名詞句の内部構造を考慮したキーワードのスコア付け

2013年3月15日

京都大学

村脇 有吾　　黒橋 禎夫

言語処理学会第19回年次大会@名古屋

名詞句の内部構造を考慮した~~キーワード~~のスコア付け
キーフレーズ

2013年3月15日

京都大学
村脇 有吾　　黒橋 禎夫

言語処理学会第19回年次大会@名古屋

bag-of-wordsを超える何か

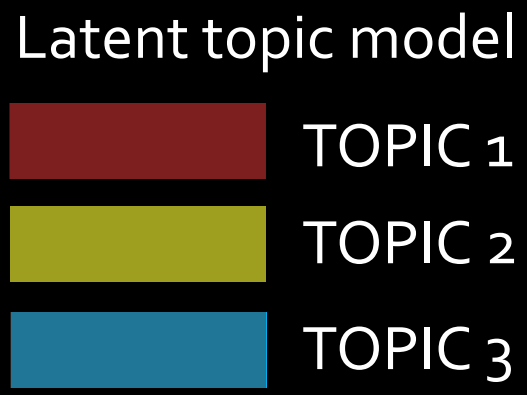
テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by …

bag-of-wordsを超える何か

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...



bag-of-wordsを超える何か

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by …

bag-of-wordsを超える何か

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

単語より長い
意味的なまとまり

キーフレーズ抽出すら
21世紀になってもbag-of-words

キーフレーズ抽出すら
21世紀になってもbag-of-words

キーフレーズ抽出すら
21世紀になってもbag-of-words

キーフレーズ抽出すら
21世紀になってもbag-of-words

教師なしキーフレーズ抽出

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

教師なしキーフレーズ抽出

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

システム出力

elliptic system
optimal control problems
distributed systems
....

教師なしキーフレーズ抽出

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

完全一致による評価

	システム出力	正解リスト
×	elliptic system	elliptic equations
○	optimal control problems	optimal control problems
○	distributed systems	distributed systems
	....	....

スコアに基づくキーフレーズ抽出

キーフレーズ候補	スコア
new optimal control problems	35.8
accurate computational discretization schemes	35.7
quadratic minimized function	26.9
computational finite-element schemes	25.2
optimal control	18.0
conjugate conditions	17.9
elliptic system	17.2
…	

スコアに基づくキーフレーズ抽出

スコア上位N語を出力

キーフレーズ候補	スコア
new optimal control problems	35.8
accurate computational discretization schemes	35.7
quadratic minimized function	26.9
computational finite-element schemes	25.2
optimal control	18.0
conjugate conditions	17.9
elliptic system	17.2
…	

実験設定

- Inspecデータセット (Hulth 2003)
  - 英語500論文
  - 表題と要旨 (平均134語)
  - キーフレーズ
    - シソーラスによる統制語
    - 非統制語 ← こちらを正解データとする
- スコア上位N語のNを変動させて recall-precision曲線を描く

実験設定

- Inspecデータセット (Hulth 2003)
  - 英語500論文
  - 表題と要旨 (平均134語)
  - キーフレーズ
    - シソーラスによる統制語
    - 非統制語 ← こちらを正解データとする
- スコア上位N語のNを変動させて recall-precision曲線を描く
- 手法間の相対評価と割りきってください!

準備: 最長名詞列と部分名詞列

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

準備: 最長名詞列と部分名詞列

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions
new optimal control problems are considered for distributed systems described by ...

1. 最長名詞列: 自動解析結果から規則により抽出される、最も長い名詞・形容詞の列

準備: 最長名詞列と部分名詞列

テキスト

computational finite-element schemes for optimal control of an elliptic system with conjugation conditions

hew optimal control problems are considered for distributed systems described by ...

1. 最長名詞列: 自動解析結果から規則により抽出される、最も長い名詞・形容詞の列
2. 部分名詞列: 最長名詞列に含まれるキーフレーズ候補
   - 不自然な候補も (e.g. control problems)

tf-idf法 (Hasan+ 2010)

$$\mathrm{tfidf}_{doc}(\mathbf{w}) = \mathrm{unit}_{doc}(\mathbf{w}) \times \mathrm{term}_{doc}(\mathbf{w})$$

$$\mathrm{unit}_{doc}(\mathbf{w}) = 1 \text{ if } \mathbf{w} \in doc,\ 0 \text{ otherwise}$$

$$\mathrm{term}_{doc}(\mathbf{w}) = \sum_{i} \mathrm{tf}_{doc}(w_i) \times \log(D / D_{w_i})$$

- キーフレーズ候補**w**のスコア: 単語tf-idfの総和（term）

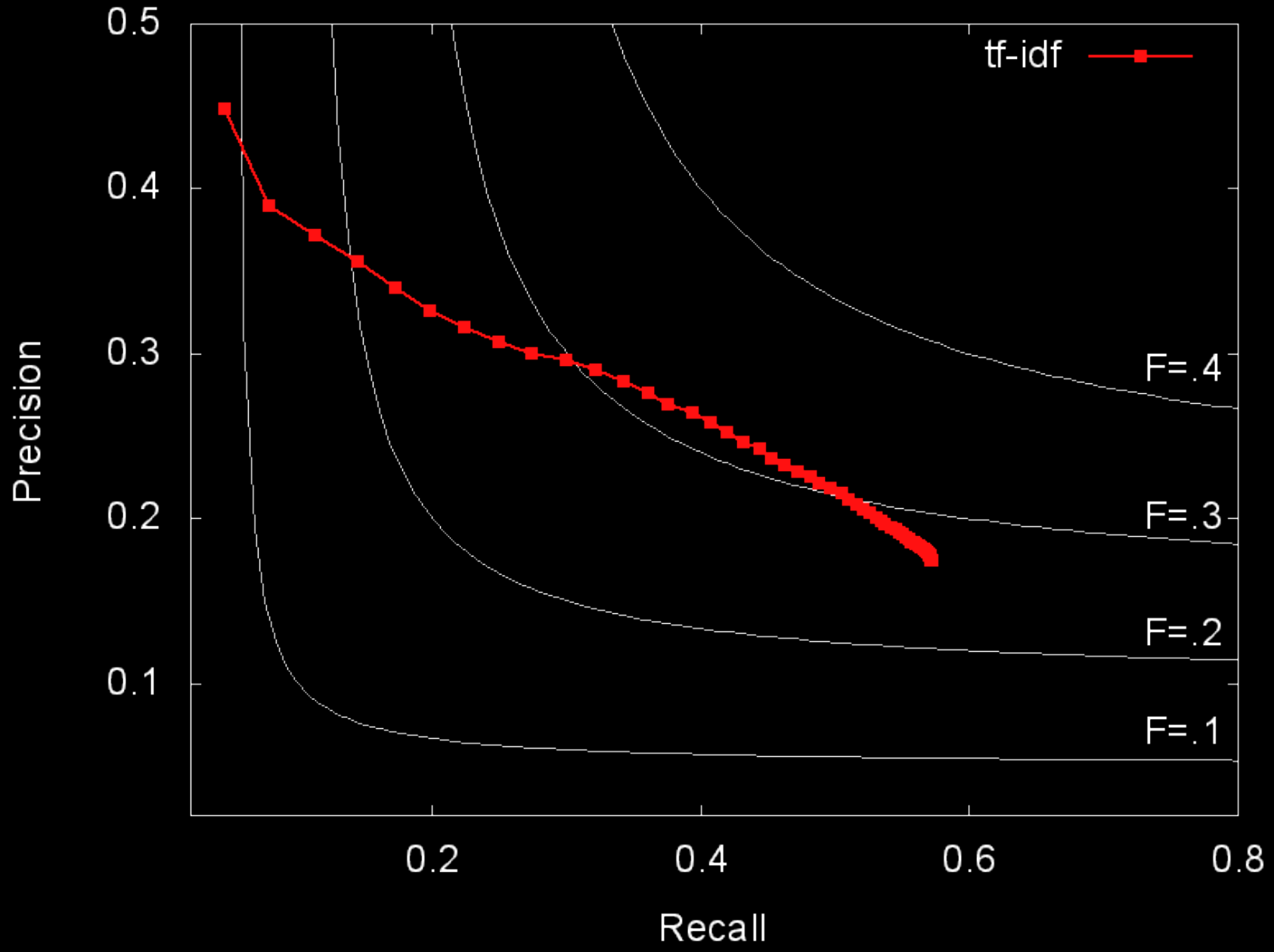
tf-idf法 (Hasan+ 2010)

$$\mathrm{tfidf}_{doc}(\mathbf{w}) = \mathrm{unit}_{doc}(\mathbf{w}) \times \mathrm{term}_{doc}(\mathbf{w})$$

$$\mathrm{unit}_{doc}(\mathbf{w}) = 1 \text{ if } \mathbf{w} \in doc,\ 0 \text{ otherwise}$$

$$\mathrm{term}_{doc}(\mathbf{w}) = \sum_i \mathrm{tf}_{doc}(w_i) \times \log(D / D_{w_i})$$

- キーフレーズ候補**w**のスコア: 単語tf-idfの総和（term）
- キーフレーズ候補: 最長名詞列のみ (unit)


