言語資源ワークショップ2022

科学技術論文における「問題」の周辺文の問題内容の抽出

平林 照雄（東京農工大学　生物システム応用科学府）
古宮 嘉那子（東京農工大学）
浅原 正幸（国立国語研究所）

概要

(※この論文中では、「問題」とは問題という語を指し、「問題内容」「解決法」とはそれぞれ取得すべき問題内容、解決法の箇所を示すものとする)
科学技術論文内での、主題を自動で抽出するシステムの作成のため、「問題」に注目した「問題」周辺文の問題内容箇所のタグ付きコーパス及び、判定分類器の作成

「問題」の語義曖昧性解消

「問題」は少なくとも2つの意味を持つ
-“problematic”:困っていること、解決したいこと
-“task”:クイズなどのお題や、課題
科学技術論文における「問題内容」と「解決法」をとる「問題」は“problematic”のみ[1]
本論文でとりあげる、「問題」を含む文は、人手により語義曖昧性解消済みである

[1]平林照雄・河野慎司・古宮嘉那子・新納浩幸 (2021).「日本語の論文コーパスにおける「問題」の語義アノテーション」言語処理学会第 27 回年次大会, pp. 1151–1155.

「問題」周辺文に問題内容アノテーション

「問題内容」を「問題」との意味関係から3種類に分類
(a)「問題」が直接指す「問題内容」
(例)従って，原則をルールとして生成しても，有効に機能しない場合があるという問題がある.【出典】『言語処理学会論文誌 LaTeX コーパス』V12/V12N02-01.tex

(b)「問題」が直接指す「問題内容」と同じ内容で言い換え
(例)この比喩的な表現の問題を解決するには，比喩に関する人間の常識的な推論が必要である．例えば，「頭が痛い」「寒気がする」「発熱がある」など，疾患・症状が比喩的に使用される例は多くある.
【出典】『言語処理学会論文誌 LaTeX コーパス』V22/V22N05-02.tex

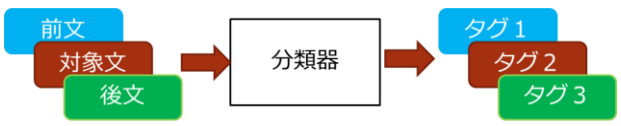
(c)「問題」が直接指す「問題内容」とは異なる内容で補足・展開
（例）この選別における問題は，選別の妥当性を確保することである．さらに，選別の対象であるがん用語の候補集合が，なるべく多くのがん用語を網羅していることを保証する必要もある.
【出典】『言語処理学会論文誌　LaTeX コーパス』V16/V16N02-01.tex

これら(a)~(c)のタグを「問題」を含む文、及びその前後文の計三文に付与
（以降この三文を**対象三文**と表記、また「問題」を含む文を**対象文**と表記）

「問題内容」が含まれる文か判定する分類器の作成

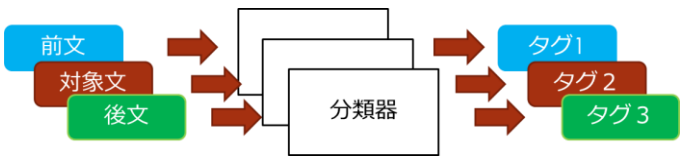
対象三文を用いて、分類器を学習
ー (i)　対象三文別々の分類器を作成するか、共通の分類器を作成するか
ー (ii)　使用するモデル　（線形のSVMかBERTによるfine-tuning）
ー (iii) (a)〜(c)のタグをどこまで正例とするか
を変化させ、8種類の実験を行う

