フレーズベースTF-IDF: 名詞句解析の応用

九州大学
村脇 有吾

2013年11月15日　NL214@屋久島

今回の内容

- 教師なしキーフレーズ抽出のstate-of-the-artは、単語TF-IDFの和による候補のスコア付け
- フレーズ単位でTF-IDFをカウントしたら、長い文書については単語TF-IDFの和に勝った
  - 候補のカウントに名詞句解析を応用
  - でも別に名詞句解析を使わなくても良かった

名詞句に着目する背景 1/2

- 教師なし係り受け解析の混迷
  - 隣にかけるベースライン (38%※) に長く勝てず
  - DMV (Klein+, 2004) がようやく勝った (46%)
  - state-of-the-artで65-70%だが...

※WSJ Sec23 10単語以下の文の係り受け精度 (Blunsom+, 2010)

名詞句に着目する背景 1/2

- 教師なし係り受け解析の混迷
  - 隣にかけるベースライン (38%※) に長く勝てず
  - DMV (Klein+, 2004) がようやく勝った (46%)
  - state-of-the-artで65-70%だが...
- 係り受け導出という問題設定が怪しいのでは?
  - 特に英語だと恣意的な係り受け規則が多い (e.g. 冠詞は主辞か?, A and Bの主辞は?)
  - 言語が系列データであることを無視し過ぎでは?

※WSJ Sec23 10単語以下の文の係り受け精度 (Blunsom+, 2010)

名詞句に着目する背景 2/2

- 系列モデル
  - N-gram言語モデルの実用上の成功
  - 言語は実際に系列として観測される
  - でも系列では説明できない構造を持っているのは明らか

名詞句に着目する背景 2/2

- 系列モデル
  - N-gram言語モデルの実用上の成功
  - 言語は実際に系列として観測される
  - でも系列では説明できない構造を持っているのは明らか

- 系列と係り受け (もしくはその代替物) をうまく両立させられないか?

名詞句の内部構造

自然　言語　処理

名詞句の内部構造

音声 ??

自然　言語　処理

名詞句の内部構造

「自然言語処理」という系列がまるごと
心的語彙に入っているはず

自然　言語　処理

名詞句の内部構造

「自然言語処理」という系列がまるごと
心的語彙に入っているはず

内部構造は再帰的

自然　言語　処理

名詞句の内部構造

「自然言語処理」という系列がまるごと
心的語彙に入っているはず

内部構造は再帰的

並列　自然　言語　処理

名詞句の内部構造

「自然言語処理」という系列がまるごと
心的語彙に入っているはず

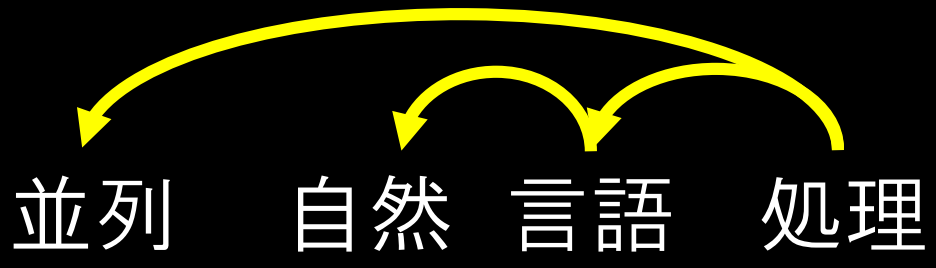
内部構造は再帰的

並列　自然　言語　処理

「並列」は「自然言語処理」と
その場で統語的に結びついている

関連研究1:
名詞句解析 = 係り受け

- 従来研究はいずれも、名詞句解析を単なる係り受けの問題とみなす



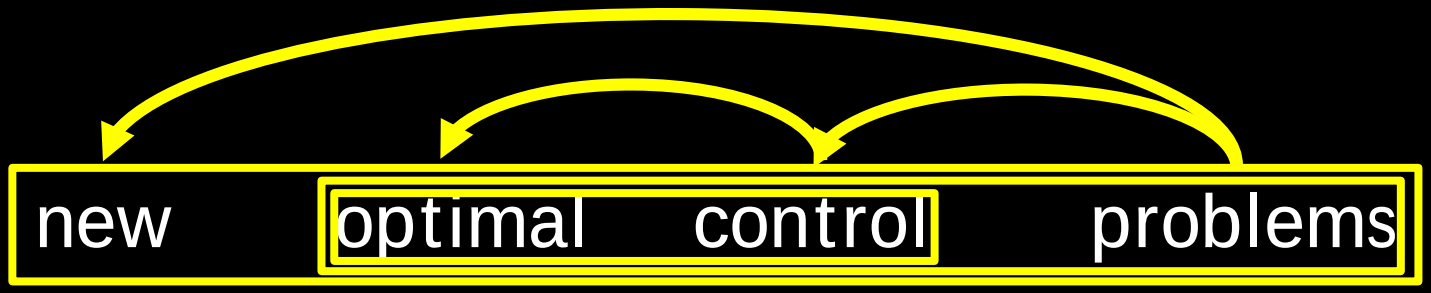
(Lauer, 1995) 以降の諸研究

関連研究2: トピックモデルの拡張 = 系列

- LDA with collocations (Griffiths+, 2007)
  - 頻出するbigramの連鎖をフレーズとみなす

- Adaptor Grammar for topical collocations (Johnson, 2010)
  - 頻出するフレーズをモデルが覚える
  - そのままでは再帰を扱えない

エッジとスパンを考慮した教師あり名詞句解析 (Murawaki+ 2012)



- 係り受け（エッジ）と句 (スパン) は一対一対応
- 両方を考慮して候補にスコア付け

score = edgeScore(new ← problems)
+ edgeScore(control ← problems)
+ edgeScore(optimal ← control)
+ spanScore(optimal control)
+ spanScore(optimal control problems)
+ spanScore(new optimal control problems)

エッジとスパンを考慮した教師あり名詞句解析 (Murawaki+ 2012)

- 少量の教師データに加えてウェブ統計を利用
  - 人間が正しく認識できるのは既に知っているから
  - 専門外の名詞句は人間でもアノテーションが難しい
- 以前トピックモデルを教師なし係り受けモデルで拡張してみたが、全然駄目だった
  - 係り受け精度が、右隣ベースラインに勝てない
  - 名詞句の内部構造を小規模テキストから導出するのは困難

名詞句解析の応用先: 教師なしキーフレーズ抽出

入力:
文書集合

出力:
各文書に対する
キーフレーズ集合

- bipartite graph
- vertices
- fault-tolerant hamiltonian laceability
- tight bounds

- collaborative software development
- mobility impairments
- accessibility

- profit analysis
- feedforward control
- state space model

単語ベースTF-IDF (Hasan+ 2010)

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

$$\mathrm{unit}_{doc}(\mathbf{w}) = \mathrm{I}(\mathbf{w} \in \mathrm{longest}(doc))$$

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

- State-of-the-art!
- キーフレーズ候補wのスコア = 単語TF-IDFの和
- キーズレーズ候補集合: 文書中で少なくとも一度最長名詞句として出現

単語ベースTF-IDF (Hasan+ 2010)

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

$$\mathrm{unit}_{doc}(\mathbf{w}) = \mathrm{I}(\mathbf{w} \in \mathrm{longest}(doc))$$

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

- State-of-the-art!
- **キーフレーズ候補wのスコア** = 単語TF-IDFの和
- キーズレーズ候補集合: 文書中で少なくとも一度最長名詞句として出現

最長名詞句と部分名詞句

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 ...

