超大規模ウェブコーパスを用いた分布類似度計算

柴田知秀　黒橋禎夫
京都大学大学院情報学研究科

06t4071f　林華

1　はじめに

- 分布類似度とは

  「似た語は似た文脈で出現する」という分布仮説に基づいて計算される語の類似度である

- 従来の手法

  単語の類似度を測る方法として、人手で構築・整備されたシソーラスを利用する方法が考えられる

新しい手法

- 従来手法の欠点

  人手で構築・整備されたリソースは低カバレージであることや一貫性を保つことが難しいといった問題がある

- 現在の手法

  近年、大量のWeb コーパスが使えるようになってきており、NLP のいろいろなタスクにおいて、大規模コーパスを使うことにより精度が向上している

これまでの研究と本研究

Lin は6400 万語のテキスト、Curran は20 億語のテキスト、相澤は4000万日本語ウェブページ を利用していた

本論文では分布類似度計算においてもコーパスサイズを大きくすることにより精度が向上するかどうかを示すため、1 億ページから得られた250 億語からなるテキストを利用する

2　共起関係の抽出

ある単語w が関係r で他の単語w’と共起していることを(w, r, w’) の3 つ組で表す

rとは、“ガ, ヲ, ニ, カラ, ト, ヘ, マデ, ヨリ, ノ”

係り受け解析済みコーパスから共起関係を抽出し、その結果を集計することにより、すべての名詞を、共起要素を並べた共起ベクトルで表す

共起関係の一例

ニ:行く　ニ:かかる. . .　ノ:指示　ガ:青くなる. . .

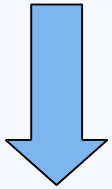
20,399,　9,284, . . .　945,　944,　. . .

「医者」の共起ベクトル

2.1　曖昧性のある係り受けの除外

解析誤りである係り受け関係から共起関係を抽出すると分布類似度計算においてノイズとなってしまう

ここで



ヒューリスティックなルールに基づき、係り受け先に曖昧性のある係り受け関係を捨て、信頼できる係り受け関係のみを用いた

2.2 語の単位

語w として、単名詞と複合名詞を考える。例えば以下の文では単名詞として「電話」、複合名詞として「携帯電話」を抽出する。

例： 携帯電話を購入した。

3　分布類似度計算

Curran は分布類似度計算をweight 関数とmeasure関数に分解した。weight 関数は頻度を別の値に変換する関数であり、measure関数は2 つのベクトル間の類似度を計算する関数である。

3.1 Weight

表 1: Weight 関数

FREQ	$f(w,c)$
MI	$\log \frac{P(w,c)}{P(w)P(c)}$
MI′	$\frac{f(w,c)}{f(w,c)+1} \cdot \frac{min(f(w,*),f(*,c))}{min(f(w,*),f(*,c))+1} \cdot \frac{P(w,c)}{P(w)P(c)}$
P_β	1 if MI $> \beta$: otherwise 0

相互情報量(MI) が広く用いられている。ただし、低頻度語の値が高くなるという問題があるので、3 番目の関数は補正をかけたものである。4 番目の関数はMI が閾値 よりも大きい場合に1、そうでない場合に0 とする関数である。

3.2 Measure

表 2: Measure 関数 ($(w,*) \equiv \{(c)|\exists(w,c)\}$)

COSINE	$\frac{\sum wgt(w_1,*)*wgt(w_2,*)}{\sqrt{\sum wgt(w_1,*)^2 * \sum wgt(w_2,*)^2}}$
LIN98	$\frac{\sum_{(w_1,*)\cap(w_2,*)} wgt(w_1,*)+wgt(w_2,*)}{\sum_{(w_1,*)} wgt(w_1,*)+\sum_{(w_2,*)} wgt(w_2,*)}$
JACCARD	$\frac{(w_1,*)\cap(w_2,*)}{(w_1,*)\cup(w_2,*)}$
SIMPSON	$\frac{(w_1,*)\cap(w_2,*)}{min((w_1,*),(w_2,*))}$
SIMPSON-JACCARD	$\frac{1}{2}(\text{JACCARD} + \text{SIMPSON})$

