文献紹介

自然言語処理研究室
後藤大明

出典

テキスト結束性を考慮したentity gridに基づく局所的一貫性モデル,横野 光 , 奥村 学,自然言語処理,Vo17,No1,pp.1_161-1_182,(2010)

本論文ではentity gridを用いたテキストの局所的な一貫性モデルに対する改善について述べる．entity grid ベースの既存モデルに対して，テキスト結束性に寄与する要素である接続関係，参照表現，語彙的結束性，また，より詳細な構文役割の分類を組み込んだモデルを提案し，その性能を検証する．語彙的結束性に関しては，語彙的連鎖を用いたクラスタリングを行う．テキスト中の文の並びに対して，より一貫性のある文の順番の判定と，人手による評価に基づいた要約テキストの比較の 2 種類の実験を行い，その結果，本論文で提案する要素が entity grid モデルの性能の改善に寄与することが明らかになった

キーワード: テキスト一貫性, テキスト結束性, テキストの評価

はじめに

テキストの評価

ROUGU,BLEU,自動要約

内容についての評価に重点

必要な情報がどれだけ含まれているか

正しく伝わることを保証するために、一貫性の評価が必要

一貫性　文章の意味的なまとまりの良さ

はじめに

Barzilayらの研究

　entity grid：要素の遷移の傾向のみを考慮しており、明示的な特徴はほとんど利用されていない

テキスト結束性

　意味のつながり、文法的つながり

　寄与する要素　参照、接続、語彙的結束性

一貫性モデルに結束性に関わる要素を組み込むことで性能を向上

entity grid

entity grid

　　テキストを行に文、列に文章中の要素と対応させた行列

要素(構文役割)は主語（S）目的語（O）その他（X）出現せず（-）

構文役割の遷移確率と構文役割の出現確率から文書ベクトルを生成

s_1 [BELL INDUSTRIES Inc.]$_S$ increased [its quarterly]$_O$ to [10 cents]$_X$ from [seven cents]$_X$ [a share]$_X$.
s_2 [The new rate]$_S$ will be payable [Feb. 15]$_X$.
s_3 [A record date]$_S$ hasn't been set.
s_4 [Bell]$_S$, based in [Los Angeles]$_X$, makes and distributes [electronic, computer and building products]$_O$.

	BELL	quarterly	10 cents	seven cents	share	rate	Feb. 15	date	Los Angeles	products
s_1	S	O	X	X	X	-	-	-	-	-
s_2	-	-	-	-	-	S	X	-	-	-
s_3	-	-	-	-	-	-	-	S	-	-
s_4	S	-	-	-	-	-	-	-	X	O

entity grid

一貫性の評価にはスコア関数を導入

テキストdから並べ替えたテキストxikよりxijの方が一貫性があるとき

$$\mathbf{w} \cdot \Phi(x_{ij}) > \mathbf{w} \cdot \Phi(x_{ik})$$

wはranking SVMによる学習で得られる

Φは遷移確率で構成された文書ベクトル

	SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	0S	0O	0X	0-	S1	O1	X1	-1	--
d_1	0	0	0	.06	0	0	0	.02	0	0	0	.08	.06	.02	.04	.02	.02	.06	.10	.02	.01	.02	.14	.32
d_2	.02	0	0	.04	0	0	0	.02	0	0	0	.08	.04	.04	.10	.02	0	0	.18	.02	.02	.02	.14	.26

提案手法

接続関係毎の遷移確率の計算
文書ベクトルへの素性の追加
意味的な類似性に基づく文中のクラスタリング

接続関係毎の遷移確率の計算

接続関係の種類毎に文中の要素の構文役割の遷移の傾向が異なる

s_1. $[e_1]_S$ ... $[e_2]_O$ ...
s_2. **そして，** $[e_2]_S$ ... $[e_3]_O$ ...
s_3. $[e_3]_S$ ... $[e_1]_X$ ...
s_4. $[e_4]_S$ ...

関係	説明	例
順接型 (35)	前文の内容を条件とするその帰結を後文に述べる型	従って
逆接型 (33)	前文の内容に反する内容を後文に述べる型	しかし
添加型 (41)	前文の内容に付け加わる内容を後文に述べる型	そして
対比型 (17)	前文の内容に対して対比的な内容を後文に述べる型	または
転換型 (19)	前文の内容から転じて，別個の内容を後文に述べる型	さて
同列型 (11)	前文の内容と同等と見なされる内容を後文に重ねて述べる型	つまり
補足型 (1)	前文の内容を補足する内容を後文に述べる型	なぜなら
連鎖型 (1)	前文の内容に直接結びつく内容を後文に述べる型	そうして

接続関係毎の遷移確率の計算

種類の分類を文脈形成の関係に基づく３グループにまとめる
各グループ毎に構文役割の遷移確率を求め、ベクトルにする

関係	対応する接続関係	説明
Group 1 （論理的結合関係）	順接型，逆接型	二つの事柄を論理的に結びつけて述べる関係
Group 2 （多角的連続関係）	添加型，対比型，転換型	二つ（以上）の事柄を別々に述べる関係
Group 3 （拡充的合成関係）	同列型，補足型，連鎖型	一つの事柄に対して拡充して述べる関係