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
F=.4
F=.3
F=.2
F=.1

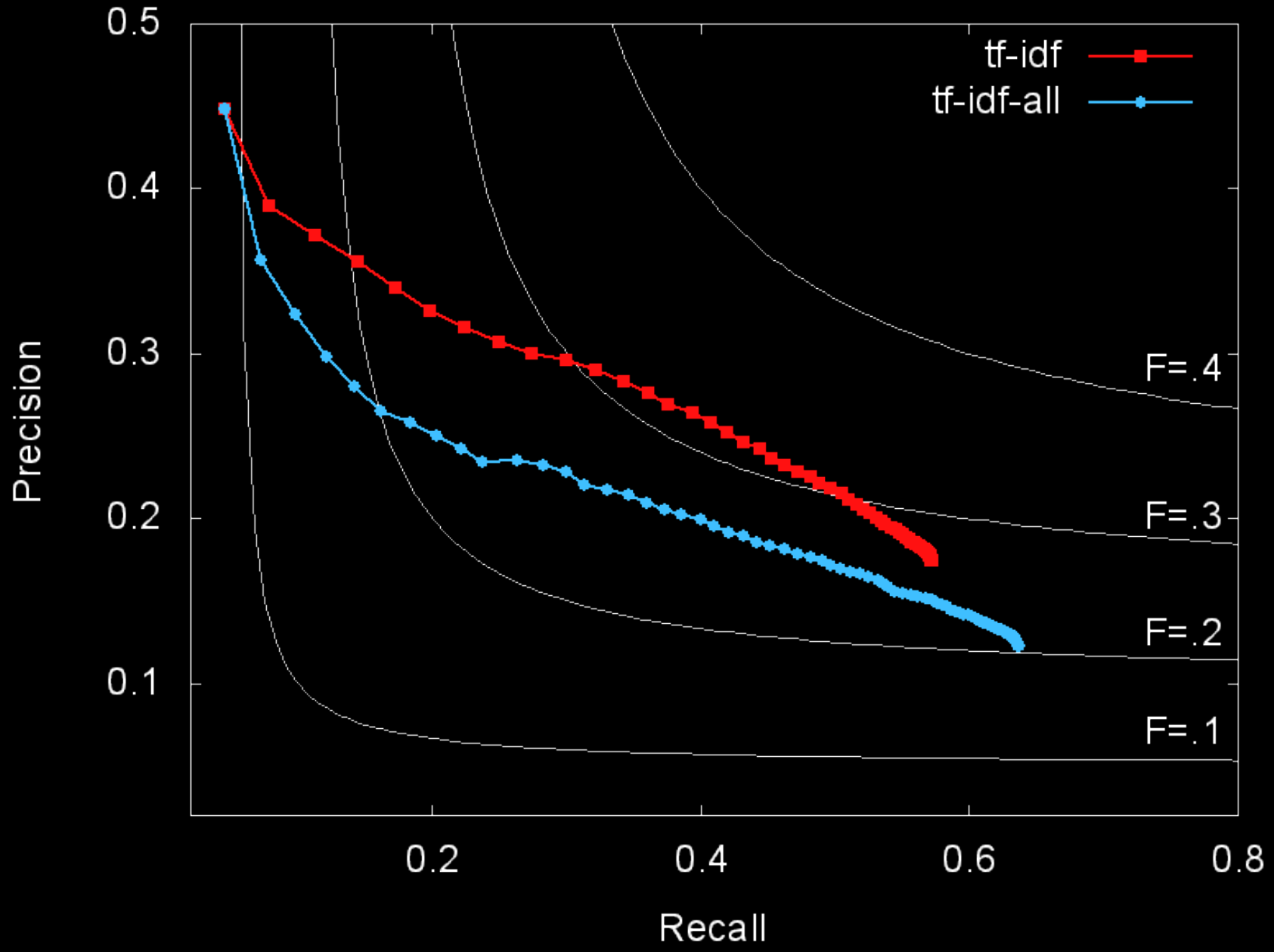
tf-idf-all

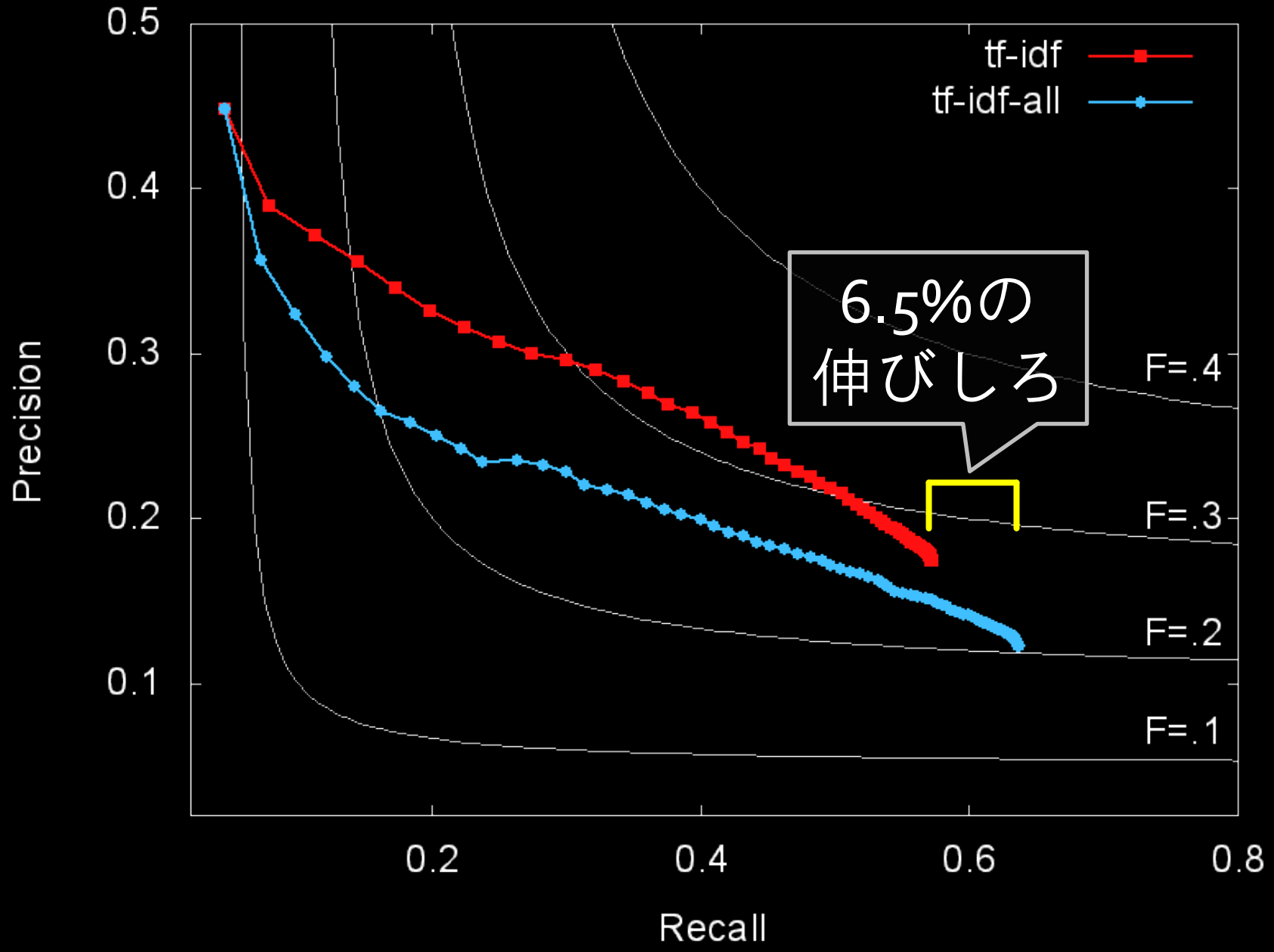
$$\mathrm{tfidfall}_{doc}(\mathbf{w}) = \mathrm{unit}_{doc}(\mathbf{w}) \times \mathrm{term}_{doc}(\mathbf{w})$$

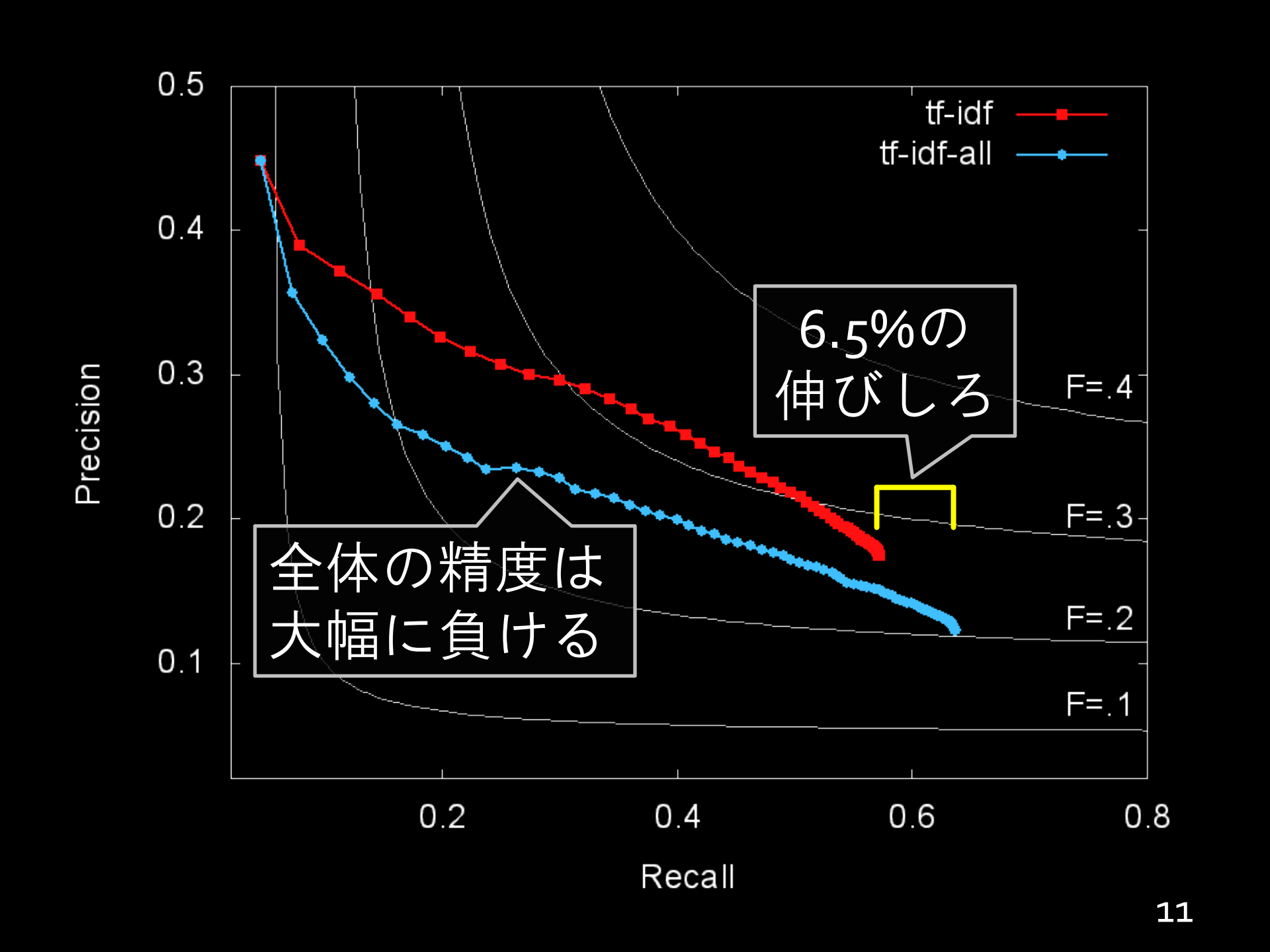
$$\mathrm{unit}_{doc}(\mathbf{w}) = 1 \text{ if } \mathbf{w} \in doc,\ 0 \text{ otherwise}$$

$$\mathrm{term}_{doc}(\mathbf{w}) = \sum_i \mathrm{tf}_{doc}(w_i) \times \log(D / D_{w_i})$$