LIN98 はLin によって提案された$measure^{P_\beta}$ 関数である[4]。JACCARD、SIMPSON、SIMSON-JACCARDはweight 関数 P=1の場合に利用する。

4 実験
4.1 分布類似度計算

- 検索エンジンTSUBAKI2で検索対象となっている日本語1 億ページ、60 億文(1 兆語)
- 60 億文をuniq した16 億文(250 億語)
- 形態素解析器JUMAN、構文解析器KNP
- この処理は150CPU を用いて約1 週間
- 共起ベクトルは表3 に示す3 セットを構築し、それぞれ5 サイズのコーパスから作成した

表 3: コーパスサイズと名詞数・平均共起要素数の関係

	(i)		(ii)		(iii)	
語の単位	単名詞		単名詞		複合名詞	
曖昧性のある係り受け	あり		なし		なし	
コーパスサイズ (文)	名詞数	平均共起要素数	名詞数	平均共起要素数	名詞数	平均共起要素数
6.3M	69,356	9.67	39,325	7.35	57,774	4.57
25M	181,218	15.80	101,341	12.83	203,379	6.14
100M	456,858	21.04	247,292	18.03	639,702	7.18
400M	1,195,702	25.29	630,725	22.47	2,100,541	7.62
1.6G	3,197,004	28.08	1,698,155	25.07	7,311,191	7.38

4.2 評価セット

相澤の評価セットを用いて分布類似度の精度を評価した

タスクI:「A やB などのC」というパターンに着目し、{(A), (B)} を類義語として抽出。分類語彙表で(A)と異なるカテゴリに属する語のうち、Web での頻度が(B) と同程度の語(D) を求め、{(A), (D)}を非類義語とする