最長名詞句と部分名詞句

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. 最長名詞句 (チャンキングで高精度に認識)

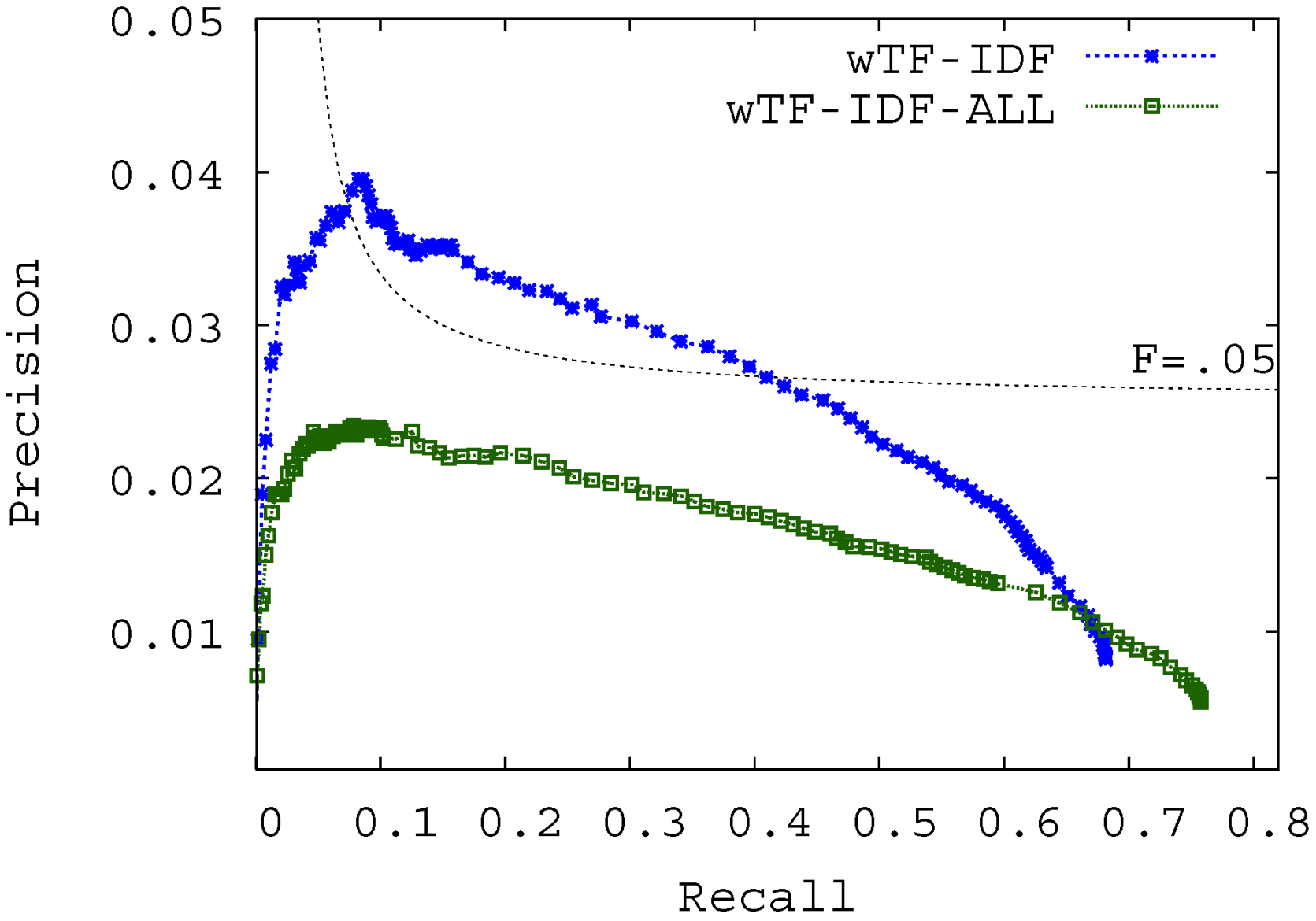
最長名詞句と部分名詞句

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

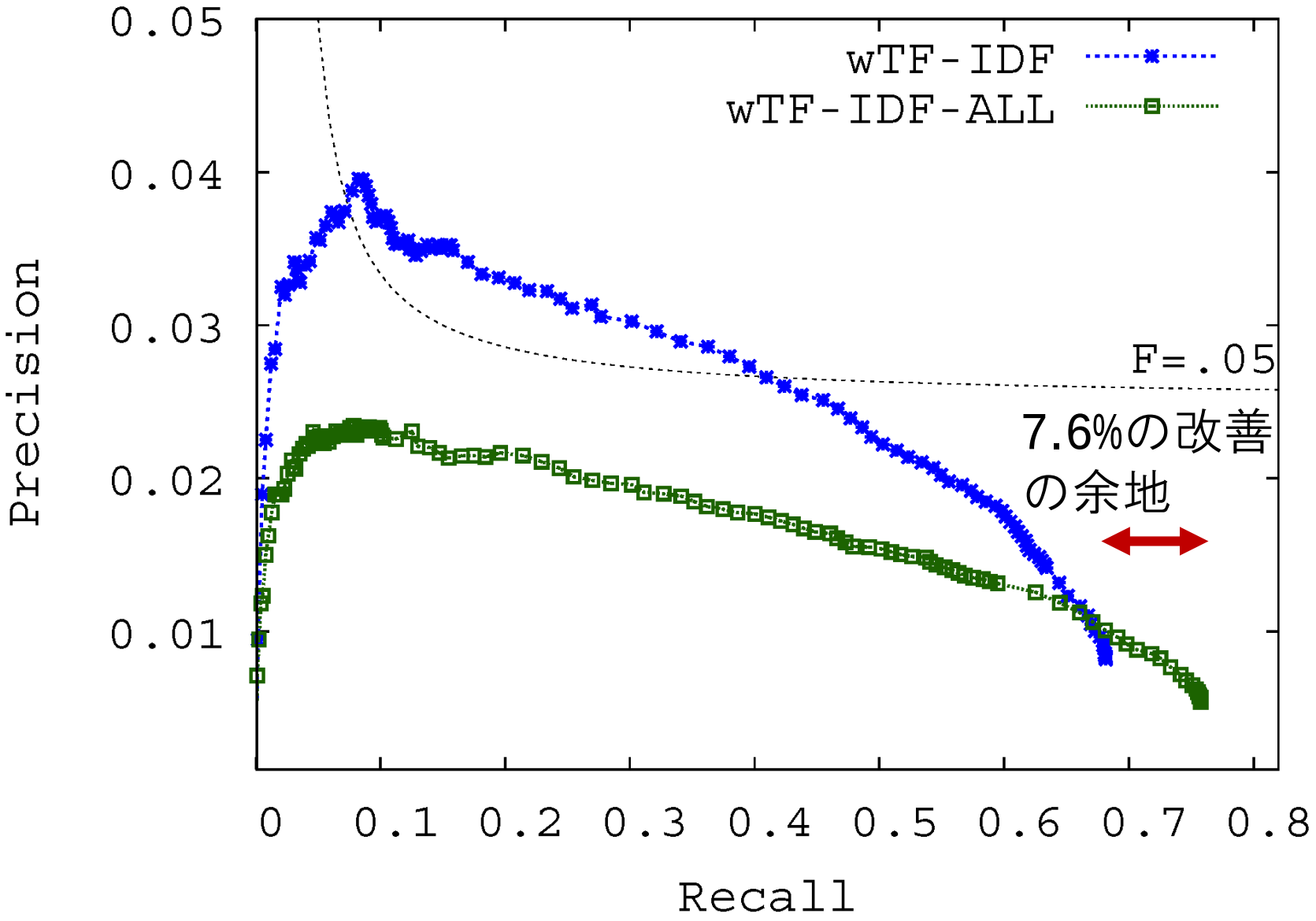
new optimal control problems are considered for distributed systems described by ...

