複数の機械翻訳を用いた言い換え認識の評価用コーパス構築に向けて

鈴木由衣 suzuki-yui@ed.tmu.ac.jp　梶原智之　小町守　首都大学東京

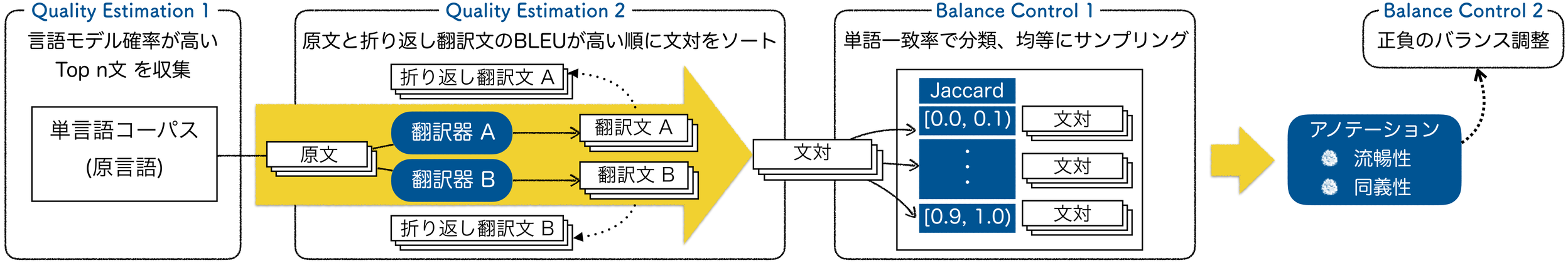
特徴

- 複数翻訳器で言い換え候補の文対を自動生成
- 非自明な言い換え文対を積極的に収集
- 日本語言い換え認識評価用コーパスを構築

背景

- 人手翻訳で言い換えを生成するのは高コスト
- 単純なヒューリスティクスでは表層的な手掛かりで解ける自明な言い換えが多く獲得される
- 日本語の言い換え認識のコーパスがない

提案手法



日本語コーパスの構築

単言語コーパス: English Wikipedia
翻訳器: Google PBMT, Google NMT
Inter Annotator Agreement : $\kappa = 0.60$
Balance Control 2: 日本語 Wikipedia から非自明な負例を追加

単語一致率 Jaccard	BC1 分類	BC1 QE2 Top n	Para.	Non-Para. (BC2)	非文	他
[0.0, 0.1)	228	200	2	1(0)	80	117
[0.1, 0.2)	2,117	200	11	14(0)	147	28
[0.2, 0.3)	14,080	200	20	9(0)	162	9
[0.3, 0.4)	51,316	200	24	15(0)	161	0
[0.4, 0.5)	100,674	200	27	16(0)	151	6
[0.5, 0.6)	134,101	200	34	16(0)	142	8
[0.6, 0.7)	100,745	200	38	13(0)	129	20
[0.7, 0.8)	55,610	200	53	52(40)	131	4
[0.8, 0.9)	26,884	200	81	83(80)	94	22
[0.9, 1.0)	8,071	200	73	73(70)	56	68
[1.0, 1.0]	6,174	0	0	0	0	0
Total	500,000	2,000	363	292(190)	1,253	282
			655			

日本語コーパスの特徴と事例

正例・負例からランダムに168文対をサンプリングして分析

パターン	%
内容語の置換*	63.1
文体の違い*	48.8
フレーズ/文の置換	25.0
機能語の置換	23.2
機能語の挿入/省略	14.3
内容語の挿入/省略	9.5
語順	6.5
片方向含意	4.2

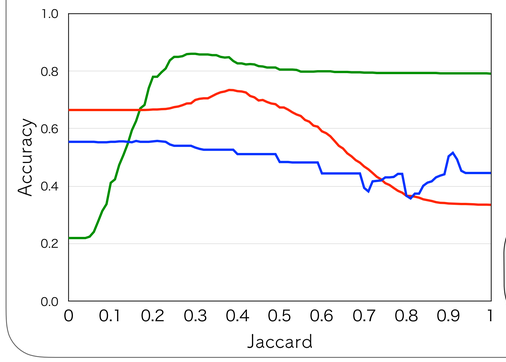
*30.2%は語種が変わる事例
[原文] *The image is clear.*
[PBMT] イメージは明確です。
[NMT] 画像はクリアです。

*文対の文体が異なる事例
PBMTは敬体
NMT は常体
を好む傾向

パターン	Jaccard	p/n		文対
内容語置換	0.60	P	PBMT	彼は共和党のメンバーでした。
			NMT	彼は共和党の一員だった。
			原文	*He was a member of the Republican Party.*
	0.90	N	PBMT	強力なローマカトリックの存在感もあります。
			NMT	強力なローマカトリックの存在もあります。
			原文	*There is also a strong Roman Catholic presence.*
フレーズ文置換	0.07	P	PBMT	めったに使われることはありません。
			NMT	まれに使われます。
			原文	*It is rarely used.*
	0.15	N	PBMT	なぜあなたは一生懸命働くのですか？
			NMT	どうしてそんなに頑張ってるの？
			原文	*Why do you work so hard?*

既存コーパスとの比較

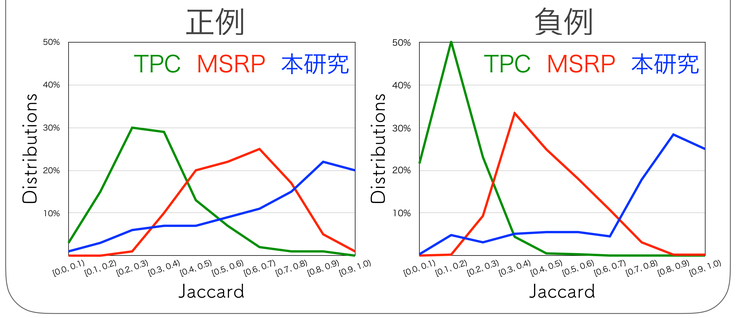
◆単語一致率の素性のみで言い換え認識したときの精度 (Accuracy)



Accuracy (max)	
TPC	0.860
MSRP	0.734
本研究	0.557

精度が低いほうが解くのが難しい言い換えコーパス

◆単語一致率ごとの事例数



本研究 : 日本語言い換え認識の評価用コーパス
MSRP : Microsoft Research Paraphrase Corpus
TPC : Twitter Paraphrase Corpus

言語処理学会第23回年次大会(NLP2017)@筑波大学　2017.03.13-17