- キーフレーズ候補**w**のスコア: 単語tf-idfの総和 (term)
- キーフレーズ候補: 部分名詞列も候補にする
  - ☺ recall上界の上昇
    (e.g. new optimal control problems → optimal control problems)
  - ☹ 不自然な候補も (e.g. control problems)


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
F=.4
F=.3
F=.2
F=.1


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
6.5%の
伸びしろ
F=.4
F=.3
F=.2
F=.1


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
F=.4
F=.3
F=.2
F=.1
6.5%の
伸びしろ
全体の精度は
大幅に負ける

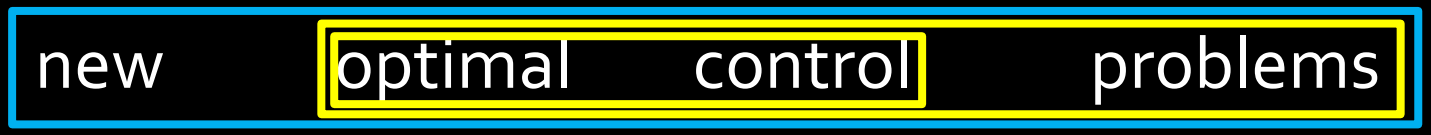
tf-idf-allは
意味的まとまりを考慮しない

	キーフレーズ候補	スコア
	new optimal control problems	25.7
	optimal control problems	23.5
☹	new optimal control	22.2
	optimal control	20.0
☹	control problems	15.5
	control	12.0
	problems	3.5
	…	

名詞句解析 (Murawaki+ 2012)

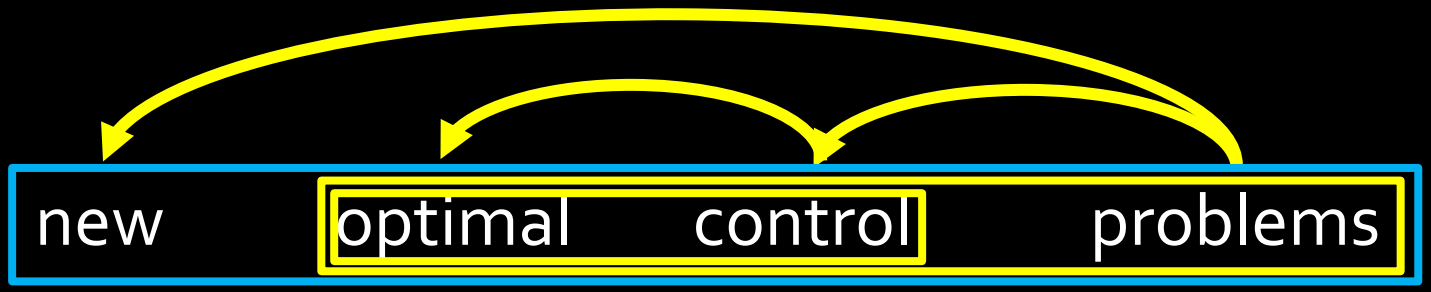
new	optimal	control	problems

名詞句解析 (Murawaki+ 2012)



- 意味的にまとまった部分名詞句を同定する

名詞句解析 (Murawaki+ 2012)



- 意味的にまとまった部分名詞句を同定する
- 主辞後置性を仮定すれば係り受けの問題
- 動的計画法で解ける
- 係り受け精度96.3% (Penn Treebankテスト)