1. 最長名詞句 (チャンキングで高精度に認識)
2. 部分名詞句

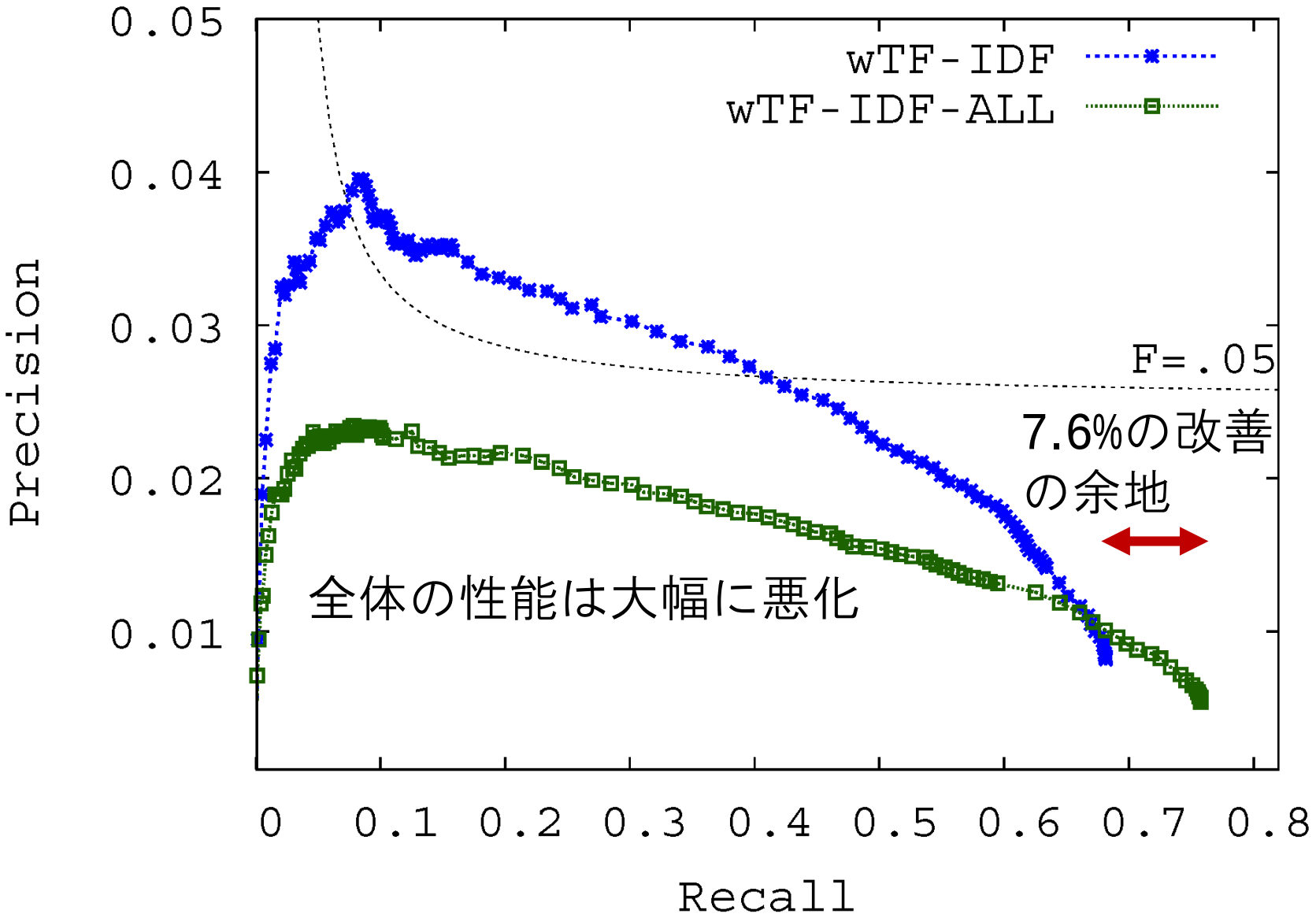
最長名詞句の制約を取っ払うと…



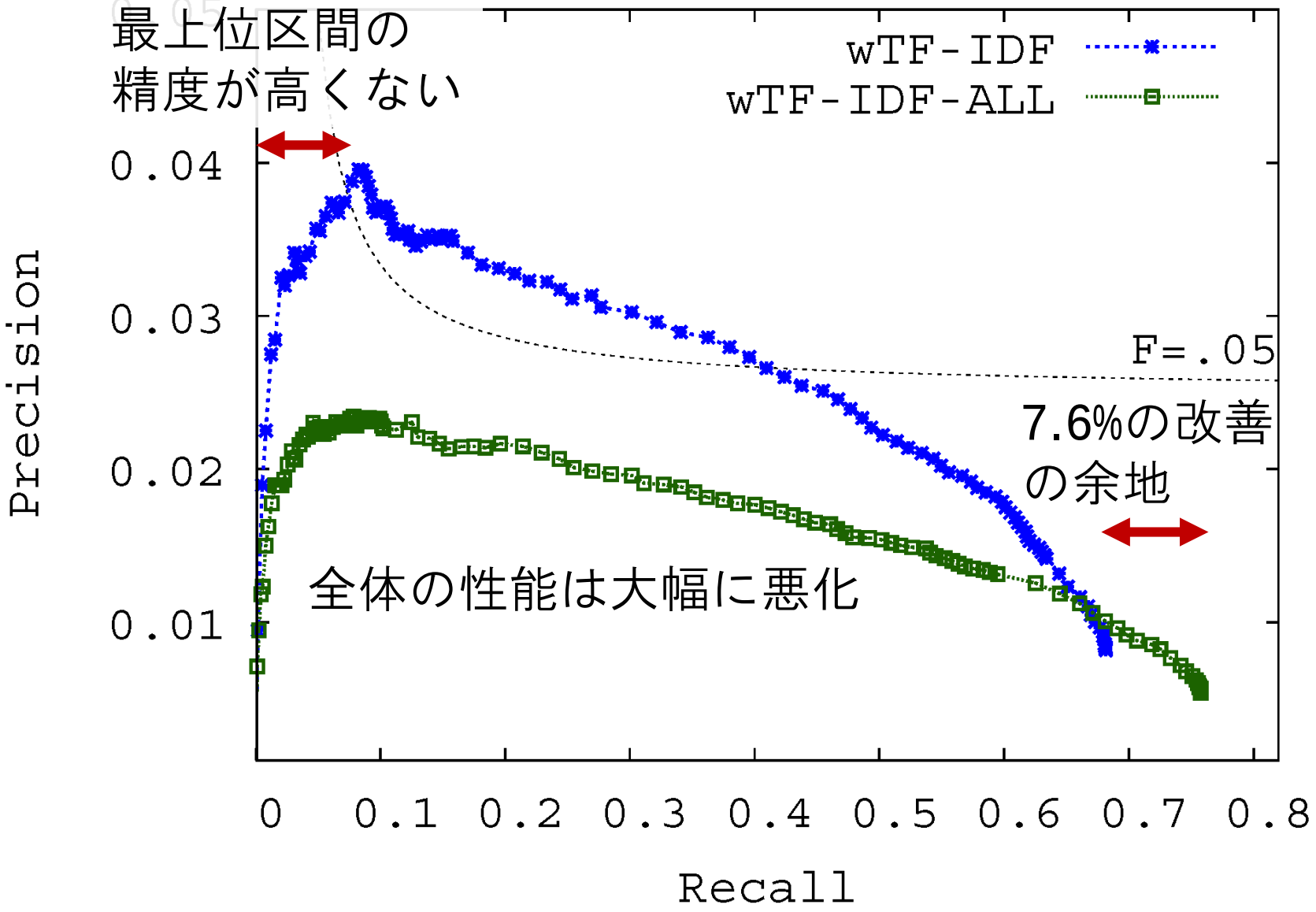
最長名詞句の制約を取っ払うと…



最長名詞句の制約を取っ払うと…



最長名詞句の制約を取っ払うと…



キーフレーズの性質

- 構造的に誤ったフレーズ候補を除きたい
  - control problems
- 最長名詞句は構造的に正しいが、キーフレーズとは限らない
  - new optimal control problems
- new optimal control problemsより optimal control problems を優先したい