SS_{G_1}	$\cdots$	--_{G_1}	SS_{G_2}	$\cdots$	S-_{G_2}	$\cdots$	--_{G_2}	SS_{G_3}	$\cdots$	S-_{G_3}	$\cdots$	--_{G_3}
0	$\cdots$	0	0	$\cdots$	.25	$\cdots$	.25	0	$\cdots$	.06	$\cdots$	.13

参照表現

「あの」、「この」が出現した文の全文に専行詞がなければ２文間のつながりは悪いと考える

指示形容詞を含む参照表現が正しく機能している割合をベクトルの素性として追加

語彙的結束性に基づいた文中の要素クラスタリング

既存モデルでは要素間の関係は反映されていない

１文の各要素を意味的なクラスタリングによってまとめ、得られたクラスタを１つの要素として扱う

語彙体系体系の意味体系を利用した手法と語彙的連鎖を利用した手法を利用

日本語語彙体系を利用したクラスタリング

特定の階層で同じ意味属性を有する要素を同じ要素にクラスタリング

要素が複数の意味属性を持つとき、所属する数の多かった意味属性をその要素の意味属性とする

語彙的連鎖を利用したクラスタリング

Mochizukiらの手法に基づいて語彙的連鎖を求める

共起スコアはコサイン尺度
クラスタ間Ci,Cjの類似度　X,Y 単語

$$sim(C_i, C_j) = \max_{X,Y} cos(X \in C_i, Y \in C_j)$$

文の要素それぞれをクラスタとし、しきい値以上のクラスタをマージ

語彙的結束性に基づいた文中の要素クラスタリング

１つのクラスタに複数の構文役割が割り当てられる

手法1(1st)：構文役割の優先順位に基づいて決定する
手法2(comb)：クラスタ中の役割すべてを利用し、遷移は組み合わせを考える

・元のentity grid

	e_1	e_2	e_3
s_1	s	o	-
s_2	-	o	s

(s-:1,oo:1,-s:1)

・手法1(1st)

	e_1	c_1
s_1	s	o
s_2	-	s

(s-:1,os:1)

・手法2(comb)

	e_1	c_1
s_1	s	o-
s_2	-	os

(s-:1,oo:1,os:1,-o:1,-s:1)

構文役割の拡張

助詞に注目して構文役割を決定

既存モデルにはなかった主題と述部要素を追加

役割間の優先順位関係　H>S>O>R>X

構文役割	該当する助詞
主語 (S)	“が”
目的語 (O)	“を”，“に”
主題 (H)	“は”
述部要素 (R)	述部に係る他の助詞
その他 (X)	上記以外の助詞
出現せず (-)	文中に出現しない

