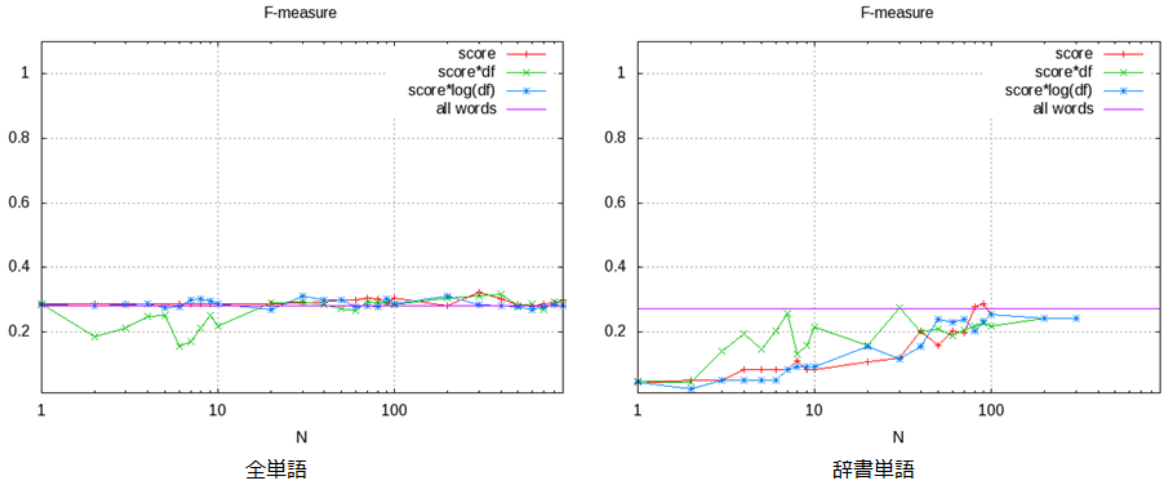

図 3：Peak A　F-measure

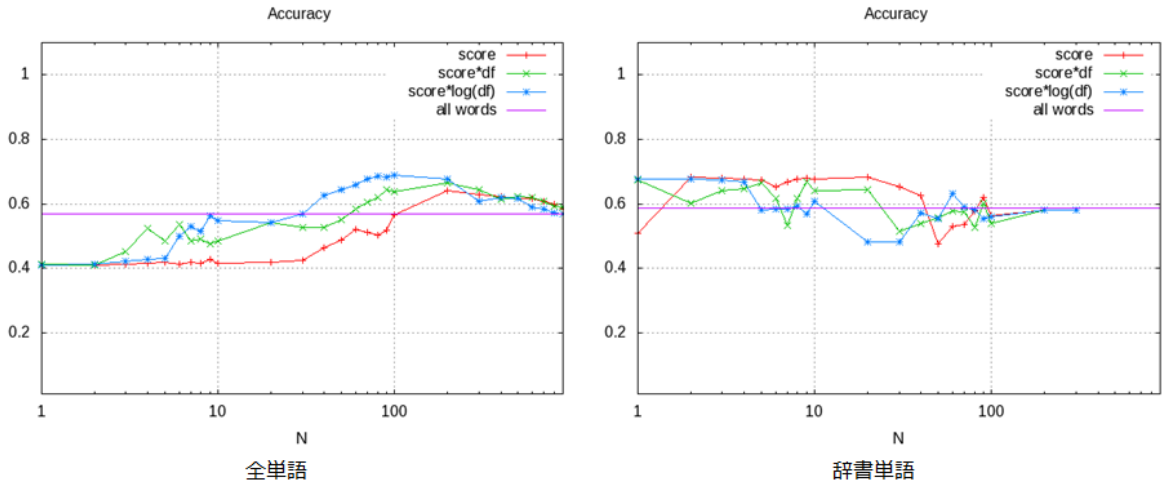

図 4：Peak B　Accuracy

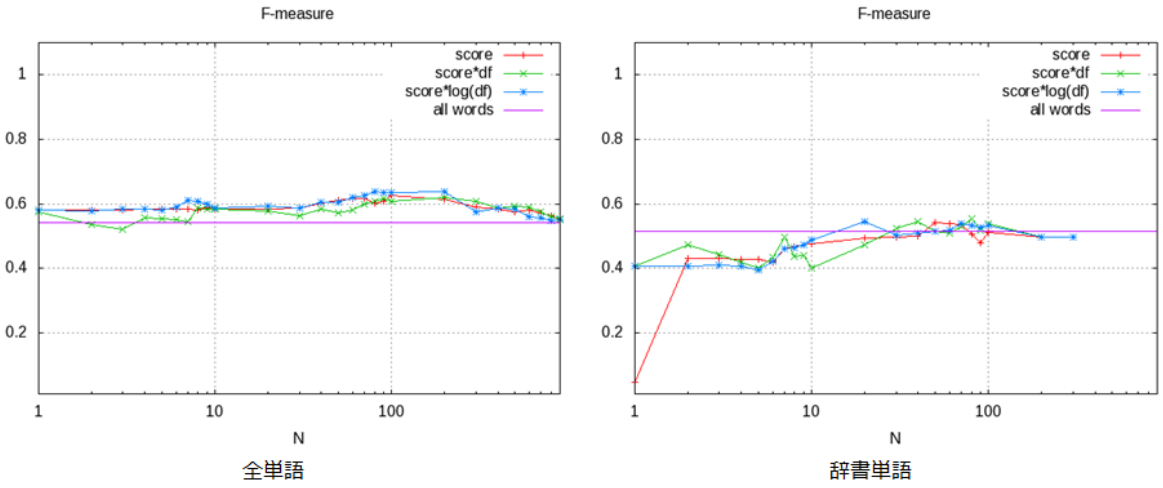