キーフレーズの性質

- 構造的に誤ったフレーズ候補を除きたい
  - control problems

文法性

- 最長名詞句は構造的に正しいが、キーフレーズとは限らない
  - new optimal control problems
- new optimal control problemsより optimal control problems を優先したい

キーフレーズの性質

- 構造的に誤ったフレーズ候補を除きたい
  - control problems
- 最長名詞句は構造的に正しいが、キーフレーズとは限らない
  - new optimal control problems
- new optimal control problemsより optimal control problems を優先したい

文法性

語彙性

名詞句解析による擬似頻度

new optimal control problems

freq. 1.00

名詞句解析による擬似頻度

new optimal control problems

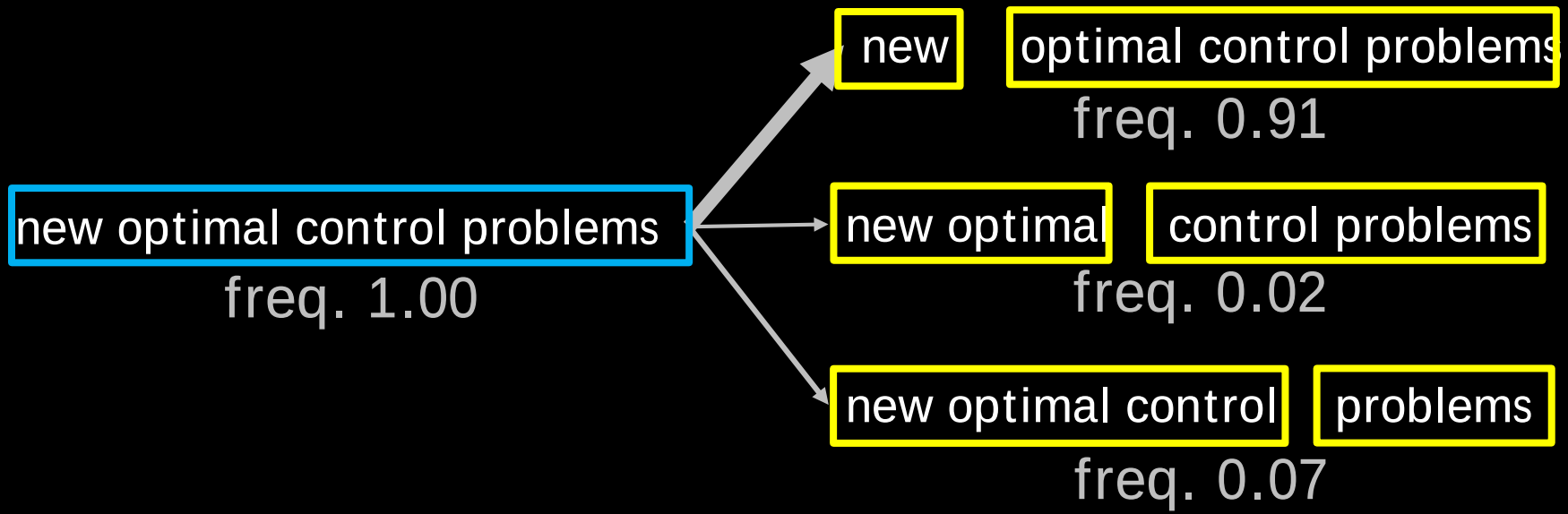
freq. 1.00

new | optimal control problems

new optimal | control problems

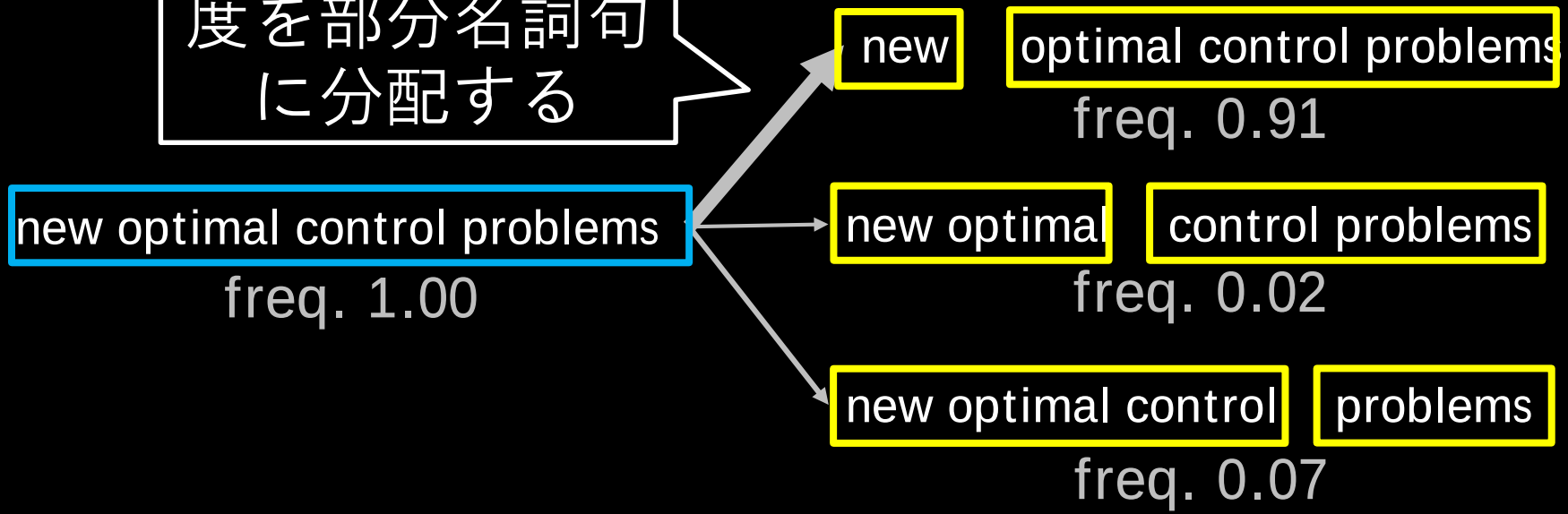
new optimal control | problems

名詞句解析による擬似頻度

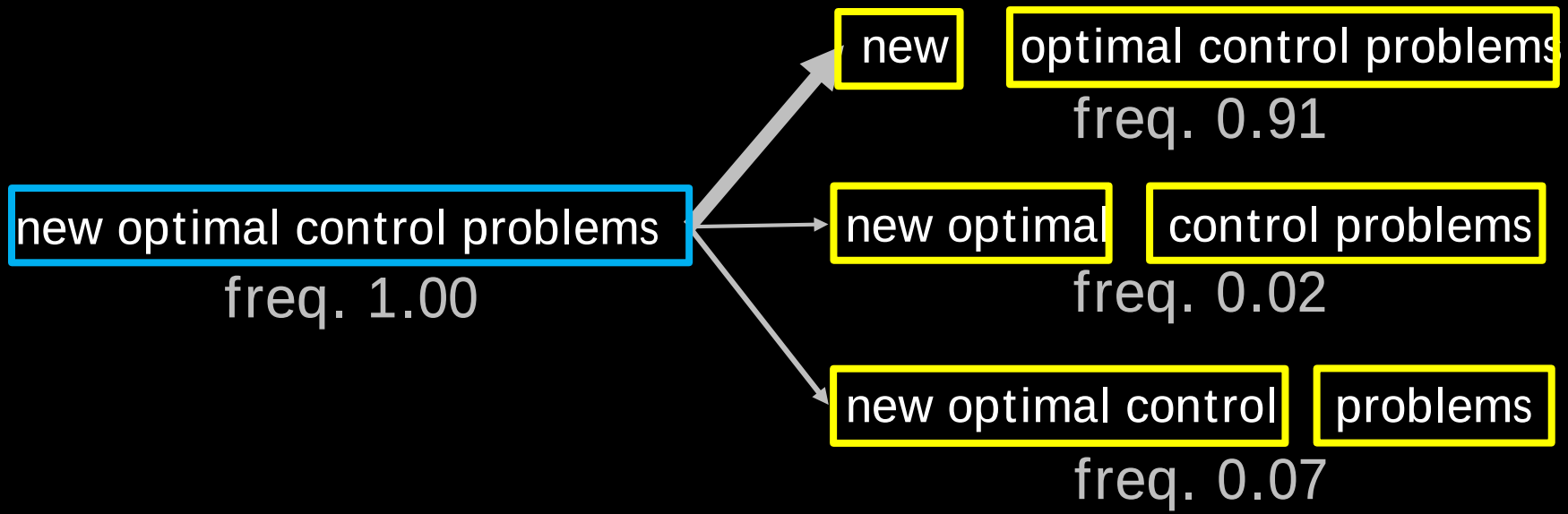


名詞句解析による擬似頻度

最長名詞句の頻度を部分名詞句に分配する

