言語横断LDAモデルを用いた
統計的機械翻訳システム

西尾 拓　福富 崇博　貞光 九月　山本 幹雄

従来の統計的機械翻訳システム

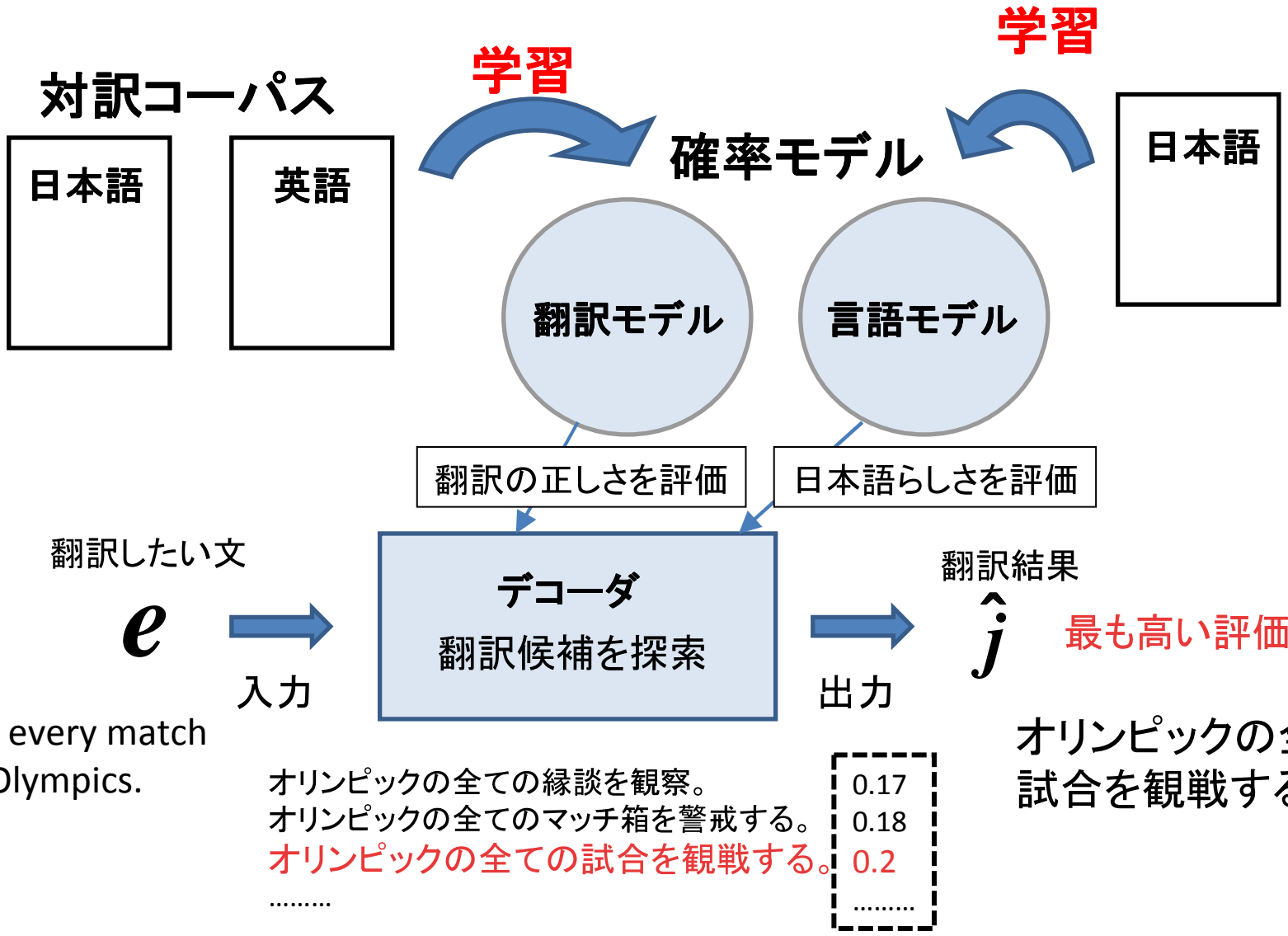
翻訳モデル ＋ **言語モデル**

提案手法

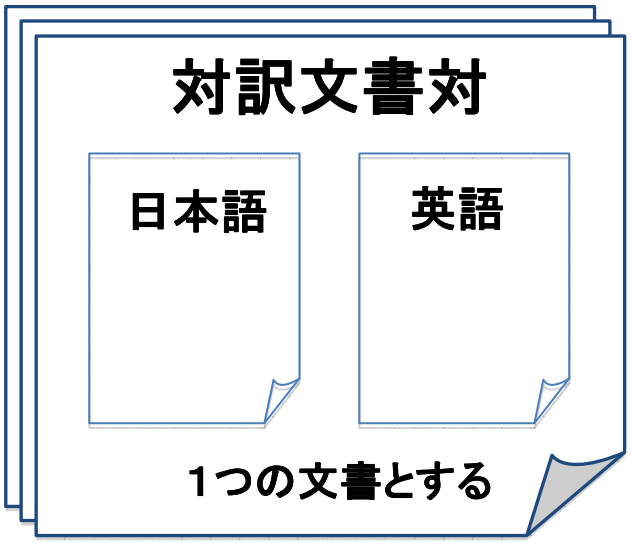
＋ **言語横断LDAモデル**

言語横断LDAモデルの捉えるトピック情報を
統計的機械翻訳システムに持ち込むことで、
翻訳システム全体の翻訳精度向上を実現する。

統計的機械翻訳システム



言語横断LDAモデル



共通のトピックにおける
両言語の単語出現確率を学習

