自 然 言 語 処 理 106－6
（1995．3．9）

ビジネス文書作成問題における誤り抽出方法

安藤智　澤邉一秀　松岡誠　上田弓子　重永信一
松下電器産業（株）　研究本部

我々は、ビジネス文書の作成問題における個別添削支援を目的とした教育支援システムCEN(Composition Exercise tutoring system on computer Networks)の開発を行っている。文書作成問題に対する教育支援システムにおいては、学習者の解答に含まれる様々な誤りを効率的にかつ高精度に抽出することが重要である。そのためには、解答と正解を効率的に照合し、精度よく対応付ける必要がある。ここでは、文書の文字表現とレイアウト情報を効果的に使用して、解答文書と正解文書との間での照合および対応付けを行い、解答文書中の誤りを抽出する方法について報告する。

A Method of Detecting Errors on Business Letter Composition

Satoshi Ando　Kazuhide Sawabe　Makoto Matsuoka　Yumiko Ueda　Shin-ichi Shigenaga
Corporate Research Division, Matsushita Electric Industrial Co., Ltd.

This paper describes a method of detecting errors on CEN(Composition Exercise tutoring system on computer Networks), which we are developing. In CAI system for composition, It is a most significant problem to detect errors efficiently and correctly in a student's answer. It is necessary to find efficiently in advance the part of the right answer corresponding to the part of the student's answer. We propose a method of detecting errors in a student's answer by deciding correspondence of a student's answer with a right anser by evaluating similarity..

1　はじめに

文書作成問題は、採点支援や課題演習の添削支援など、計算機支援が強く望まれている領域である。我々は、この要望に応えるために、ビジネス文書の文書作成問題を対象に個別添削支援を目的とした教育支援システムCEN(Composition Exercise tutoring system on computer Networks)の開発を行っている[1]。

計算機によって文書内容を抽出する試みとしては、特定の文書を対象として、文字表現とレイアウトに関する情報を用いて文書内容を抽出する試みがなされている[2][3]。

文書作成問題を対象とした教育支援システムにおいては、文書内容の抽出に加えて、誤り箇所を抽出しなければならない。すなわち、文字表現やレイアウトの誤り、無効な文や語句を識別しながら解答の文書内容を抽出することが必要である。

本稿では、文字表現やレイアウト情報を効果的に用いた、CENにおける解答文書の誤り抽出方法について報告する。

2　CEN の概要

2.1　目的

CENの目的は、文書作成問題において、個々の学習者の解答文書の誤り箇所を適切に抽出して、その誤りに対して一貫した添削指導を行い、対象文書の書法を教授することである。