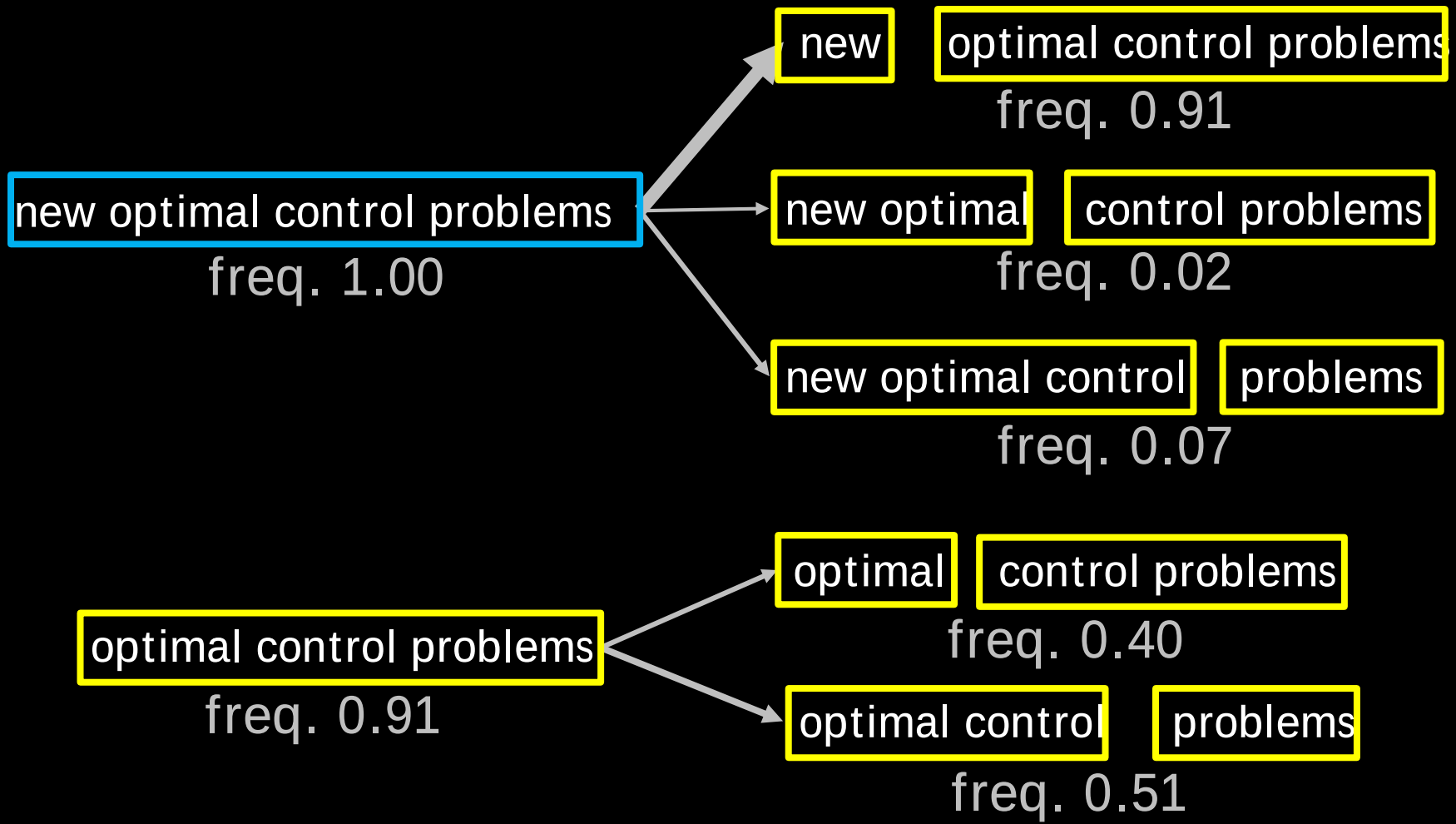


名詞句解析による擬似頻度

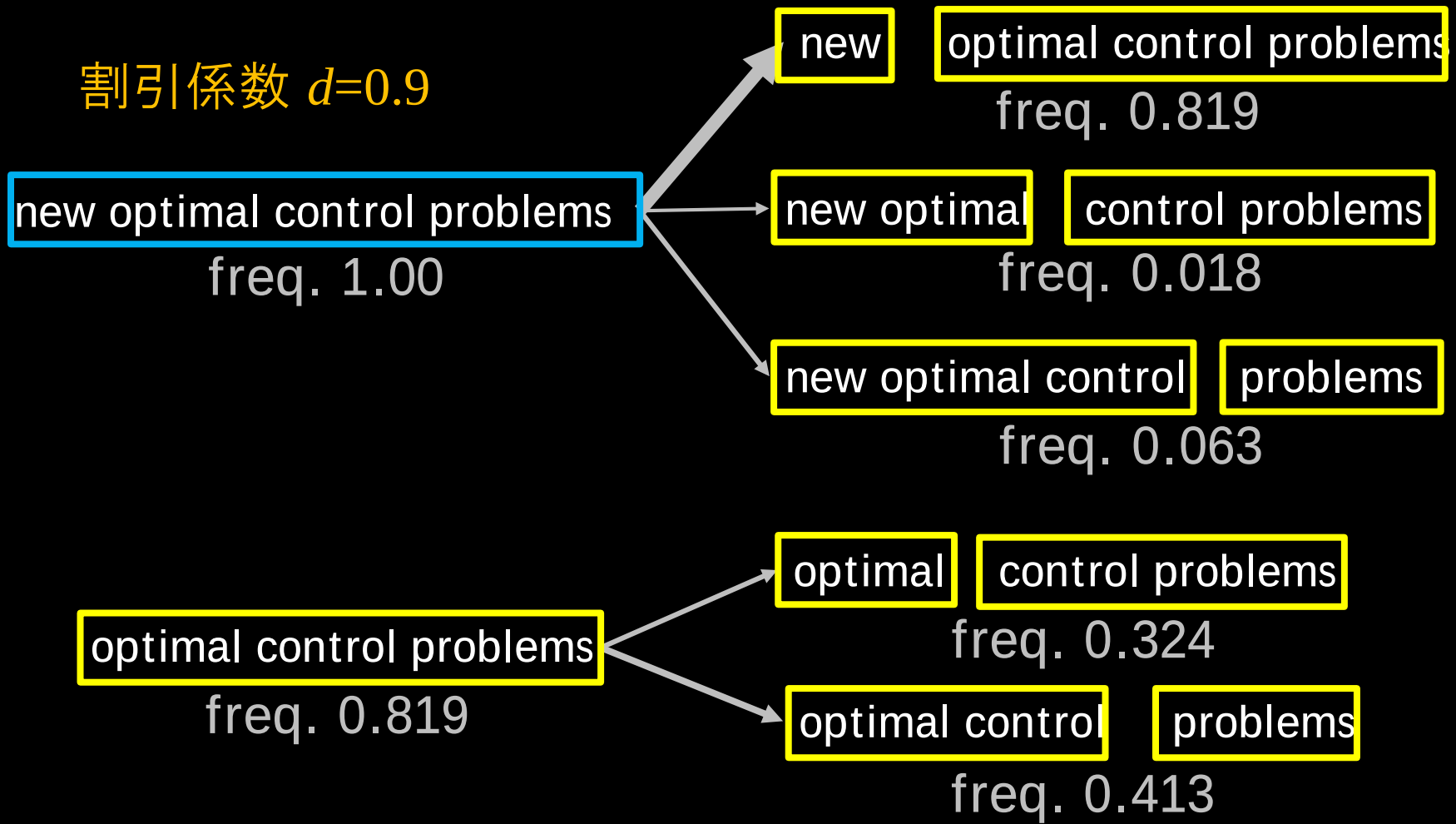


optimal control problems

名詞句解析による擬似頻度



名詞句解析による擬似頻度



名詞句解析による擬似頻度

	擬似カウント	
new optimal control problems	1.0	
optimal control problems	0.819	
new optimal control	0.063	
optimal control	0.467	(0.413 + 0.054)
control problems	0.342	(0.018 + 0.324)

⋮

名詞句解析による擬似頻度

	擬似カウント	
new optimal control problems	1.0	
optimal control problems	0.819	
new optimal control	0.063	
optimal control	0.467	(0.413 + 0.054)
control problems	0.342	(0.018 + 0.324)

⋮

文法性問題は解消された

名詞句解析による擬似頻度

	擬似カウント	
new optimal control problems	1.0	
optimal control problems	0.819	
new optimal control	0.063	
optimal control	0.467	(0.413 + 0.054)
control problems	0.342	(0.018 + 0.324)