tf-term: 名詞句解析に基づく部分名詞列への頻度配分

new optimal control problems

頻度 1.00

tf-term: 名詞句解析に基づく部分名詞列への頻度配分

new optimal control problems

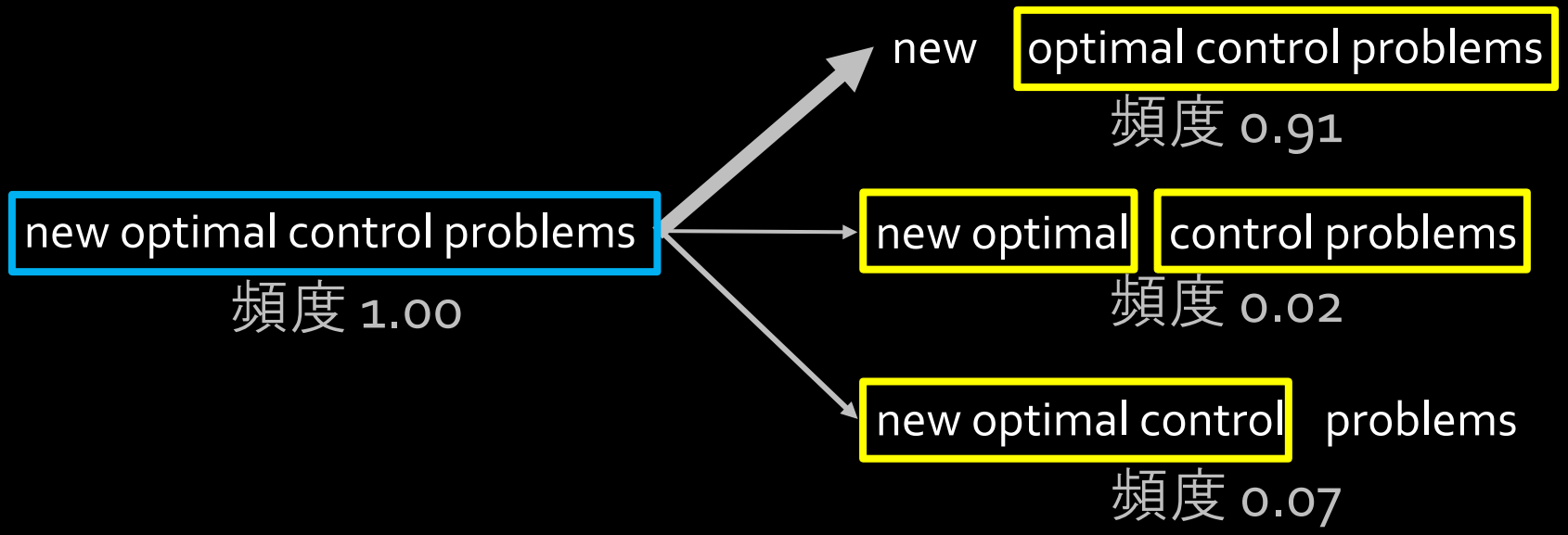
new optimal control problems

頻度 1.00

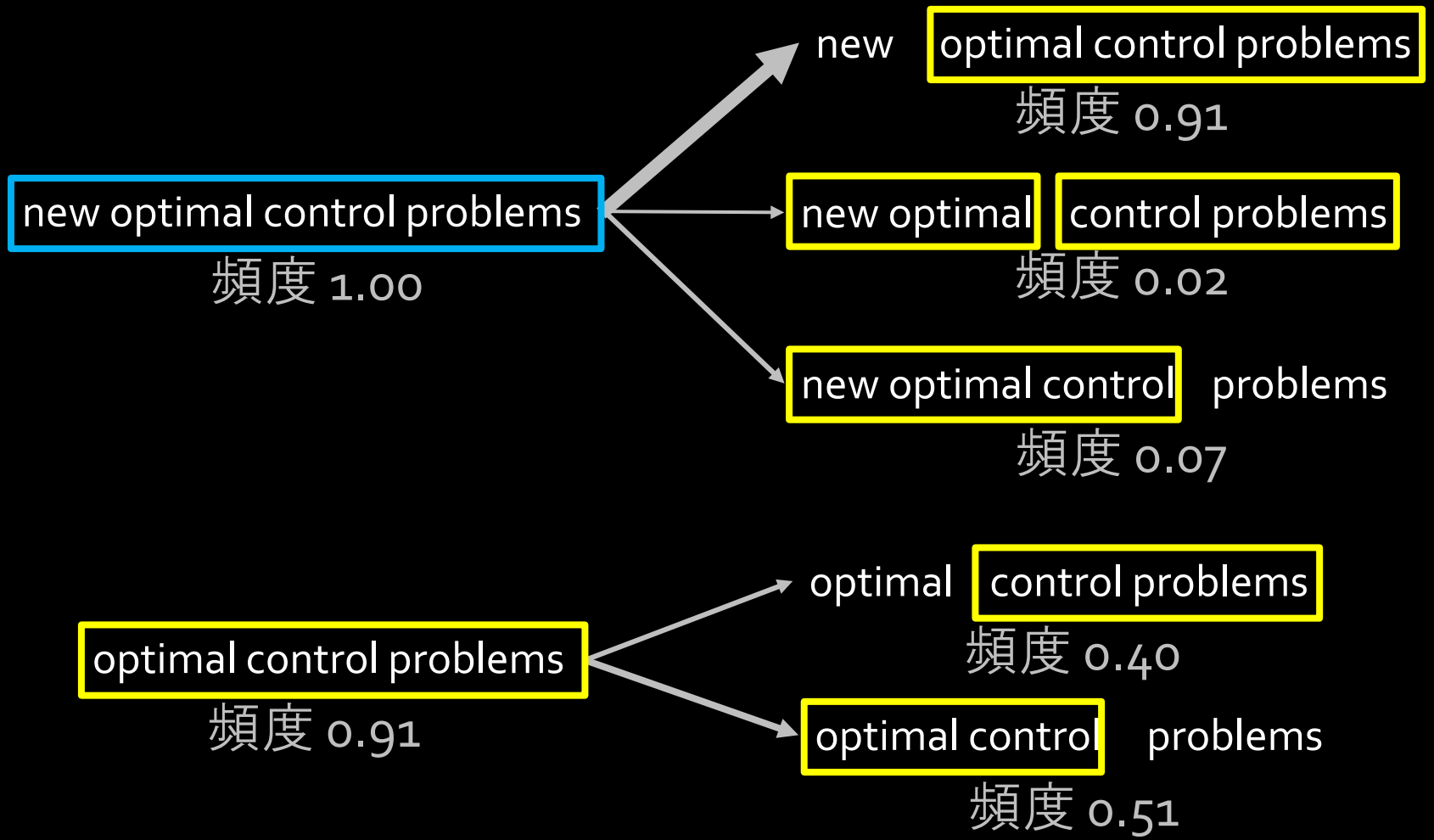
new optimal control problems

new optimal control problems

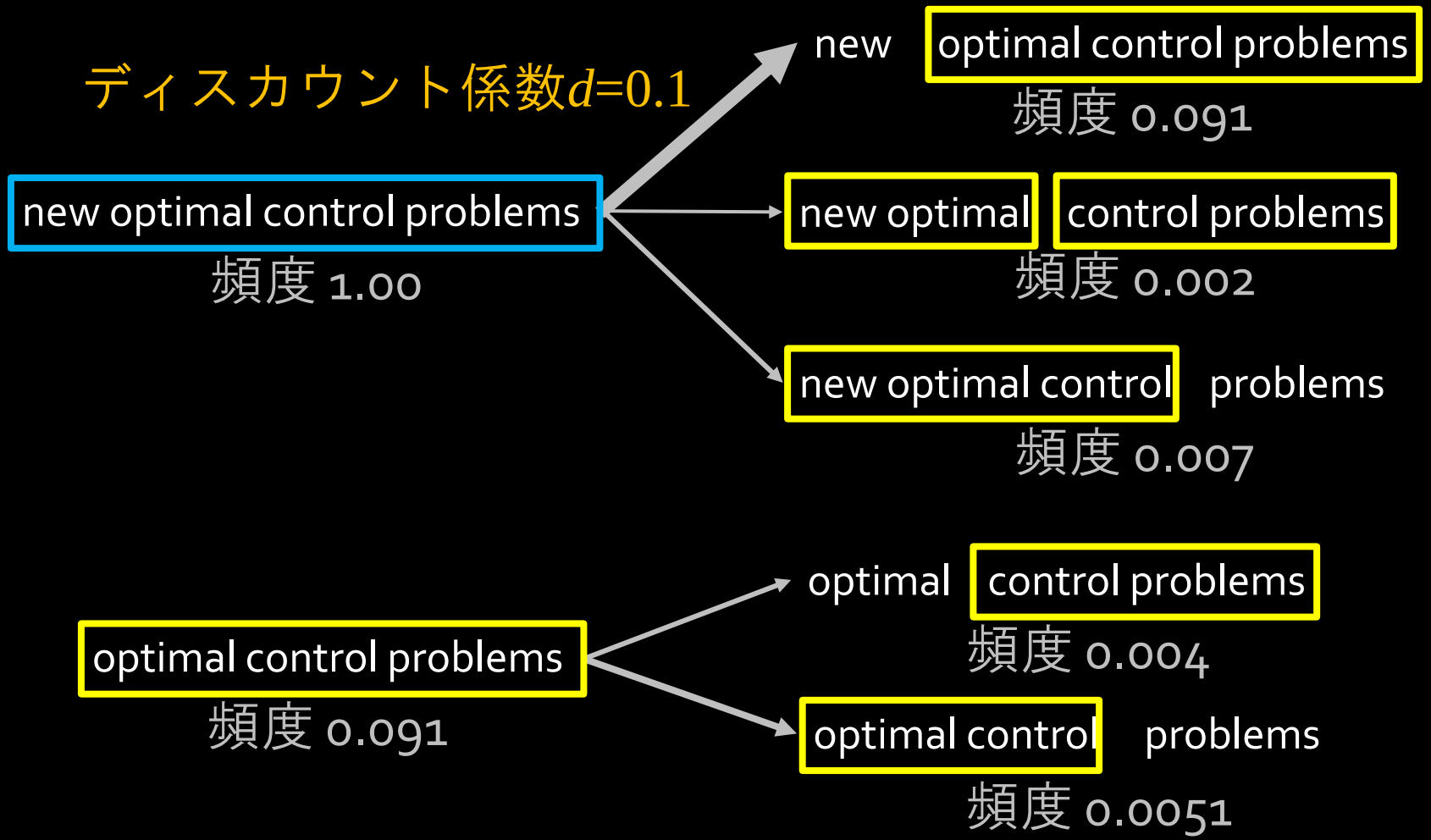
tf-term: 名詞句解析に基づく部分名詞列への頻度配分



tf-term: 名詞句解析に基づく部分名詞列への頻度配分



tf-term: 名詞句解析に基づく部分名詞列への頻度配分



名詞列のidf相当尺度?

スコア(**w**) = tfterm(**w**) × idf相当尺度(**w**)

名詞列のidf相当尺度?

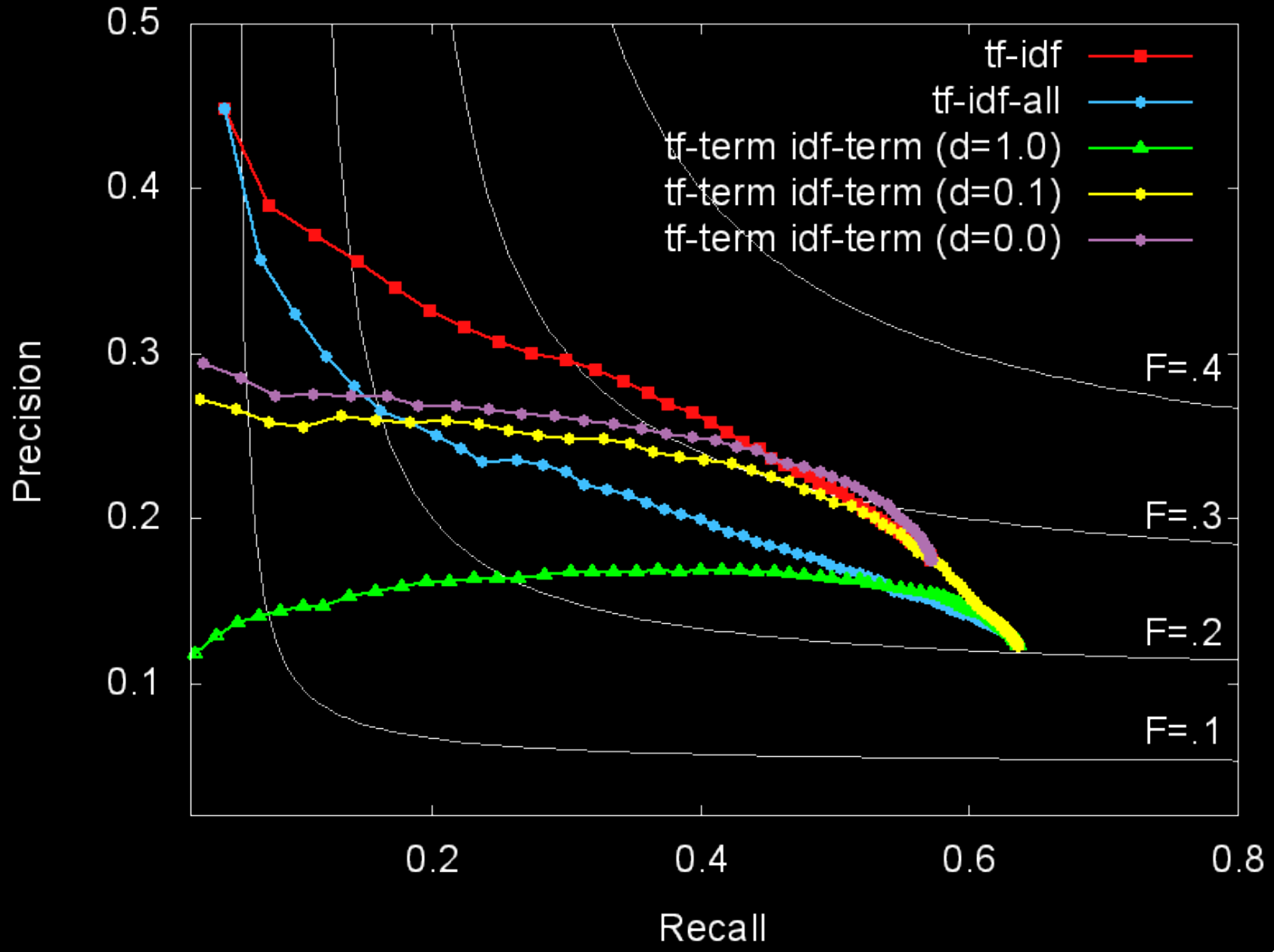
スコア(**w**) = tfterm(**w**) × idf相当尺度(**w**)