図 5：Peak B　F-measure

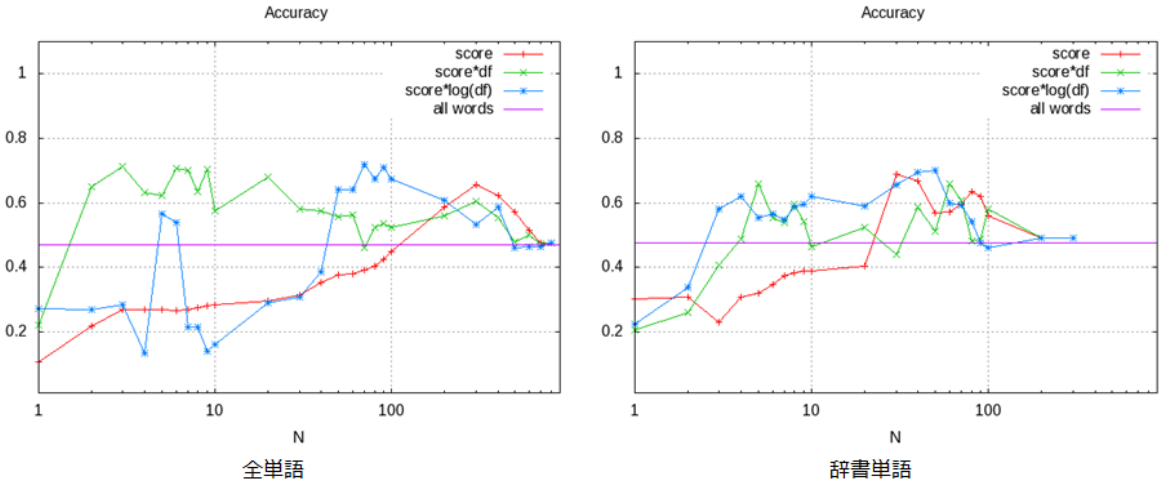

図 6：長期在院　Accuracy

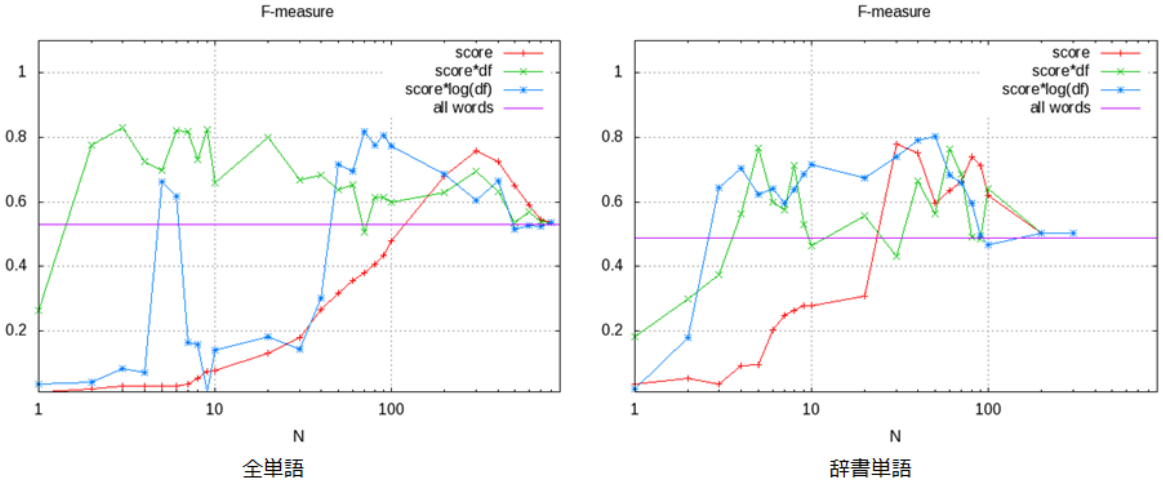

図 7：長期在院　F-measure