⋮

語彙的フレーズに高いスコアを与えるには?

文法性問題は解消された

フレーズベースTF-IDF

- TF: 文書内の擬似カウントの総和
  - キーフレーズ: 中～大
  - 一般フレーズ: 大
  - 非文法的フレーズ: 小
  - 非語彙的フレーズ: 小

フレーズベースTF-IDF

- TF: 文書内の擬似カウントの総和
  - キーフレーズ: 中～大
  - 一般フレーズ: 大
  - 非文法的フレーズ: 小
  - 非語彙的フレーズ: 小

new optimal control problemのような非語彙的フレーズは頻出しない

フレーズベースTF-IDF

- TF: 文書内の擬似カウントの総和
  - キーフレーズ: 中～大
  - 一般フレーズ: 大
  - 非文法的フレーズ: 小
  - 非語彙的フレーズ: 小
- IDF: 出現文書数の逆数 (のlog)
  - キーフレーズ: 中
  - 一般フレーズ: 小
  - 非文法的フレーズ: 中～大
  - 非語彙的フレーズ: 大

new optimal control problemのような非語彙的フレーズは頻出しない

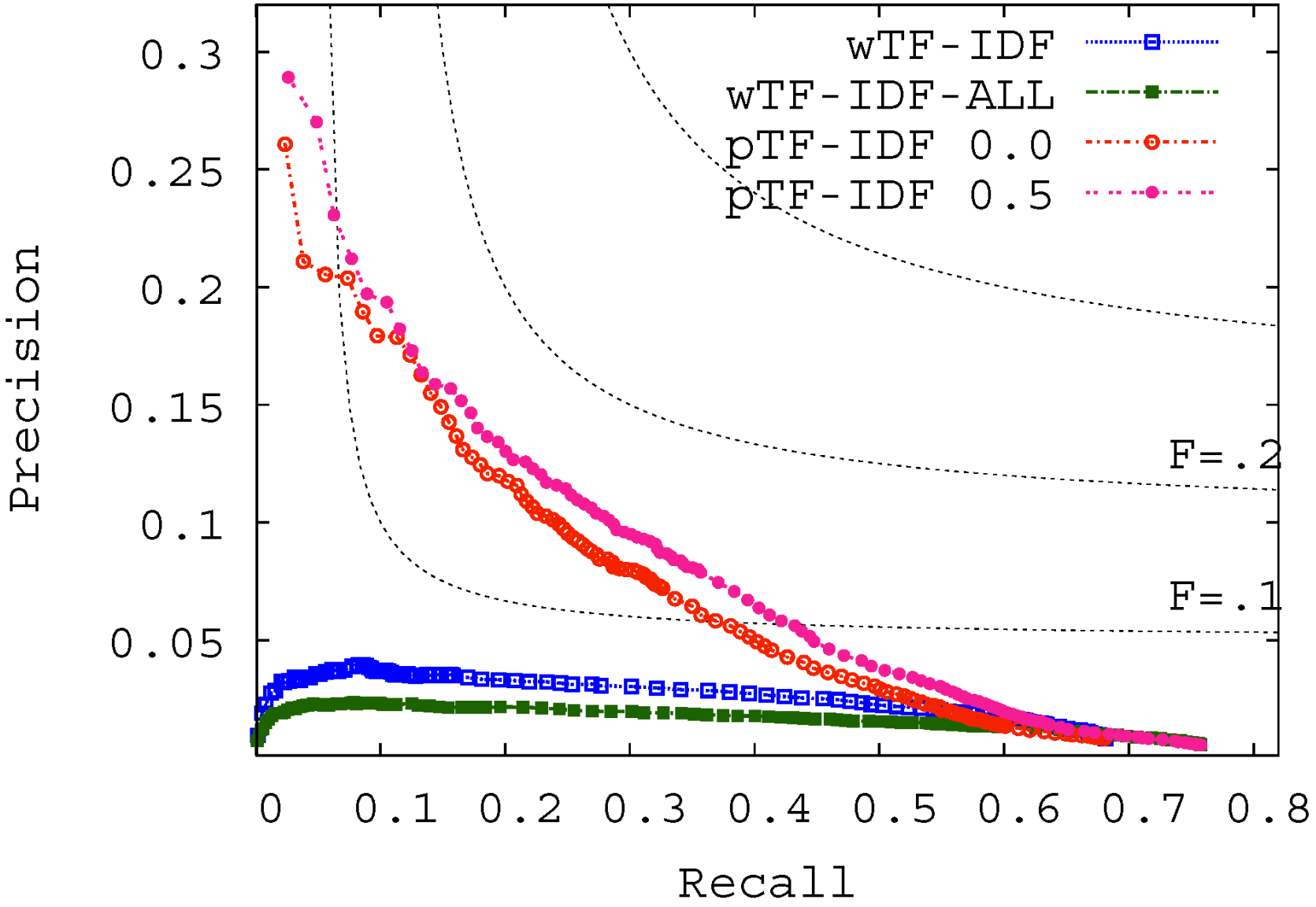
フレーズベースTF-IDF

- TF: 文書内の擬似カウントの総和
  - キーフレーズ: 中～大
  - 一般フレーズ: 大
  - 非文法的フレーズ: 小
  - 非語彙的フレーズ: 小
- IDF: 出現文書数の逆数 (のlog)
  - キーフレーズ: 中
  - 一般フレーズ: 小
  - 非文法的フレーズ: 中～大
  - 非語彙的フレーズ: 大