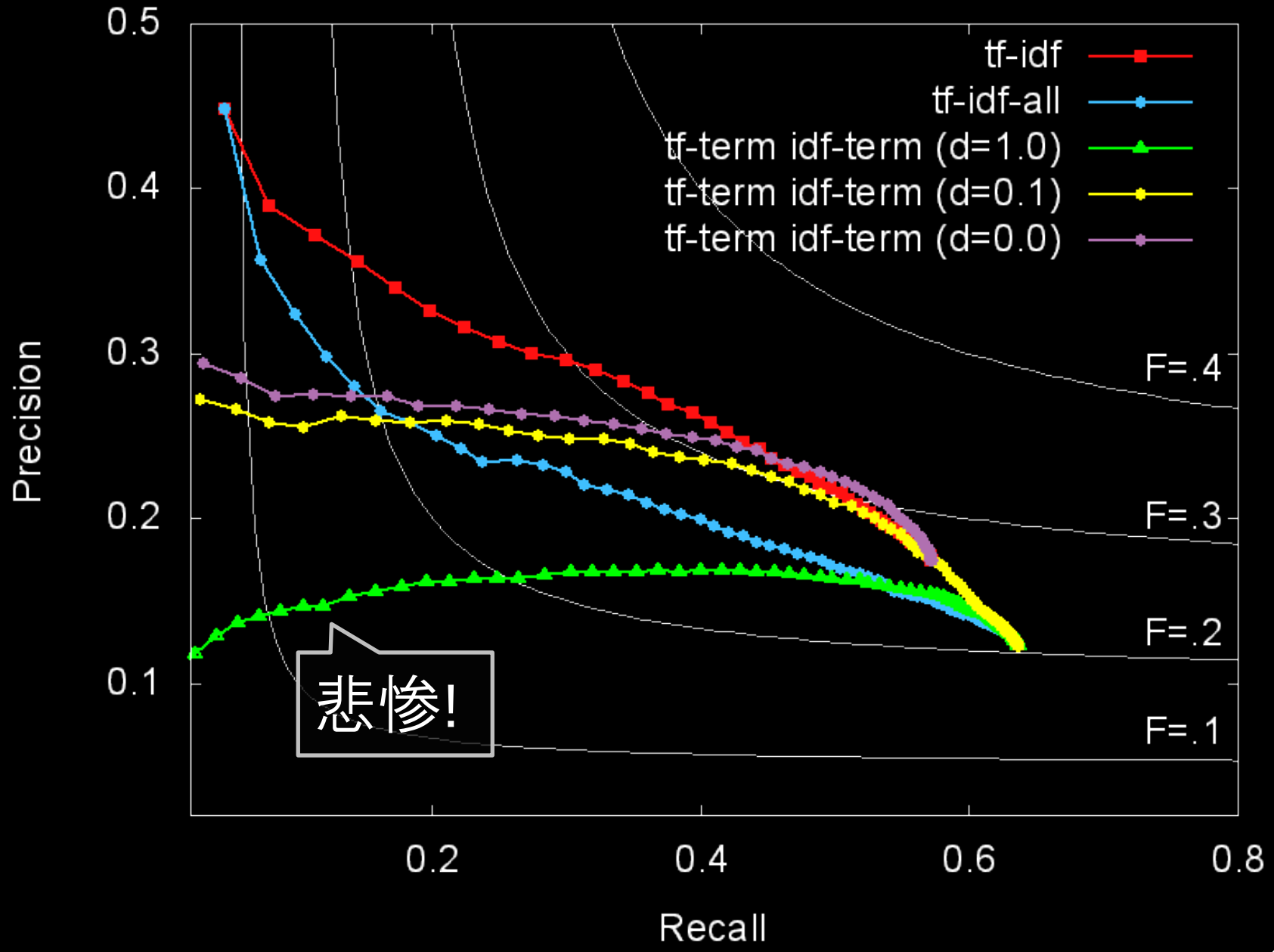
1. 名詞列のidfを直接算出 (tf-term idf-term)

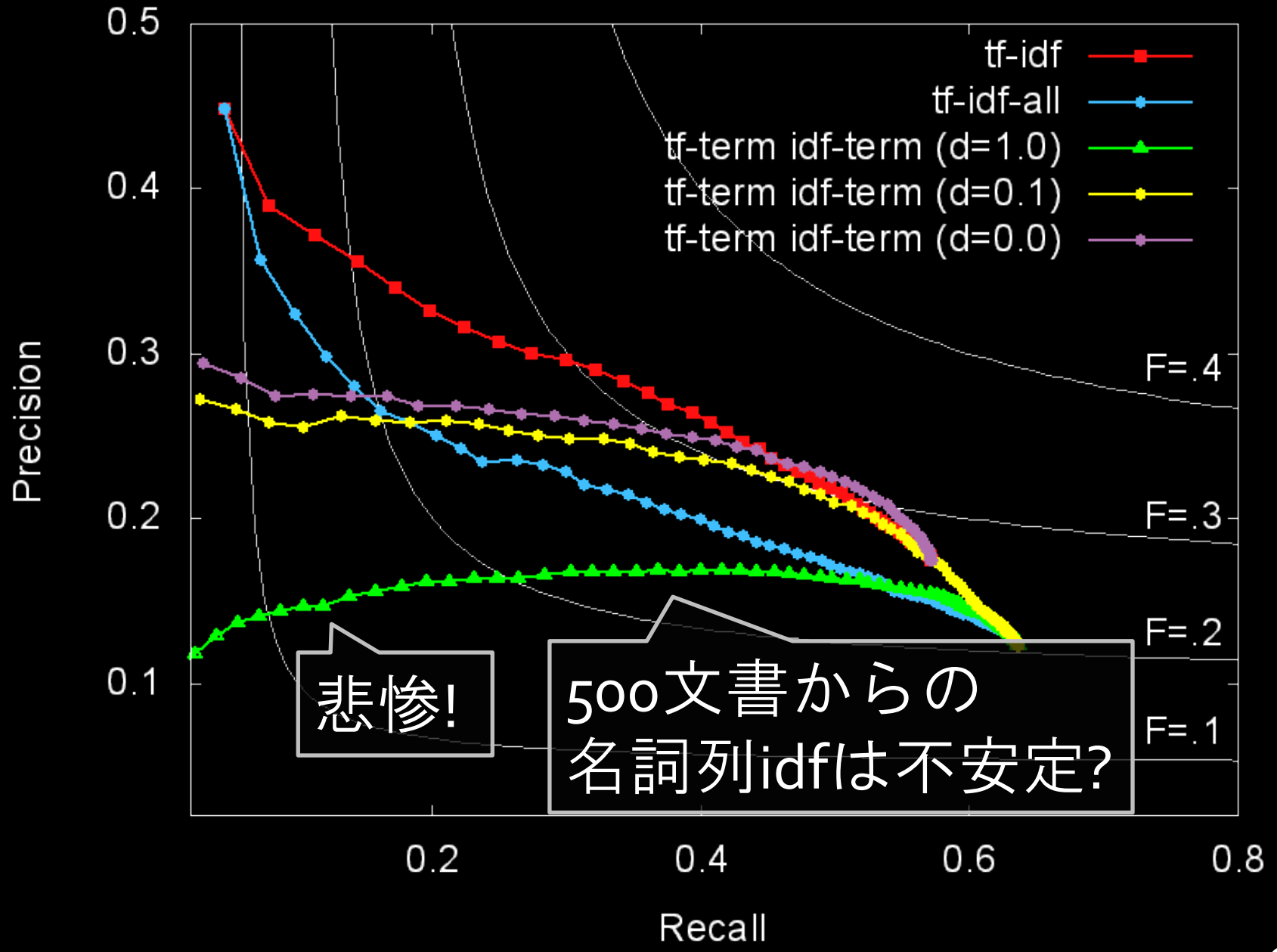
$$\mathrm{idfterm}(\mathbf{w}) = \log(D / D_{\mathbf{w}})$$

$$D_{\mathbf{w}} = \left|\left\{doc \,\middle|\, \mathrm{tfterm}_{doc}(\mathbf{w}) > 0\right\}\right|$$

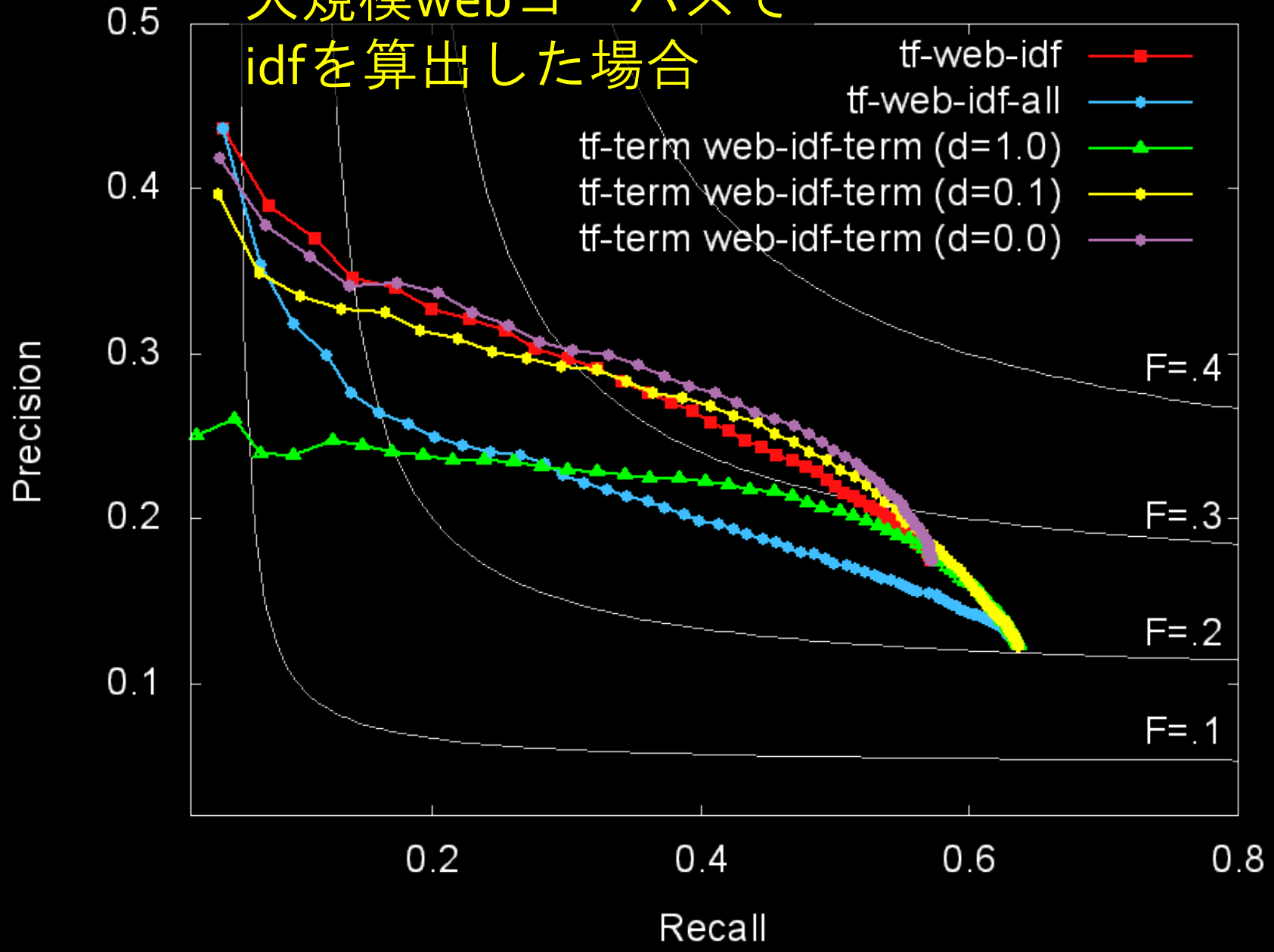
2. …


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-term (d=1.0)
tf-term idf-term (d=0.1)
tf-term idf-term (d=0.0)
F=.4
F=.3
F=.2
F=.1