タスクII:「A やB などのC」というパターンに着目し、(C) が(A) や(B) の上位概念になっているかどうかを人手で判定

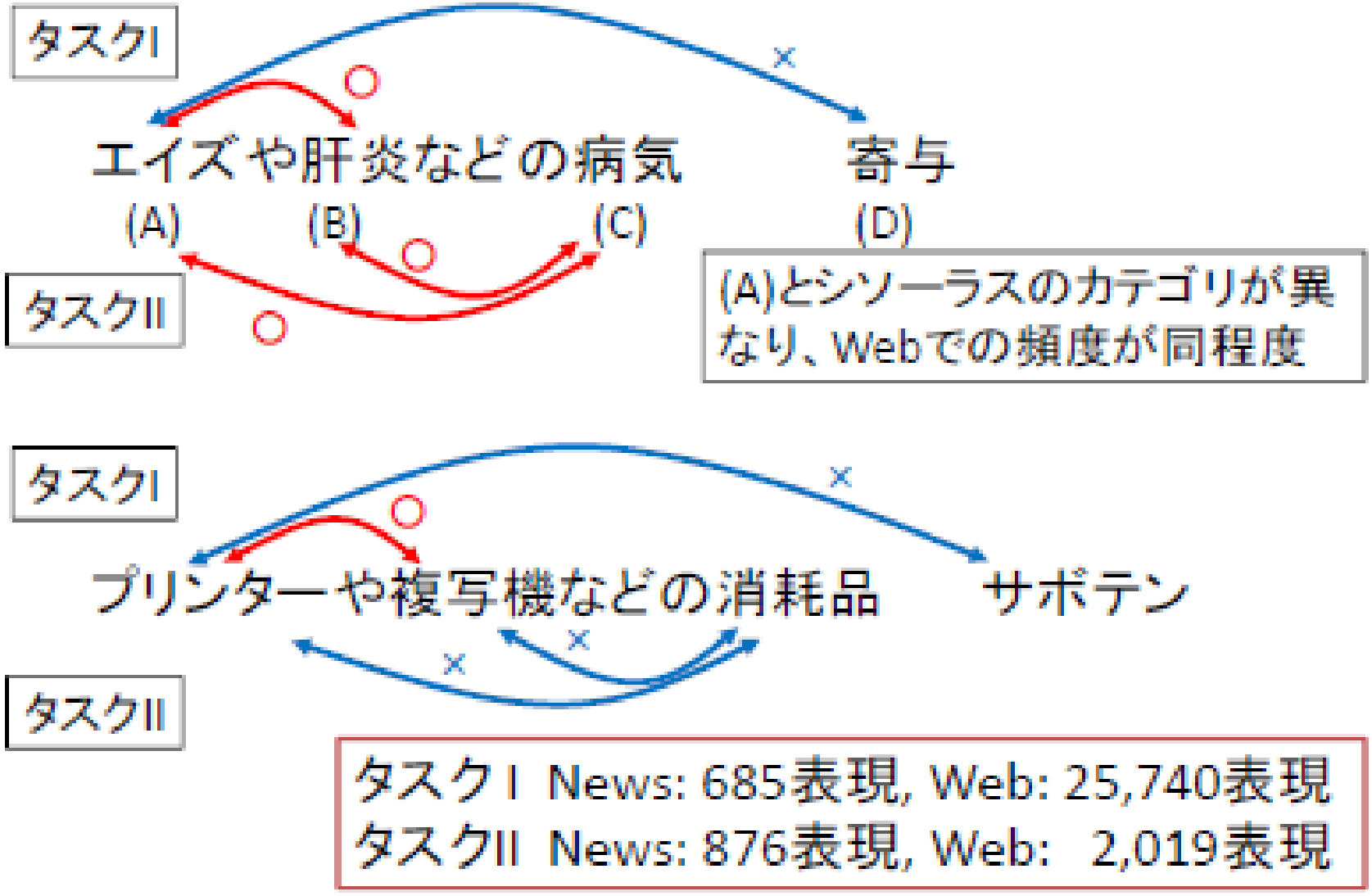


図 2: 評価セットの作成方法

4.3　評価

- 　予備実験の結果、weight 関数とmeasure 関数のあらゆる組み合わせのうち、表4 にあげた5 つの類似度尺度の精度が比較的よかった。
- これら5 つの類似度について、F 値を4 つの評価セットを用いて評価した。
- F 値の評価は、類似度閾値 を0.01 から0.40 まで0.01 ずつ変化させた時の最大値を求めて行なった

表 4: 5 つの類似度尺度の F 値 (括弧内の数字は最大の F 値の時の類似度閾値を示す)

類似度尺度	weight	measure	F 値			
			タスク I		タスク II	
			新聞	Web	新聞	Web
P-S	P_β	SIMPSON	0.985 (0.13)	0.973 (0.13)	0.807 (0.19)	0.876 (0.17)
P-J	P_β	JACCARD	0.981 (0.04)	0.945 (0.03)	0.743 (0.04)	0.805 (0.02)
P-SJ	P_β	SIMPSON-JACCARD	**0.988** (0.08)	**0.975** (0.08)	**0.817** (0.13)	**0.878** (0.11)
Lin98[4]	MI	LIN98	0.985 (0.08)	0.949 (0.06)	0.748 (0.10)	0.805 (0.06)
Lin02[5]	MI′	COSINE	0.984 (0.14)	0.955 (0.13)	0.758 (0.16)	0.818 (0.12)
相澤 [6]			0.982	0.971	0.752	0.862