LDA学習

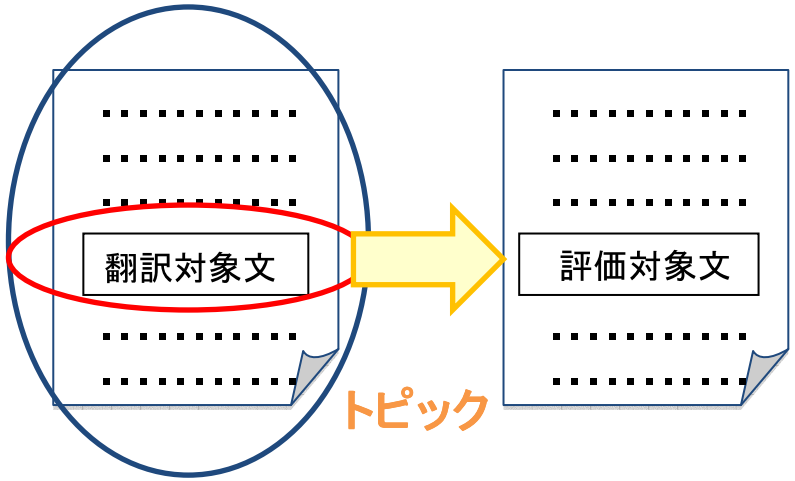
トピック	出現確率の高い単語
WEB	キーワード, 文章, 手続き, サーバ, クライアント
	menu, job, message, password, script, icon,client
画像	シェーディング, dpi, ポリゴン, 画素, 輝度, RGB
	gray, gamma, picture, luminance, dct, rgb
機械工学	内圧, 弁, 噴射, 点火, 燃焼, NOx, アクセル, 燃料
	manifold, spark, ignition, purge, exhaust, nox, fuel