(i)-(1) 対象三文共通の分類器を作成する時



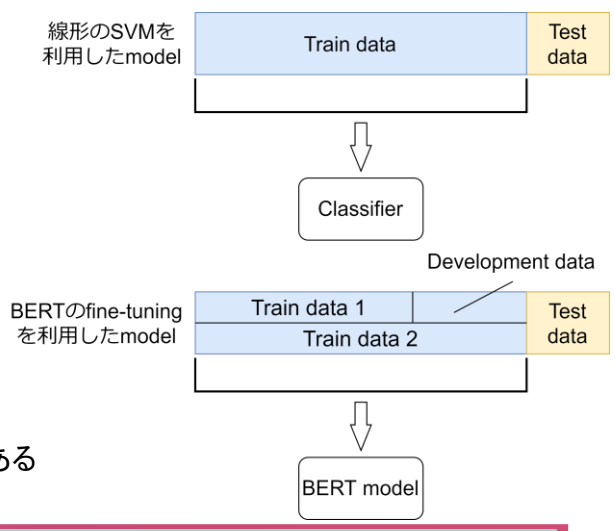
それぞれ独立した文として入力
この時(iii)の条件は(a)〜(c)すべてのタグを正例とする

(i)-(2) 対象三文別々の分類器を作成する時



対象三文の前後文には、それぞれ対象文の前後文であるという情報を入力

(ii) 使用するモデルを変える



(iii)　(a)〜(c)のタグをどこまで正例とするか

A　(a)のタグのみを正例とした時

	「問題」を含む文		前文		後文		総計
	正例	負例	正例	負例	正例	負例	
学習データ	48	31	10	69	1	78	237
開発データ	16	11	3	24	0	27	81
テストデータ	16	11	2	25	0	27	81
総数	80	53	15	118	1	132	399

B　(a)と(b)のタグを正例とした時

	「問題」を含む文		前文		後文		総計
	正例	負例	正例	負例	正例	負例	
学習データ	48	31	11	68	7	72	237
開発データ	17	10	5	22	2	25	81
テストデータ	15	12	3	24	2	25	81
総数	80	53	19	114	11	122	399

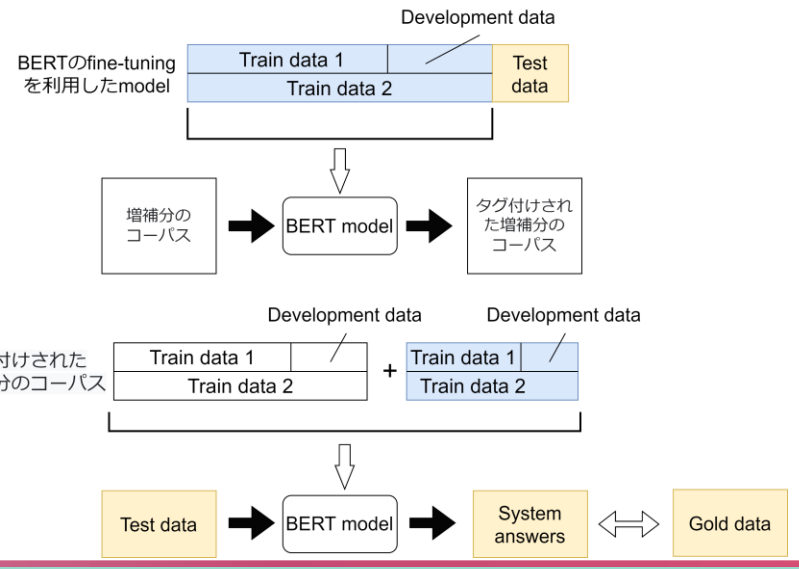
C　(a)と(b)と(c)のタグを正例とした時

	「問題」を含む文		前文		後文		総計
	正例	負例	正例	負例	正例	負例	
学習データ	45	34	15	64	10	69	237
開発データ	20	7	4	23	6	21	81
テストデータ	19	8	7	20	5	22	81
総数	84	49	26	107	21	112	399