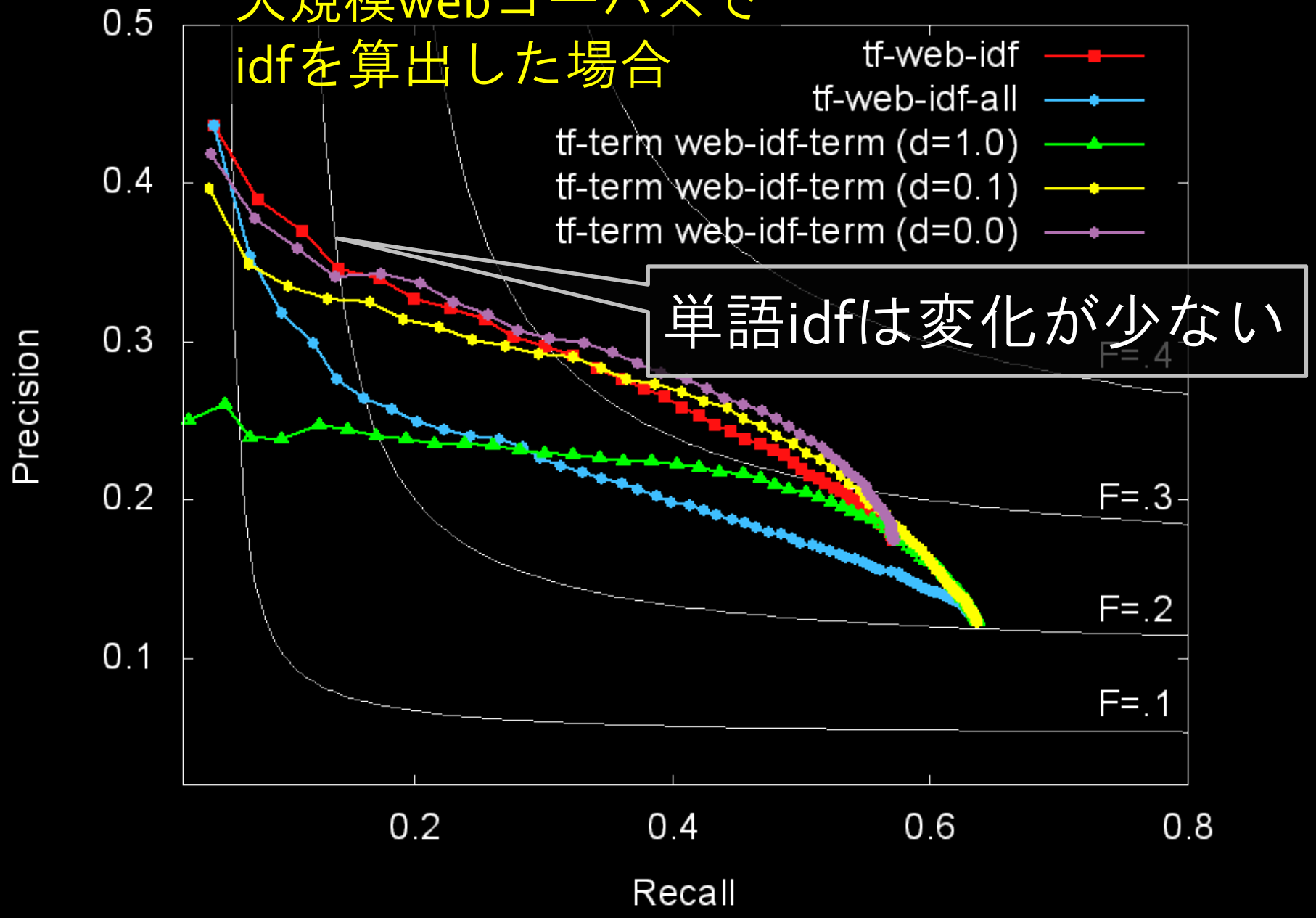

0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-term (d=1.0)
tf-term idf-term (d=0.1)
tf-term idf-term (d=0.0)
F=.4
F=.3
F=.2
F=.1
悲惨!

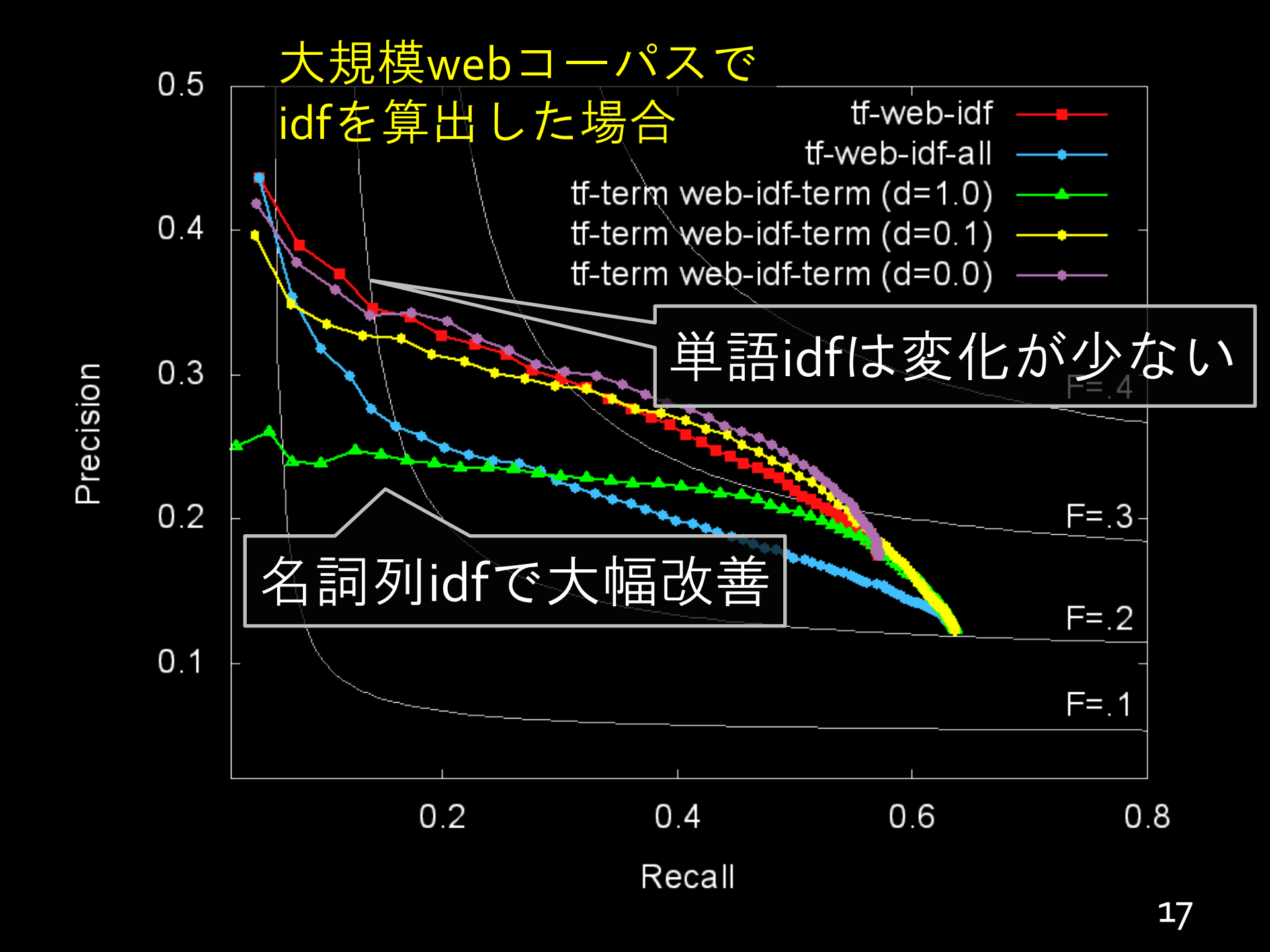

0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-term (d=1.0)
tf-term idf-term (d=0.1)
tf-term idf-term (d=0.0)
F=.4
F=.3
F=.2
F=.1
悲惨!
500文書からの
名詞列idfは不安定?