4.4　各要素間の関係

- コーパスサイズ（文数）とF値の関係
- 曖昧性のある係り受けの除外の効果
- 語の単位（単名詞と複合名詞）
- 共起要素の格とF値の関係

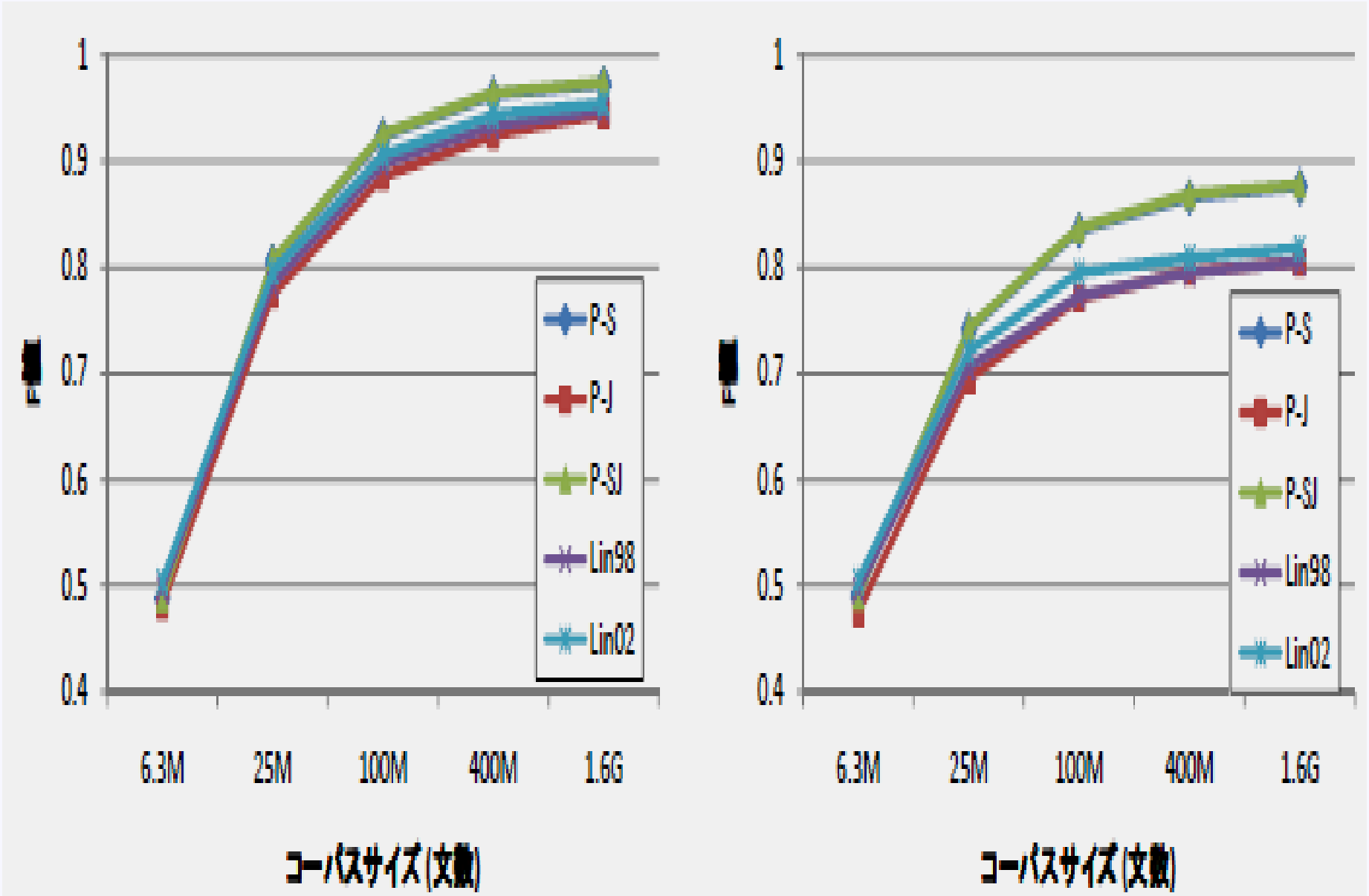


図 3: コーパスサイズと F 値の関係 (左: タスク I, 右: タスク II)