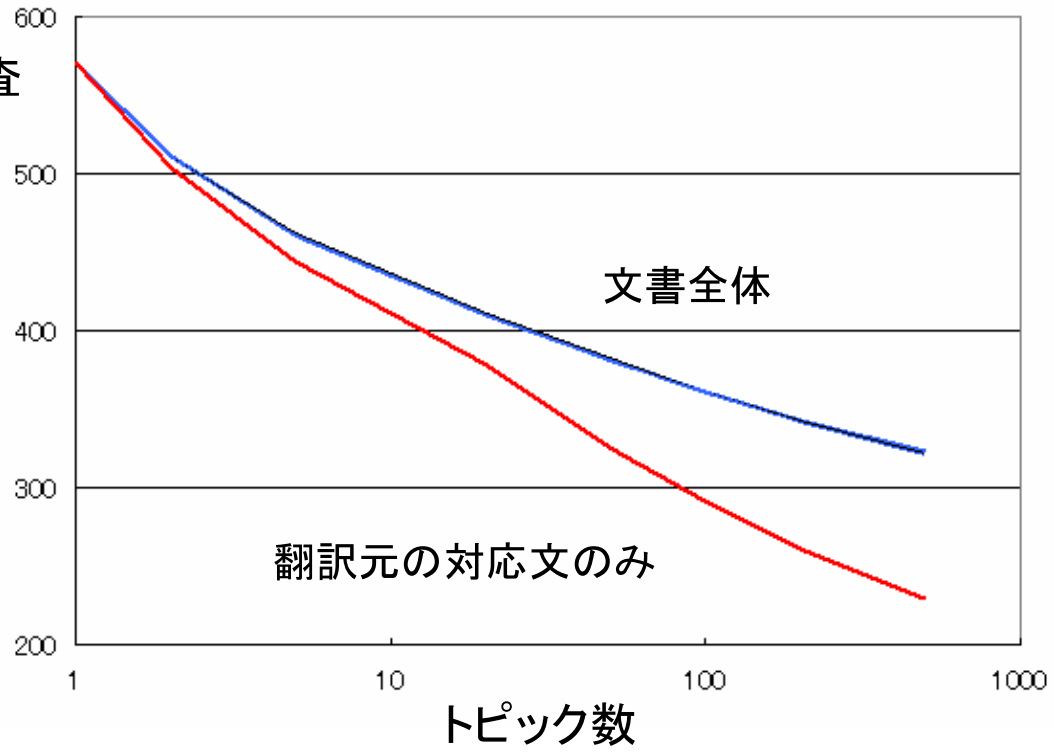
実験条件

統合手法	リランキング法	log-linear model
使用コーパス	特許文180万文（NTCIR7）	
	英語語彙サイズ　121,815	
	日本語語彙サイズ　139,491	
トピック数	500	2,5,10,20,50,100,200
デコーダ	Moses	Pharaohコピー
Reordering table	あり	なし
言語モデル	SRI Language Modeling Toolkit により学習した5gramモデル	
翻訳モデル	フレーズベースモデル	
N-best	100	ー