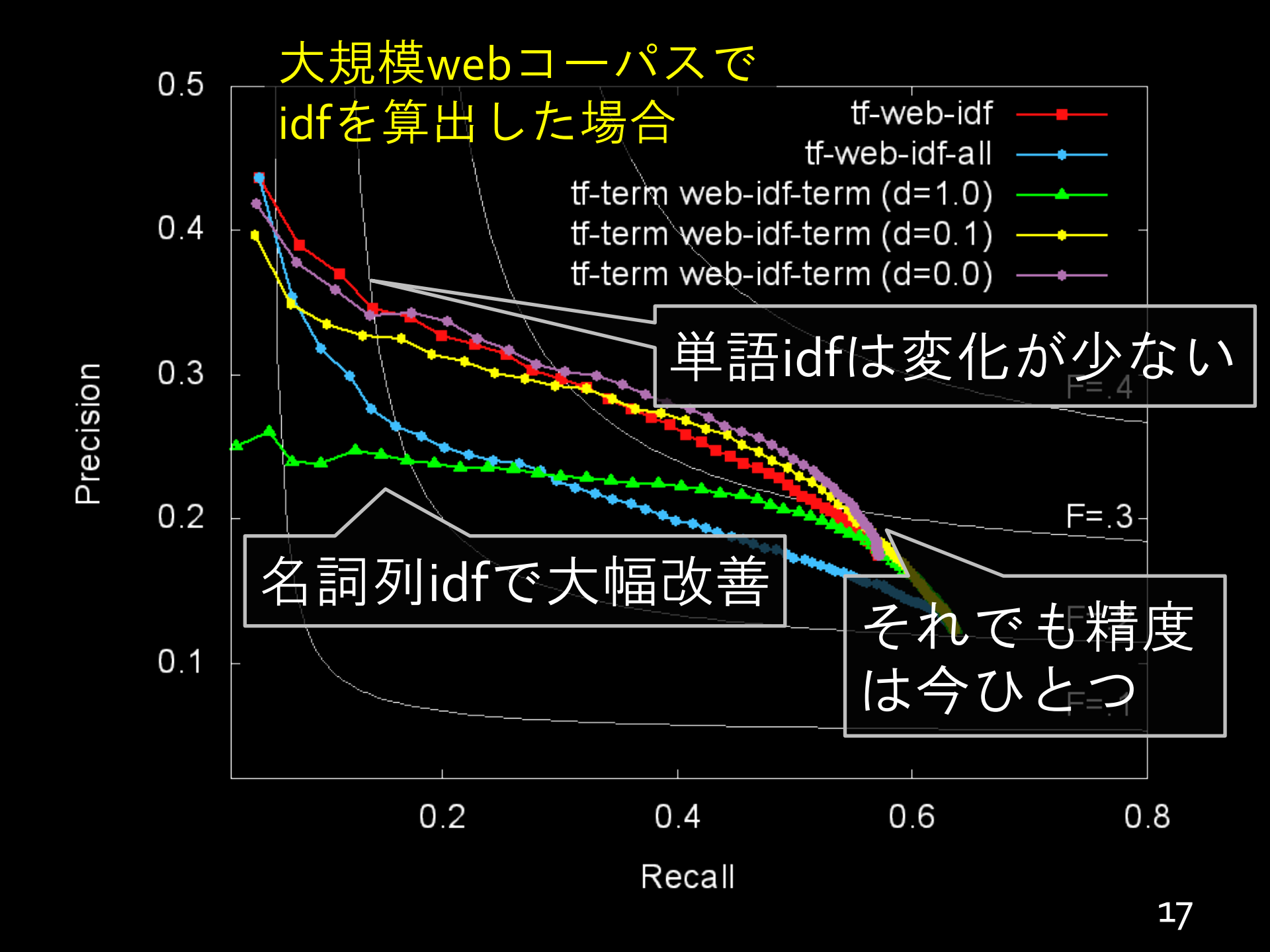
大規模webコーパスで idfを算出した場合



大規模webコーパスで idfを算出した場合




大規模webコーパスで
idfを算出した場合
tf-web-idf
tf-web-idf-all
tf-term web-idf-term (d=1.0)
tf-term web-idf-term (d=0.1)
tf-term web-idf-term (d=0.0)
単語idfは変化が少ない
名詞列idfで大幅改善
F=.4
F=.3
F=.2
F=.1
Precision
Recall
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8


大規模webコーパスで
idfを算出した場合
tf-web-idf
tf-web-idf-all
tf-term web-idf-term (d=1.0)
tf-term web-idf-term (d=0.1)
tf-term web-idf-term (d=0.0)
単語idfは変化が少ない
名詞列idfで大幅改善
それでも精度
は今ひとつ
F=.4
F=.3
Precision
Recall
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8

名詞列のidf相当尺度?

スコア(w) = tfterm(w) × idf相当尺度(w)

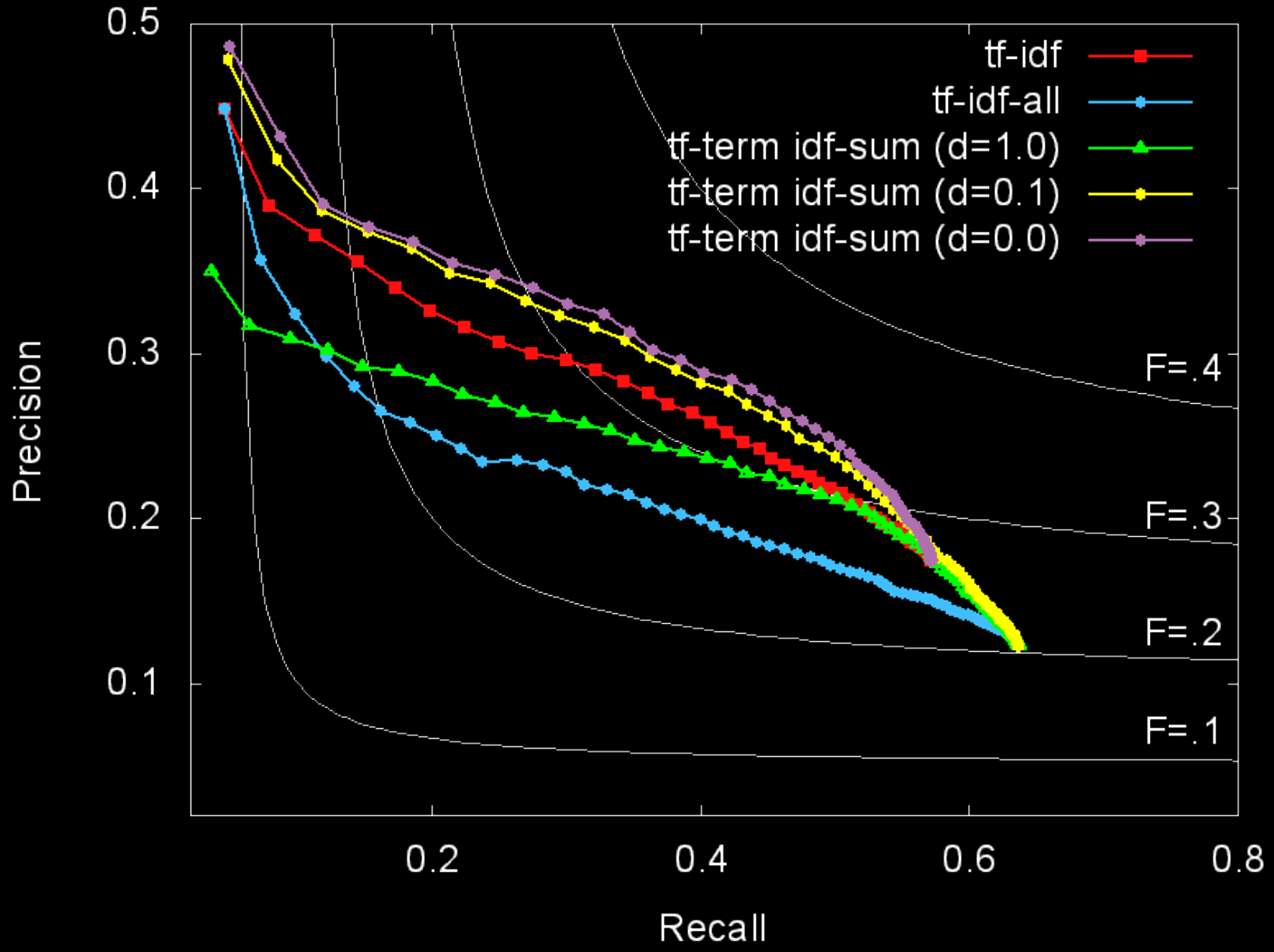
1. 名詞列のidfを直接算出 (tf-term idf-term)

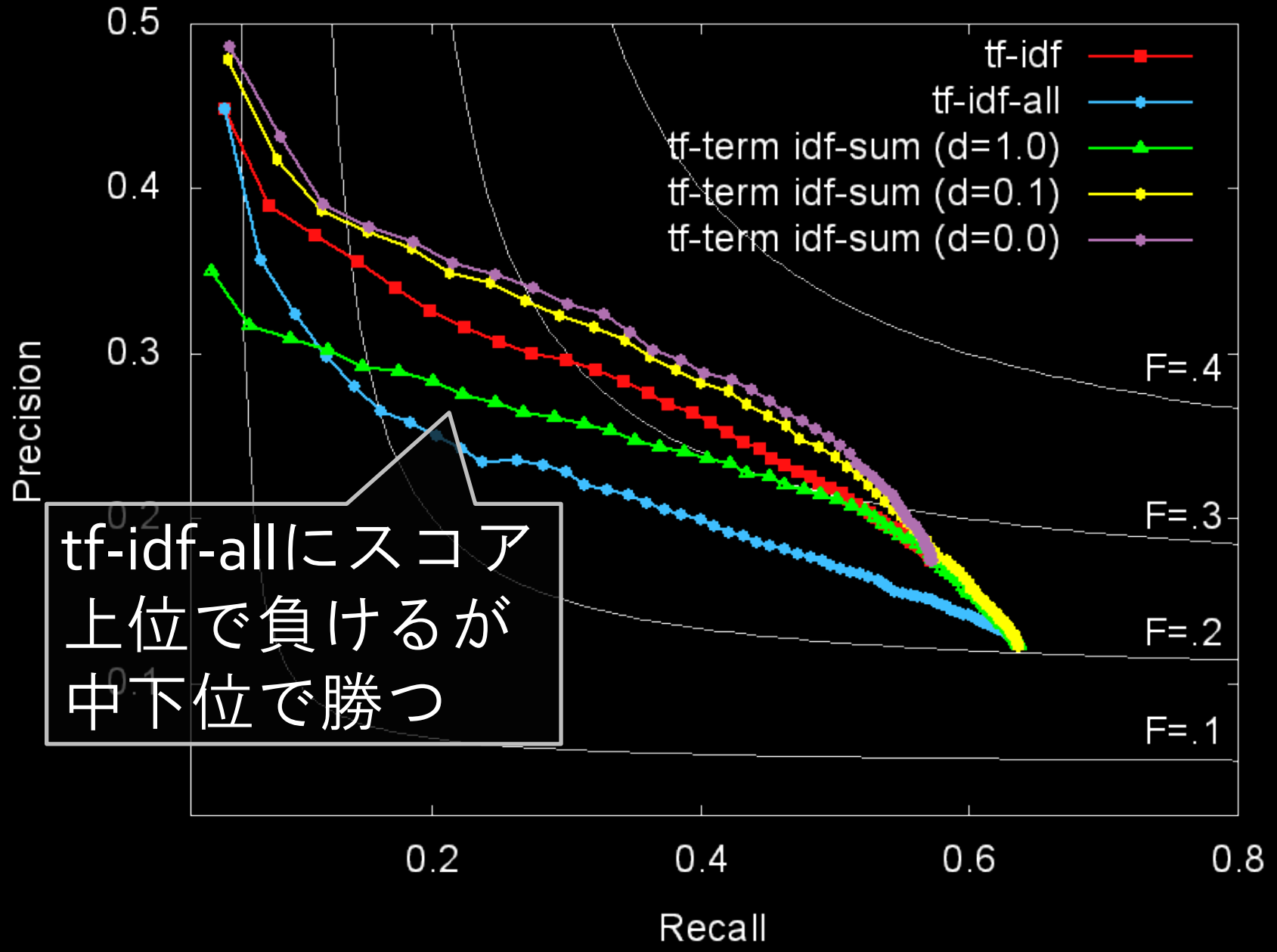
$$\mathrm{idfterm}(\mathbf{w}) = \log(D / D_{\mathbf{w}})$$