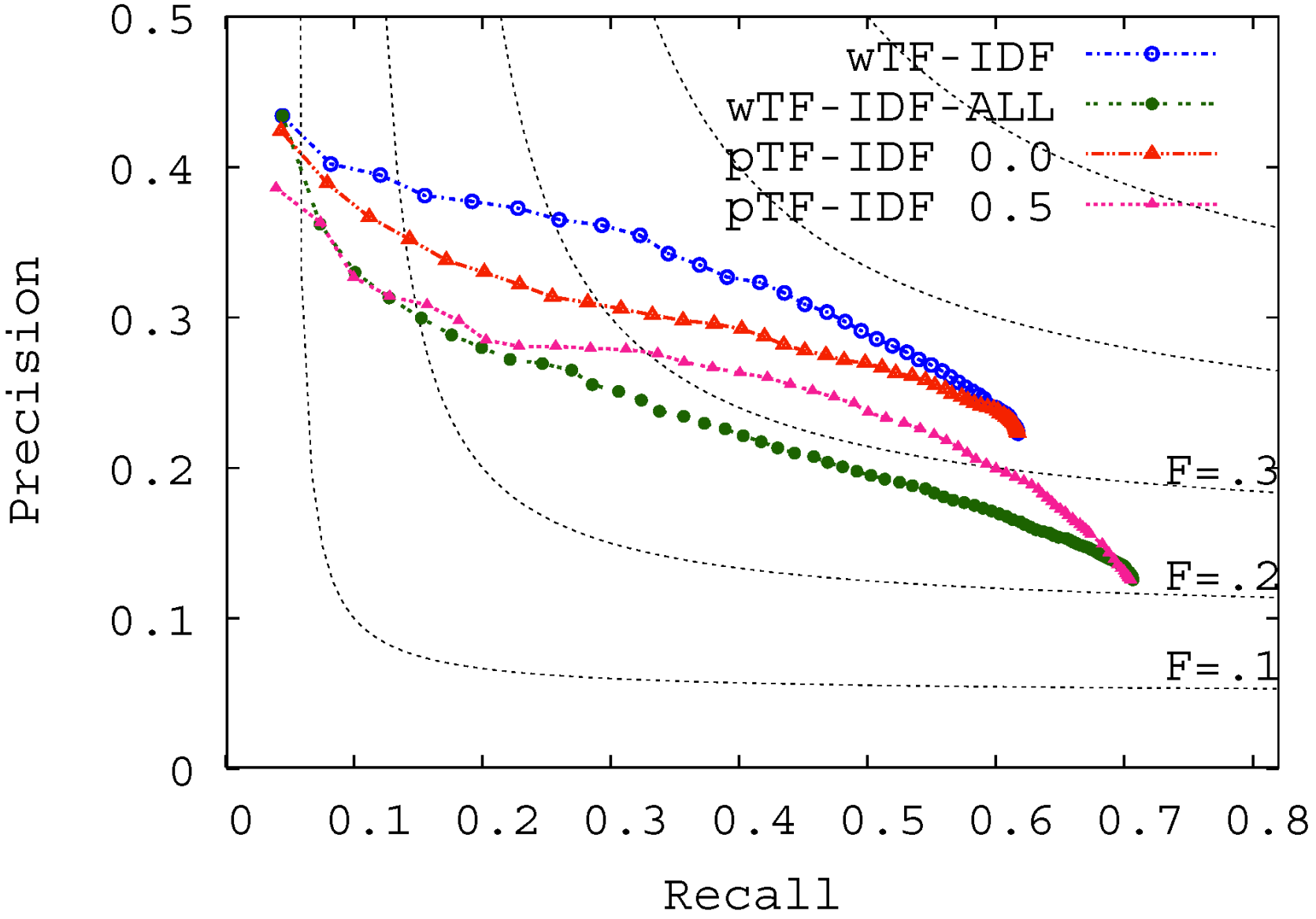
new optimal control problemのような非語彙的フレーズは頻出しない

この性質は好ましくないが、TFと掛け合わせると大丈夫?