言語横断LDAモデル
適応範囲の違いによる比較

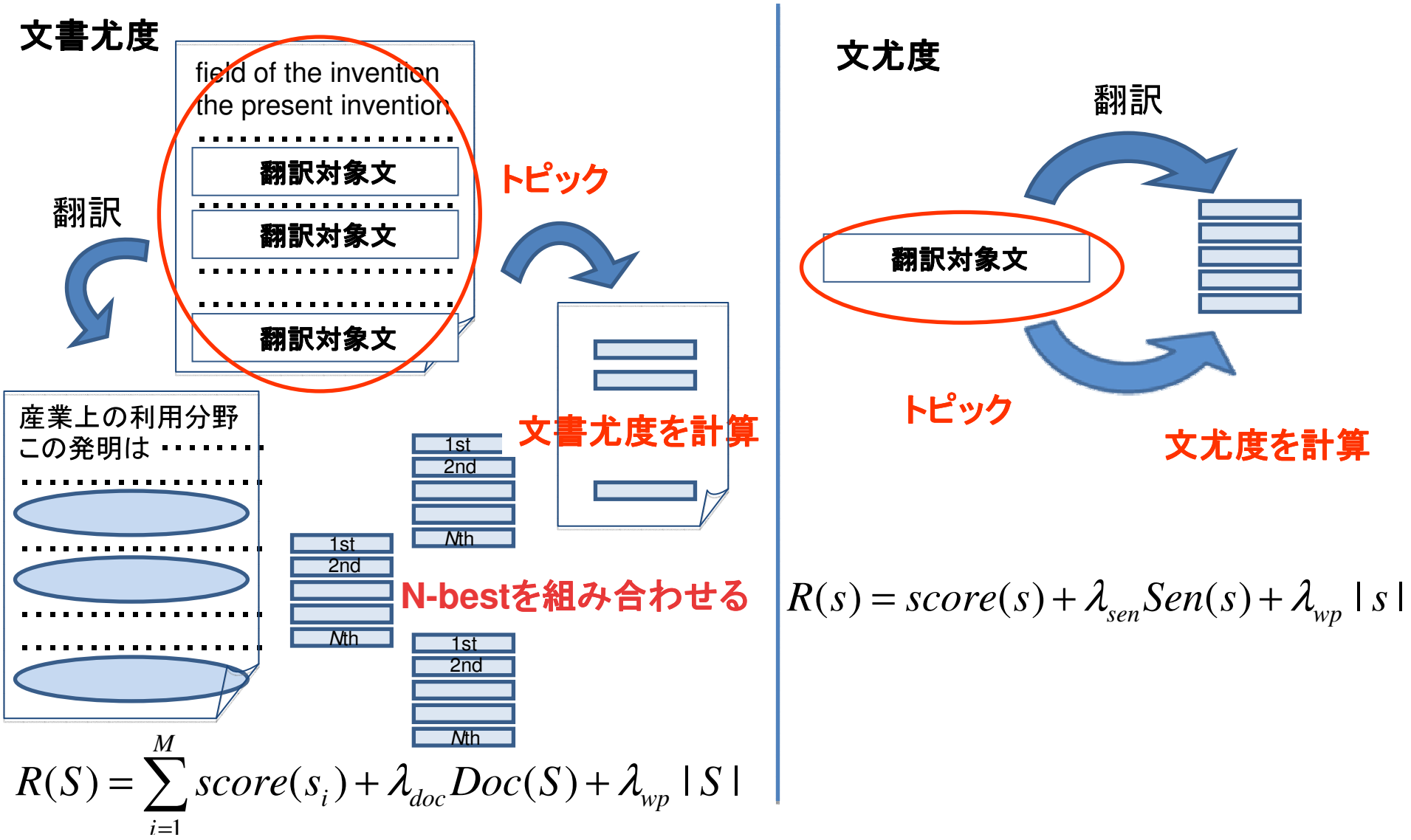
各適応範囲のパープレキシティの推移を調査





適応の範囲を翻訳元の対応文のみに絞った場合に
言語モデルとしての性能が最も良くなる

リランキング法
文書尤度/文尤度