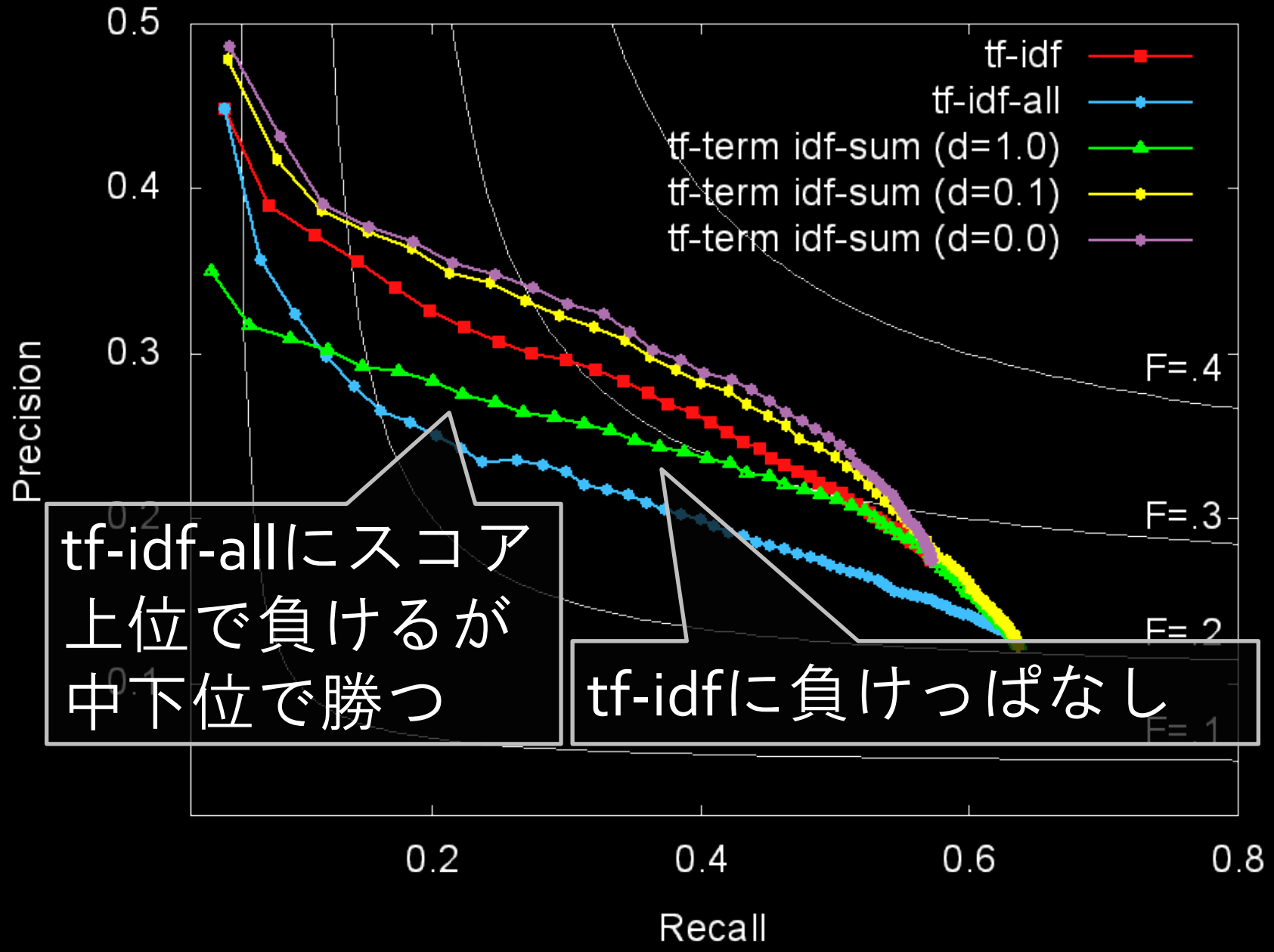
$$D_{\mathbf{w}} = \left|\left\{doc \middle| \mathrm{tfterm}_{doc}(\mathbf{w}) > 0\right\}\right|$$

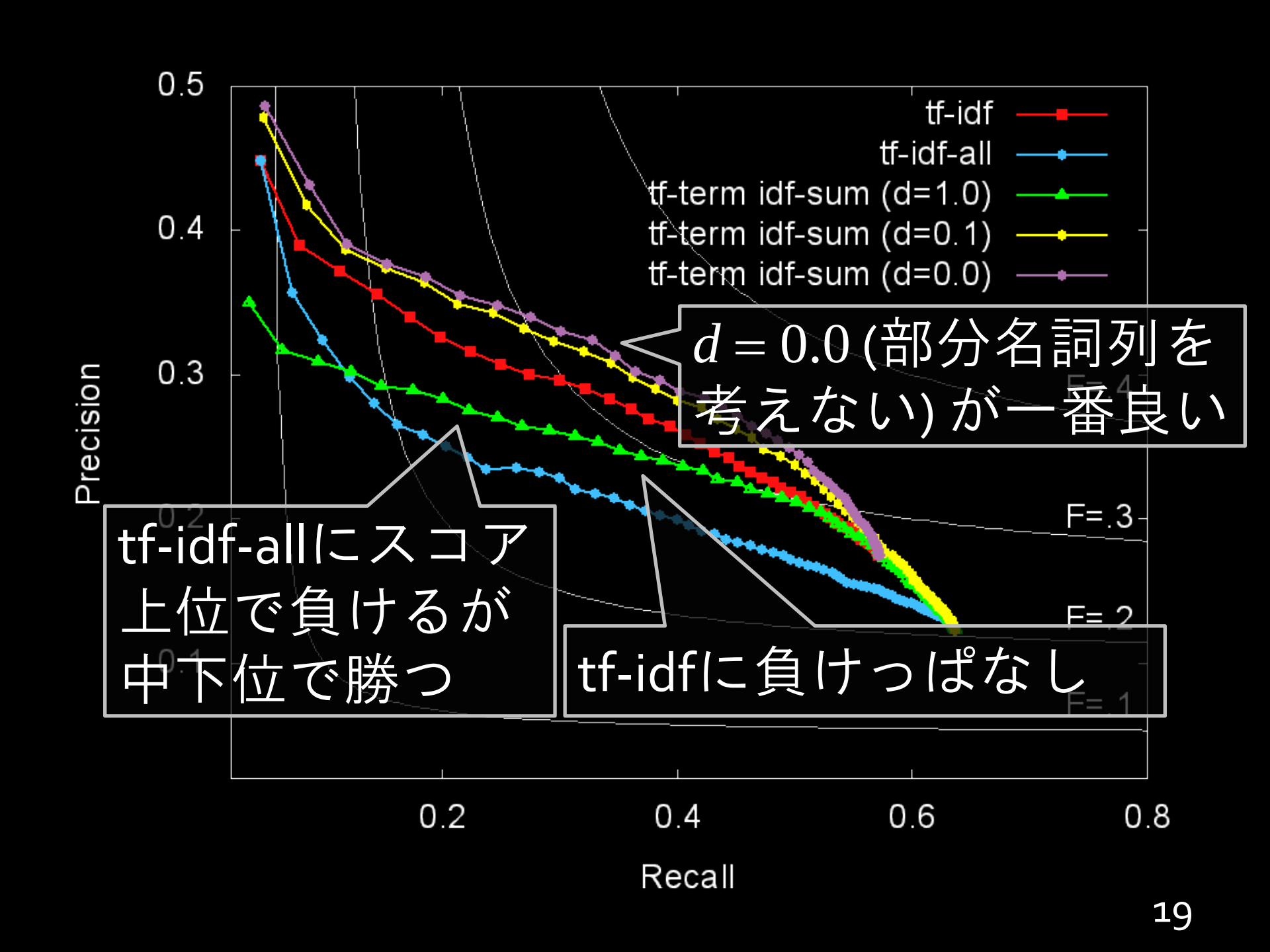
2. 単語idfの総和 (tf-term idf-sum)

$$\mathrm{idfsum}(\mathbf{w}) = \sum_{i} \log(D / D_{w_i})$$


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-sum (d=1.0)
tf-term idf-sum (d=0.1)
tf-term idf-sum (d=0.0)
F=.4
F=.3
F=.2
F=.1


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-sum (d=1.0)
tf-term idf-sum (d=0.1)
tf-term idf-sum (d=0.0)
F=.4
F=.3
F=.2
F=.1
tf-idf-allにスコア
上位で負けるが
中下位で勝つ


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-sum (d=1.0)
tf-term idf-sum (d=0.1)
tf-term idf-sum (d=0.0)
F=.4
F=.3
F=.2
F=.1
tf-idf-allにスコア上位で負けるが中下位で勝つ
tf-idfに負けっぱなし


0.5
0.4
0.3
0.2
0.1
Precision
0.2
0.4
0.6
0.8
Recall
tf-idf
tf-idf-all
tf-term idf-sum (d=1.0)
tf-term idf-sum (d=0.1)
tf-term idf-sum (d=0.0)
F=.4
F=.3
F=.2
F=.1
$d = 0.0$ (部分名詞列を考えない) が一番良い
tf-idf-allにスコア上位で負けるが中下位で勝つ
tf-idfに負けっぱなし

議論

- キーフレーズ抽出 ≒ 文書要約
  - optimal control problems をキーフレーズとして採用したら、optimal control は不要
  - 今回のデータセットには長い名詞列をキーフレーズとして採用する傾向が見られる

tf-idf-all	tf-term idf-sum (d=0.1)	tf-term idf-sum (d=1.0)
new optimal control problems	new optimal control problems	new optimal control problems
optimal control problems	optimal control	optimal control
☹ new optimal control	control	optimal control problems
optimal control	optimal control problems	control
☹ control problems	☹ new optimal control	☹ control problems
control	☹ control problems	problems
problems	problems	☹ new optimal control
…	…	…

まとめと今後の課題

- 教師なしキーフレーズ抽出ではtf-idf法がいまだ強い
- 名詞句の内部構造を考慮した名詞列tfを提案
- 名詞列のidf相当尺度で頑健なものとは?