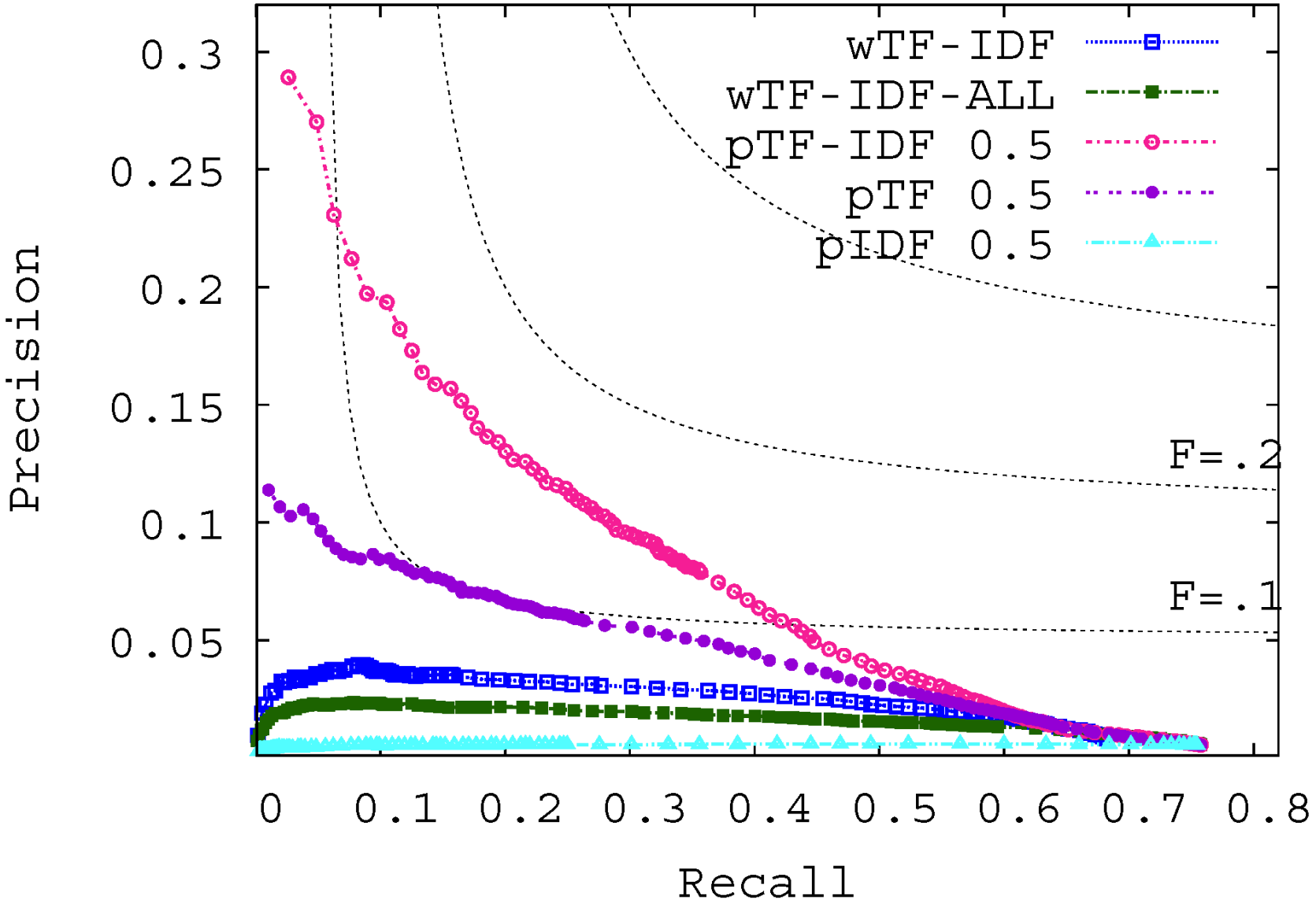
長い文書 (予稿集) での結果



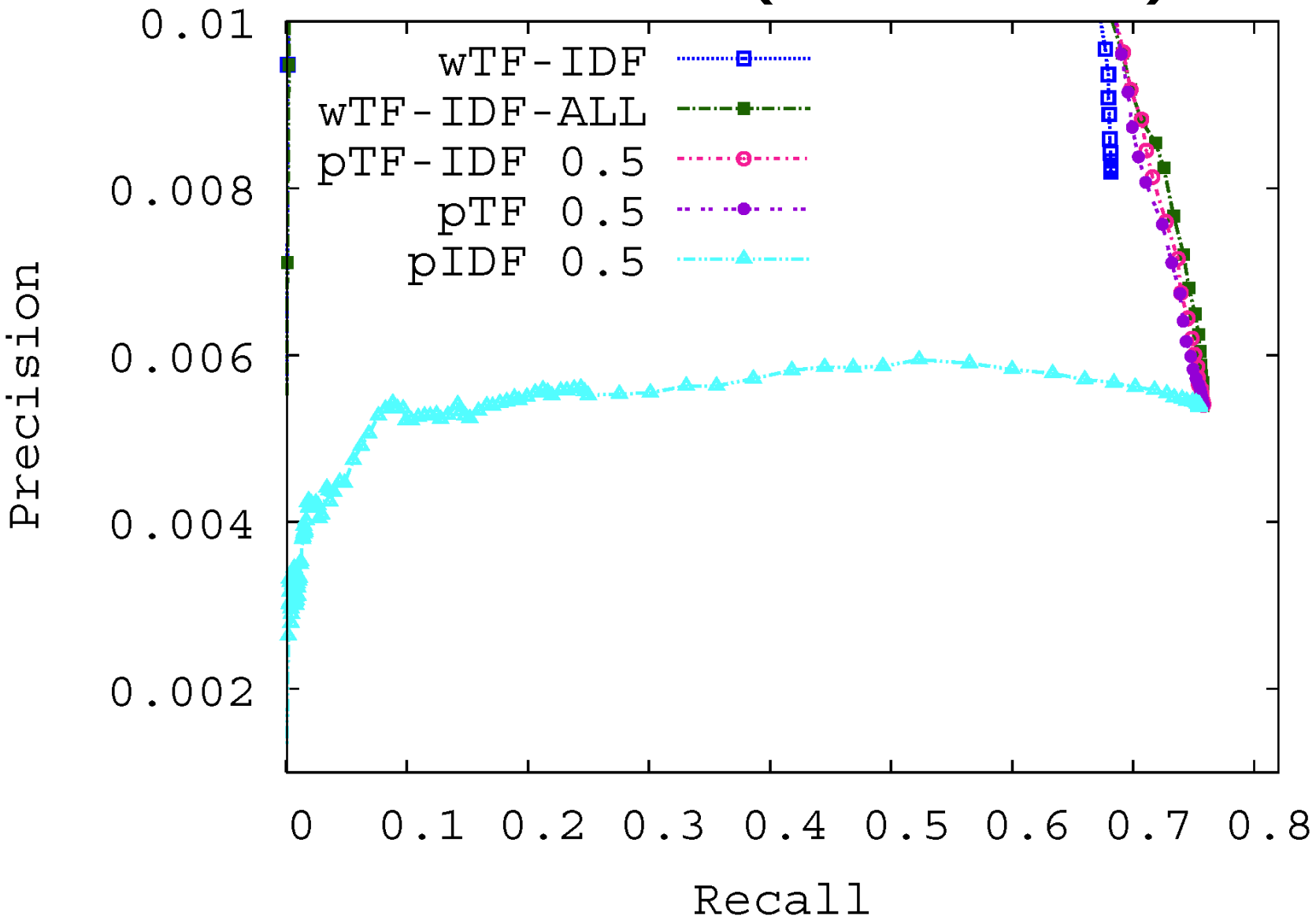
短い文書 (要旨) での結果


wTF-IDF
wTF-IDF-ALL
pTF-IDF 0.0
pTF-IDF 0.5
F=.3
F=.2
F=.1
Precision
Recall
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

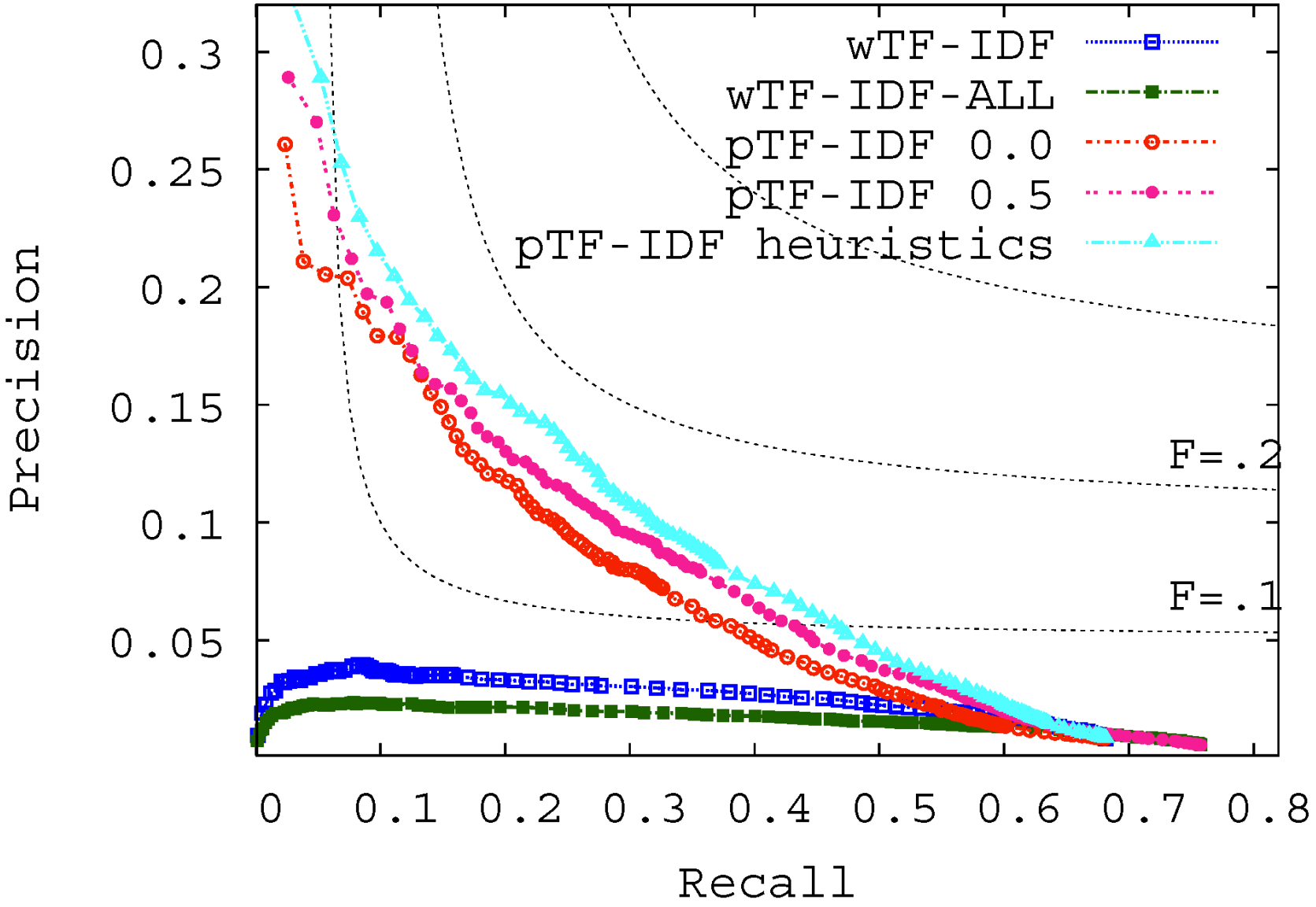
TF/IDF単独との比較 (長い文書)



IDF部分を拡大 (長い文書)



最長名詞句集合を候補集合として、単純にフレーズを数えたら最高性能



まとめと今後の課題

- フレーズ単位でTF-IDFをカウントしても、長い文書ならうまく動いた
- 短い文書に対しても頑健にするには?
- 名詞句では系列と係り受けの両方の性質が重要
- 名詞句以外でも、系列と係り受けを同時に考える方法は?