実験　テキストの並び替え

オリジナルのテキストと文の順番を並べ替えたテキストを比較

朝日新聞コーパス2003年の記事から１記事あたり１０文以上で構成されているものから２０個の並び替えテキストを生成

10分割交差検定を行いオリジナルのテキストの方が一貫性があると判定されたペアの割合で評価

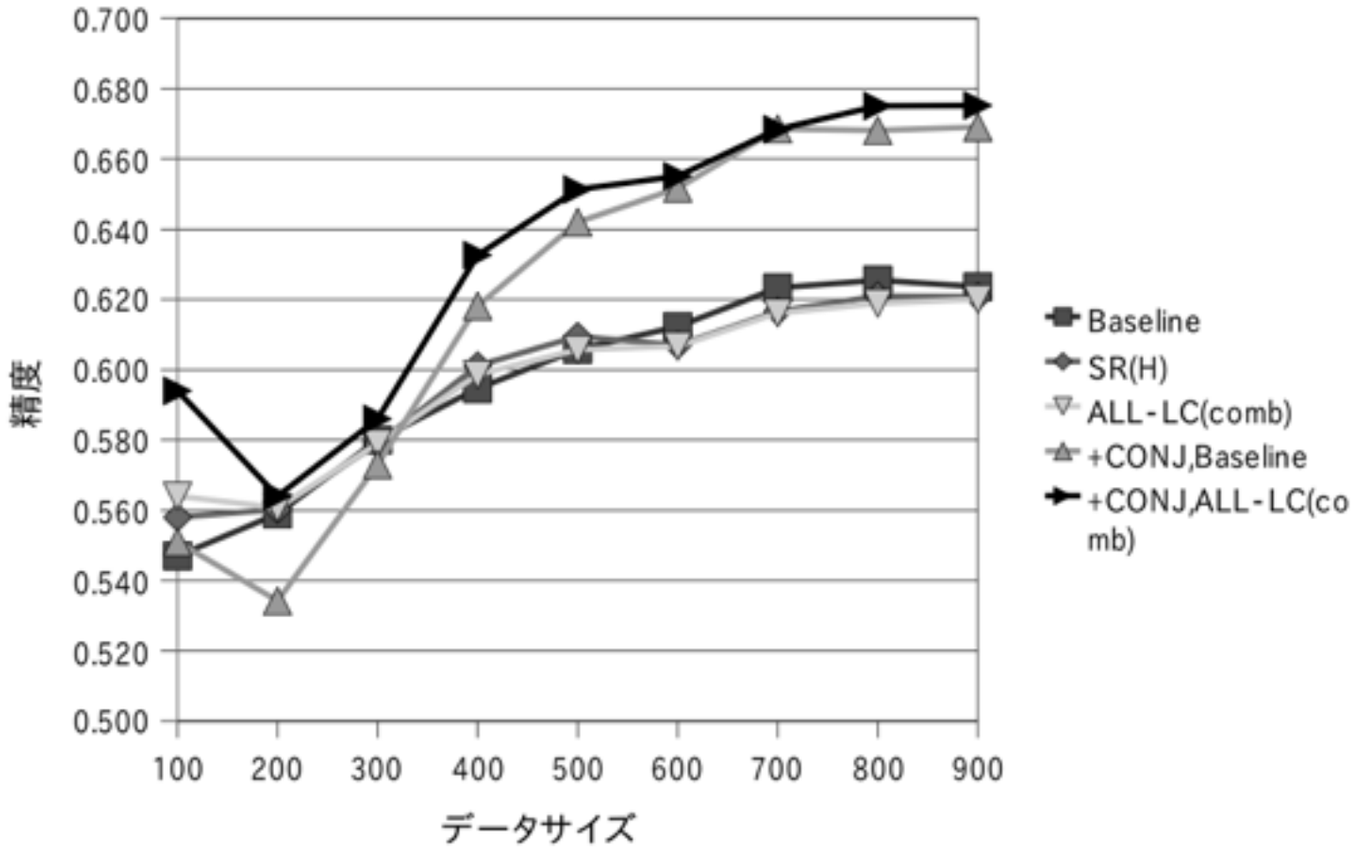
結果

モデル	説明
Baseline	(Barzilay and Lapata 2008) モデル
CONJ のみ	接続関係毎の遷移確率の計算
REF	指示表現に関する素性の追加
SC(1st)	日本語語彙大系によるクラスタリング (1st)
SC(comb)	日本語語彙大系によるクラスタリング (comb)
LC(1st)	語彙的連鎖によるクラスタリング (1st)
LC(comb)	語彙的連鎖によるクラスタリング (comb)
SR(H)	構文役割の拡張
ALL-SC(1st)	SR(H)+REF+SC(1st)
ALL-SC(comb)	SR(H)+REF+SC(all)
ALL-LC(1st)	SR(H)+REF+LC(1st)
ALL-LC(comb)	SR(H)+REF+LC(comb)

	mix(100)		mix(300)	
	no CONJ	+CONJ	no CONJ	+CONJ
Baseline	0.547	–	0.580	–
CONJ のみ	–	0.551	–	*0.573*
REF	*0.546*	0.550	*0.579*	*0.573*
LC(1st)	0.567	0.582*	0.585	0.677**
LC(comb)	0.549	0.557	**0.614**	**0.773****
SR(H)	0.558	0.562	*0.579*	*0.560***
ALL-LC(1st)	**0.574**	0.577*	*0.574*	*0.563*
ALL-LC(comb)	0.564	**0.594****	*0.579*	0.586

結果

モデル	100	200	300	400	500	600	700	800	900
Baseline	0.547	0.559	0.580	0.595	0.606	0.612	0.623	0.626	0.624
SR(H)	0.558	0.560	*0.579*	0.601	0.610	*0.607*	*0.617*	*0.621*	*0.621*
ALL-LC(comb)	0.564	0.561	*0.579*	0.599	0.606	*0.607*	*0.616*	*0.619*	*0.620*
+CONJ, Baseline	0.551	*0.534***	*0.573*	0.618**	0.642**	0.652**	**0.669****	0.668**	0.669**
+CONJ, ALL-LC(comb)	**0.594****	**0.564**	**0.586**	**0.633****	**0.651****	**0.655****	0.668**	**0.675****	**0.675****



実験　要約文書の比較

２種類の自動要約により生成された要約を比較し、どちらが一貫性があるかを判定

NTCIR-4のサブタスクTSC3に提出された要約657件
要約の読みやすさの評価結果が付与されている

モデル		all (6434)	0~0.5 (2986)	0.5~1.0 (2014)	1.0~1.5 (930)	1.5~2.0 (334)
no CONJ	Baseline	0.502	0.500	0.502	0.505	0.515
	ALL-LC(comb)	**0.589****	**0.532***	**0.592****	**0.669****	**0.775****
+CONJ	CONJ のみ	0.507	0.508	0.502	*0.503*	0.536
	ALL-LC(comb)	0.546**	0.510	0.547**	0.604**	0.689**

まとめ

日本語に特化した構文役割をentity gridモデルに組み込むことで、結束性を考慮した一貫性モデルを提案し、その有用性を検証した

精度はまだまだ高いとはいえないので改良の余地がある