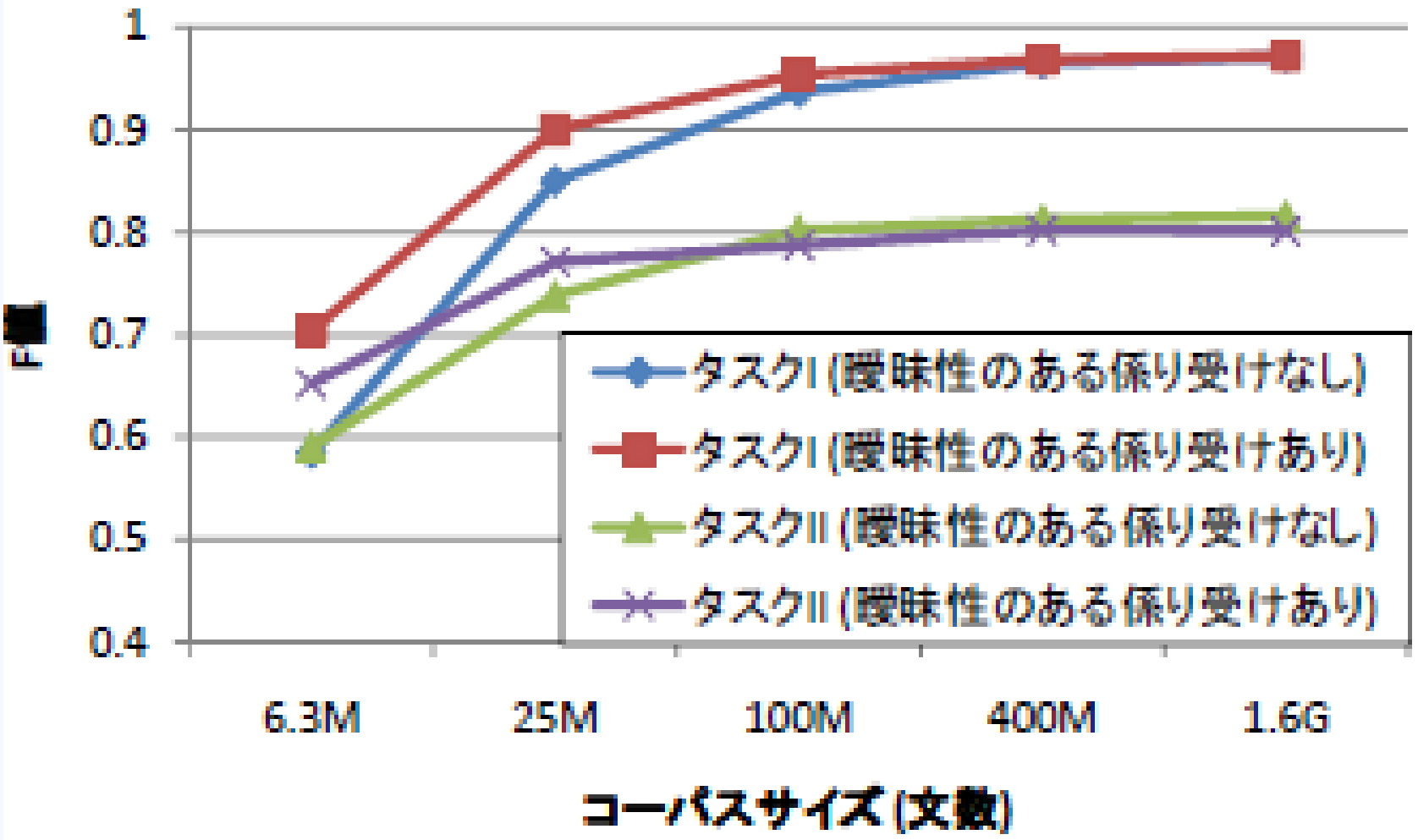


図 4: 曖昧性のある係り受けの除外の効果

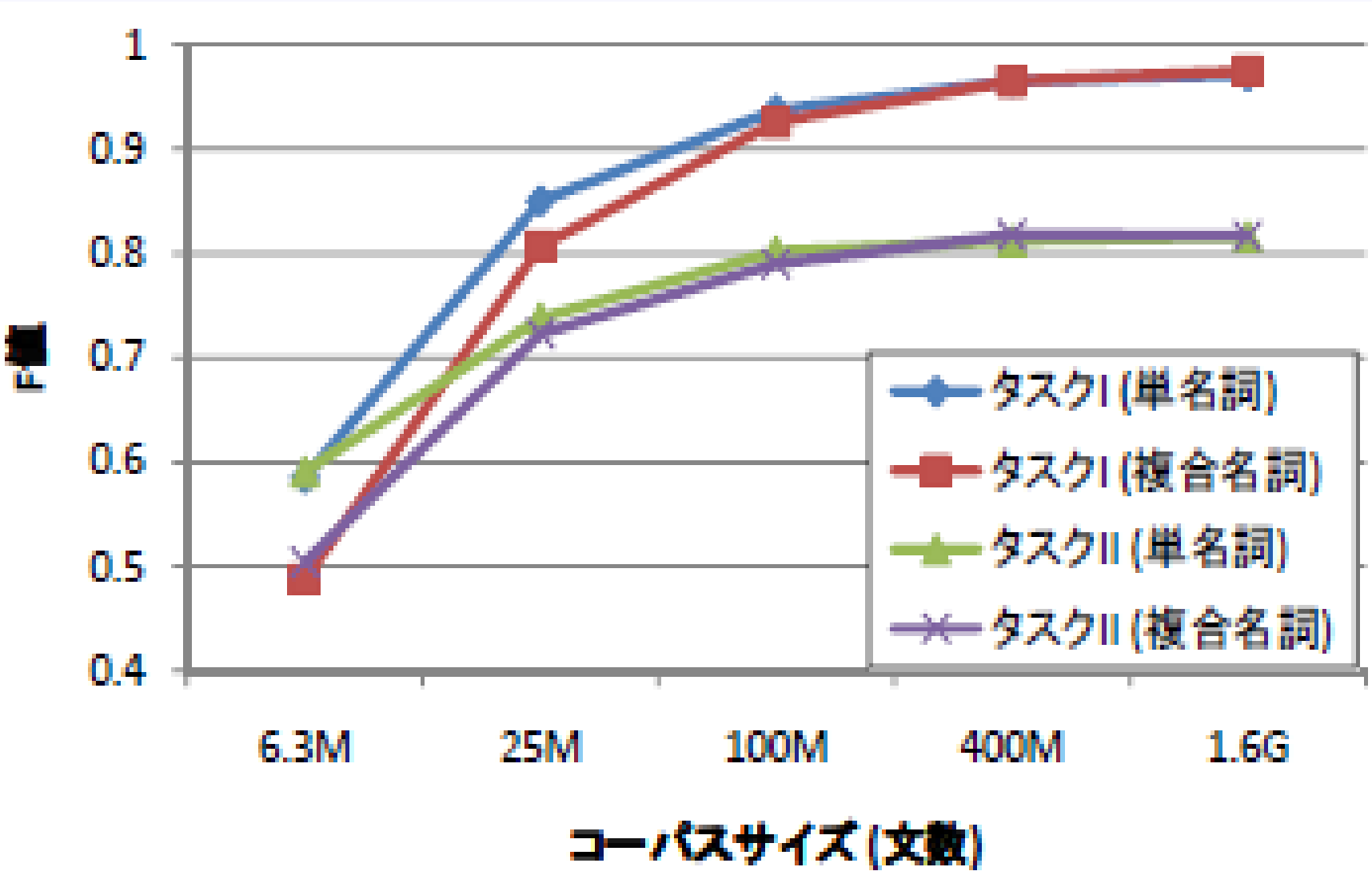


図 5: 語の単位 (単名詞と複合名詞)

格	F 値			
	タスク I		タスク II	
— ガ	0.971	(0.07)	0.870	(0.10)
— ヲ	0.970	(0.07)	0.873	(0.10)
— ニ	0.971	(0.07)	0.871	(0.11)
— カラ	0.971	(0.08)	0.873	(0.10)
— ト	0.971	(0.07)	0.872	(0.10)
— ヘ	0.971	(0.08)	0.872	(0.10)
— マデ	0.971	(0.08)	0.872	(0.10)
— ヨリ	0.971	(0.08)	0.872	(0.10)
— ノ	0.966	(0.08)	0.867	(0.11)
すべて	**0.971**	(0.07)	**0.872**	(0.10)
+ デ	0.971	(0.07)	0.870	(0.11)

表 5: 共起要素の格と F 値の関係

5　おわりに

本稿では超大規模ウェブコーパスを用いて分布類似度を計算する手法について述べた。コーパスサイズを大きくするにつれて精度が向上することを示した。

今後は分布類似度を用いて大規模格フレームを構築する予定である。