log-linear modelによる統合

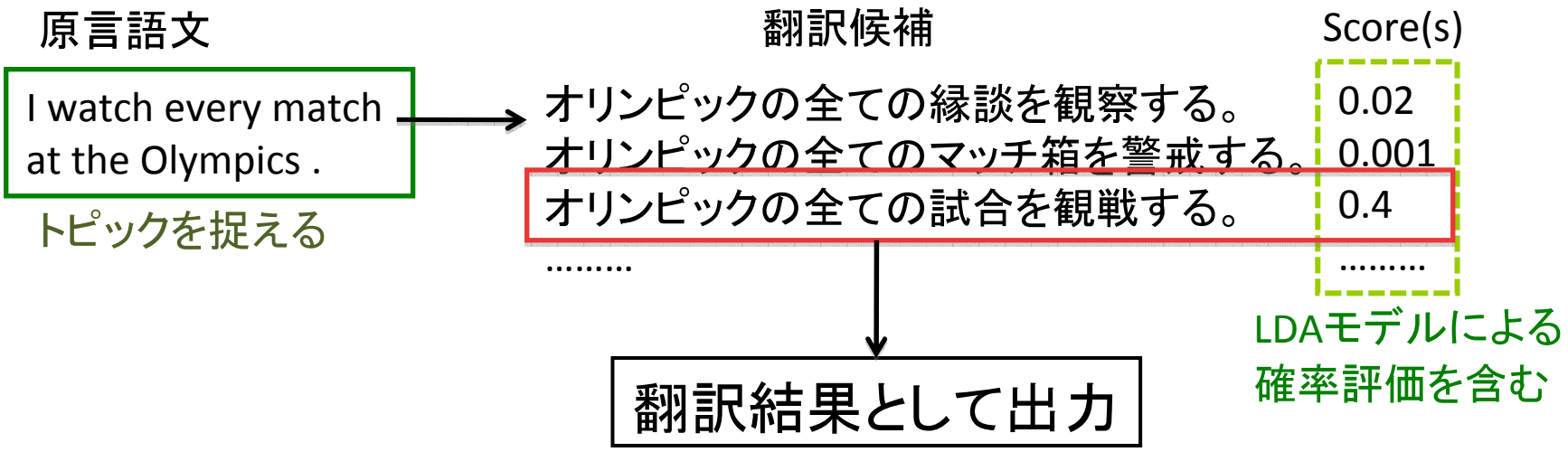
従来の翻訳システムにおける翻訳候補の評価式を変更する

翻訳らしさの評価式

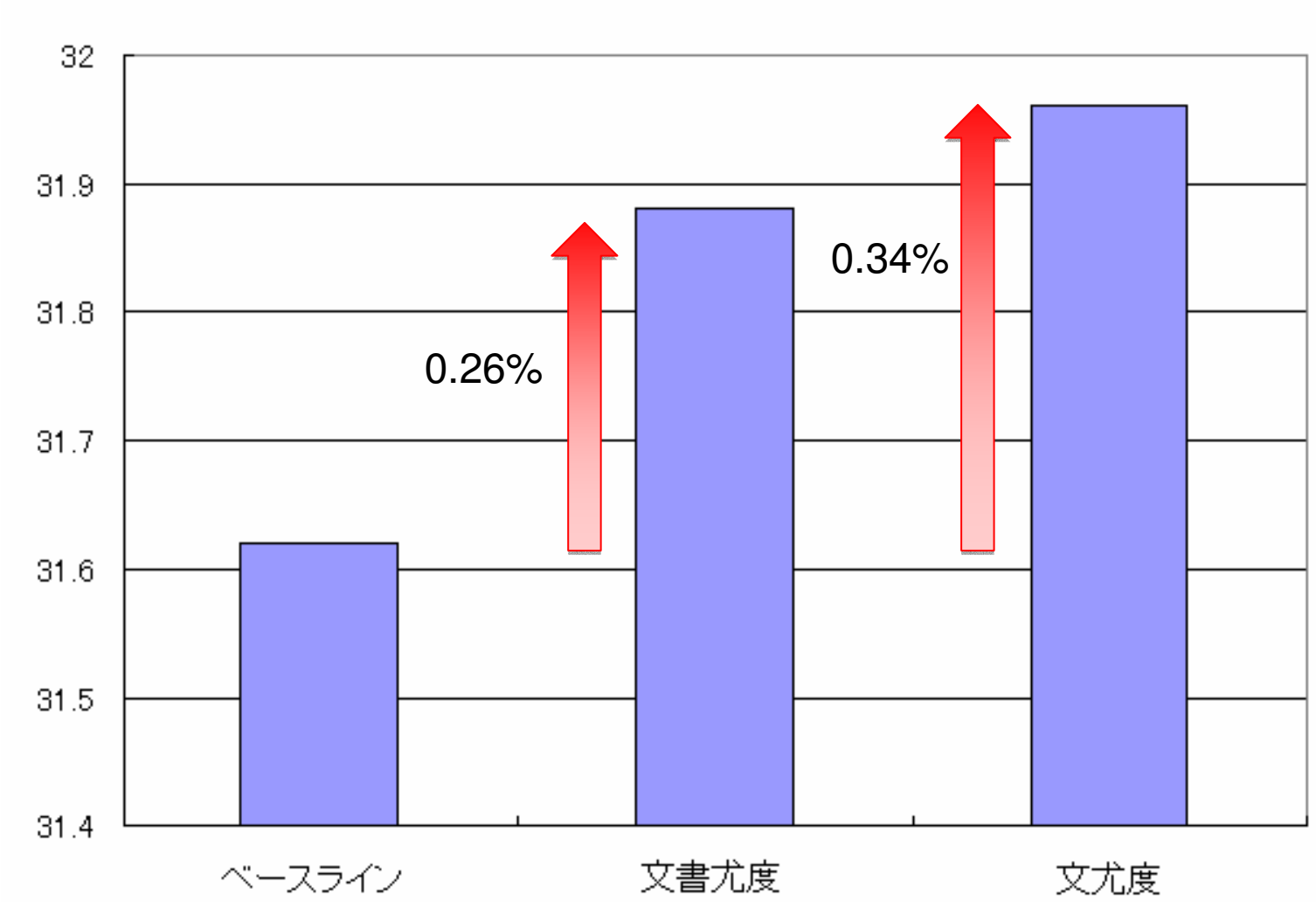
従来手法で用いられる特徴関数

$$score(s) = \lambda_{LM} L(s) + \lambda_{TM} T(s) + \lambda_D D(s) - \lambda_{WP} |s| - \lambda_{UNK} \bullet unk(s) + \lambda_{LDA} LDA(s)$$