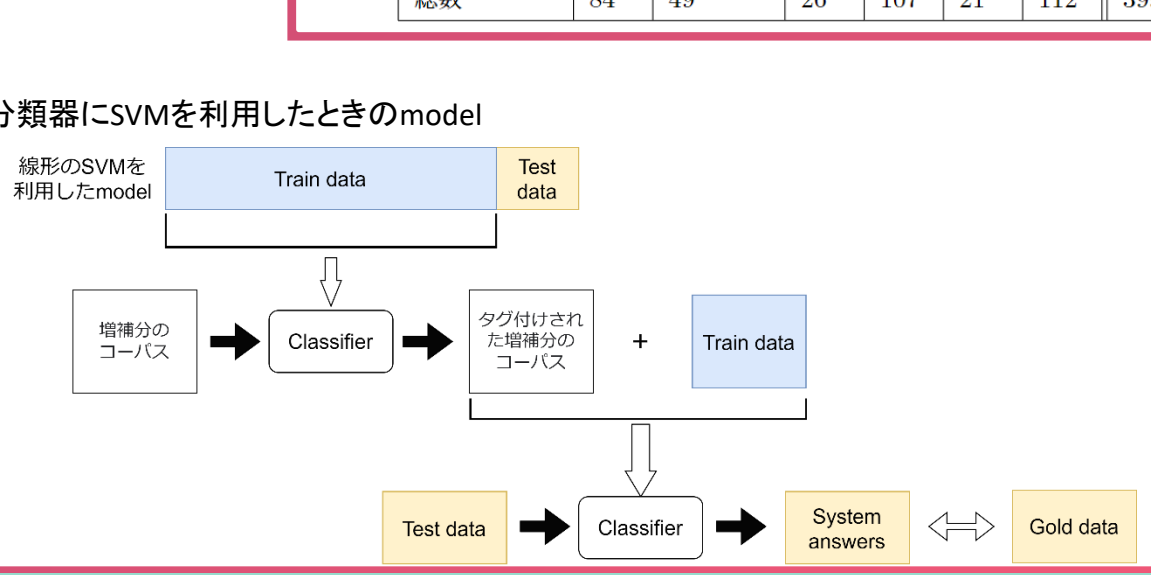
作成した分類器によるコーパスの増補と自己学習

それぞれ学習した分類器でコーパスを増補した後、増補したコーパスを加え、同条件で再学習
この時使用したモデルにより、再学習の手法が異なる

分類器にBERTのfine-tuningを利用したときのmodel



分類器にSVMを利用したときのmodel



実験設定・実験結果

SVMはScikit-learn ライブラリの linearSVC モデルをデフォルト値で使用

BERTの事前学習済みモデルとして
“cl-tohoku/bert-base-japanese-v2”を利用

分類器の入力として、事前学習済み
BERT モデルから出力された分散表現を用いる

対象三文内で異なる分類器を作成する時、
対象文から作成される分類器を「抽」、
対象文の前文から作成される分類器を「前」、
対象文の後文から作成される分類器を「後」と表記

コーパス増補前の結果を右表の左側に、
コーパス増補後の結果を右表の右側に示す

分類器の番号	(1)-SVM	(1)-BERT		(2)-SVM	(2)-BERT
正解率	0.70	0.79	A	抽 0.56	抽 0.70
				前 0.89	前 0.85
				後 1.0	後 1.0
				計 0.48	計 0.63
			B	抽 0.63	抽 0.89
				前 0.89	前 0.85
				後 0.93	後 0.89
				計 0.59	計 0.67
			C	抽 0.70	抽 0.78
				前 0.85	前 0.85
				後 0.78	後 0.78
				計 0.48	計 0.48

分類器の番号	(1)-SVM	(1)-BERT		(2)-SVM	(2)-BERT
正解率	0.72	0.69	A	抽 0.78	抽 0.67
				前 0.93	前 0.85
				後 1.0	後 1.0
				計 0.74	計 0.52
			B	抽 0.67	抽 0.56
				前 0.85	前 0.89
				後 0.85	後 0.89
				計 0.52	計 0.41
			C	抽 0.85	抽 0.70
				前 0.74	前 0.74
				後 0.85	後 0.81
				計 0.52	計 0.30