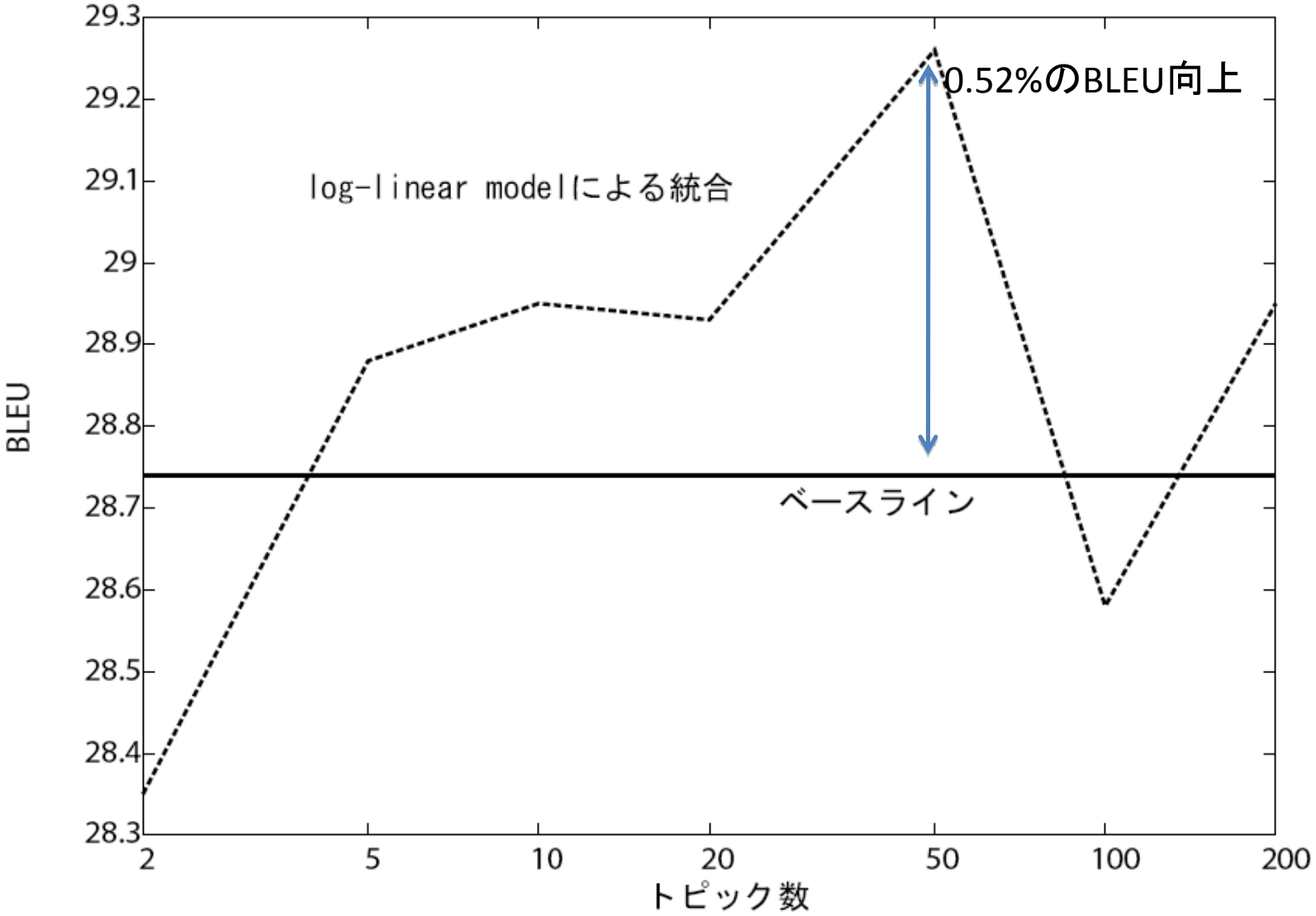
言語横断LDAモデルを追加



翻訳実験(リランキング法)
実験結果



翻訳実験(log-linear modelによる統合)
実験結果



翻訳結果(log-linear modelによる統合)
改善した例

- 英文
  - On the other hand , when the spindle motor 3 is started at one of current values i2 , i3 , and i4 , the control skips step s4 .
- 正解文
  - なお、電流値i2,i3,i4のいずれかでスピンドルモータ3が起動した場合には、ステップs4はスキップされる。
- 従来手法
  - 一方、スピンドルモータ3の一方の電流値i2,i3,i4,スキップ制御が開始されると(ステップs4)、
- 提案手法
  - 一方、スピンドルモータ3が起動されると、スキップs4で、制御電流値のi2,i3,i4である。

翻訳結果(log-linear modelによる統合)
悪化した例

- 英文
  - The reader unit 1 is further provided with an operation unit 115 for effecting various settings on the composite image input / output apparatus .
- 正解文
  - また、リーダ 部 1 には、本 複合 画像 入出力 装置 に対して 各種 設定 を 行う ため の 操作 部 115 が 設け られ ている 。
- 従来手法
  - また 、リーダ 部 1 が 設け られ た 操作 部 115 の 複合 画像 入出力 装置 の 各種 設定 を 行う 。
- 提案手法
  - また 、リーダ 部 1 の 複合 画像 入出力 装置 の 各種 設定 を 行う ため の 演算 部 115 が 設け られ ている 。

まとめと今後の課題

- まとめ
  - 言語横断LDAモデルを用いた統計的機械翻訳システム
    - リランキング法
    - log-linear model による統合
    - どちらもベースラインをわずかに上回る性能
  - 文脈適応範囲
    - 翻訳元の対応する文のみ適応で良い性能
- 今後の課題
  - ドメインが狭いとトピックが効きにくい

    →多様なドメインを含むコーパスによる実験