2.2　処理概要

CEN では、以下の手順によって、個別添削を行う。

(1)解答文書を文の区切り記号(空白、改行、句読点など)を手がかりに照合の対象となる文字列(照合対象項目)を切り出す。

(2)各照合対象項目を、正解文書の文書項目(正解文書を構成する要素で添削の対象として意味のある語句あるいは複数の語句の集まり)と照合し、対応付けを行う。

(3)正解文書項目と対応付けられた照合対象項目について、文字表現やレイアウトを検査し、誤りの抽出を行なう。

(4)各誤り箇所を画面上で指摘し、学習者の要求に応じて、段階的に詳細な添削指導を行う。

2.3　システムの構成

図1に、CENの構成を示す。

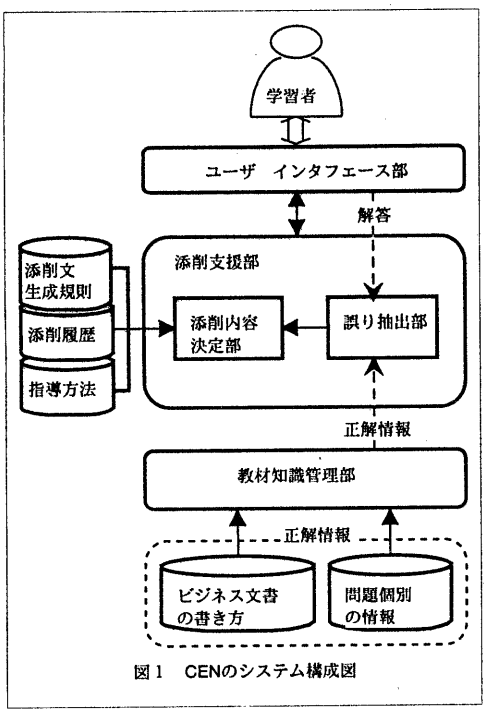

図1　CENのシステム構成図

(1)ユーザインタフェース部

学習者とのやりとりの制御(問題の提示・解答や添削要求の受け付け・添削メッセージの提示など)を行なう。

(2)添削支援部

CEN の中核となるモジュールであり、以下のサブモジュールからなる。

(a)誤り抽出部

学習者の解答を正解情報と照合し、誤り箇所を抽出する。

(b)添削内容決定部

現在の誤り状況や過去の添削履歴を用いて指導方針を決定し、その指導方針に従って解答文書上で誤り箇所を指摘する。指摘した誤りに対する学習者からの添削要求に応えて添削メッセージを提示する。

(3)教材知識管理部

個々の問題には依存しないビジネス文書のレイアウトのルールや個々の問題に対する正解の文字表現などからなる正解情報を格納・管理し、誤り抽出部にそれらの情報を提供する。

正解情報の内容を図2に示す。

(a)　レイアウトに関するルール
- 文書項目の行属性
  - 右寄せ、左寄せ、センタリング、次の行へ継続する行の種別
- 位置関係
  - 文書内の絶対的位置
  - 文書項目間の相対的位置

(b)　文字表現に関する情報
- 文書項目間の階層構造
- 文書項目における正解表現、制約条件、および誤り表現

図2　正解情報

3　誤り抽出の方法

3.1　誤り抽出における課題

CENでは自由記述の解答を扱うため、多様な誤りを含む解答文書から、いかに効率的にかつ精度よく誤りを抽出するかが課題である。そのためには、学習者の解答と正解とを効率的に照合して、学習者の解答と正解との対応付けを精度よく決定することが必要である。

3.2　誤り抽出の手順

CENの誤り抽出の手順は、次のとおりである。

(1)照合に必要な正解情報の取り込み

(2)解答文書の照合対象項目への分割

(3)照合対象項目と正解の文書項目との対応付け

(4)誤りの判定

3.2.1　正解情報の取り込み

教材知識管理部から問題に対応する正解情報の取り込みを行う。

3.2.2　照合対象項目への分割

次の手順に従って、解答文書を、照合の対象となる項目(照合対象項目)に分割する。

(1)区切り記号(空白、改行、句読点など)によって解答文書を分割し、照合対象項目として切り出す。

(2)照合対象項目の行内の位置(右寄せ、左寄せ、中央)、文書内での絶対的位置、次行に継続すべき語句があるかどうかなどを調べる。

(3)複数行に跨っていると判断される場合は、適宜連結して一つの照合対象項目として扱う。

3.2.3 正解の文書項目との対応付け

文書の文字表現やレイアウト情報を使って、照合対象項目と正解文書の文書項目(正解文書項目)を照合し、対応付けを行う。

3.2.4 誤りの判定

各照合対象項目に対して次の検査を行い、誤りを判定する。

(1)文字表現の検査

対応する正解文書項目の文字表現と比較し、異なっていれば文字表現の誤りとする。

(2)レイアウトの検査

対応する正解文書項目に関するレイアウトのルールを満たすかどうかを検査し、満たさなければレイアウトの誤りとする。

(3)制約条件チェック

対応する正解文書項目に関する制約関数を評価して制約条件を検査する。制約条件を満たしていなければ、制約に関する誤りとする。

4 照合の方法

照合および対応付けを行う上で、利用できる正解文書項目の情報としては次のものがある。

(1)文字表現に関する情報

(a)正解文書項目中に含まれる単語

(b)正解文書項目に含まれる単語の連接関係

(2)レイアウトに関する情報

(a)文書内の絶対的位置

(b)行内での相対的位置

(c)他の文書項目との相対的位置関係

4.1 実験モデルの照合方法

実験モデルでは、処理手順の単純さを重視し、照合対象項目を、正解文書の各文書項目の構成単語の任意組合せを網羅する正規表現と再帰的に照合していく方法(再帰的縦型照合方法)を試みた[1]。

4.1.1 再帰的縦型照合方法の概要

[1]では正解木の親ノードから子ノードに降下しながら照合を繰り返すことにより解答文書の木構造を構築している。

正解木の各ノードは、正解表現、子ノードの正解表現から組み上げられた正規表現(正解正規表現)および正解表現に対応する誤りを含む拡張正規表現を持っている。

照合は、正解正規表現、拡張正規表現の順で適用され、正解正規表現で一致した文字列部分に対しては拡張正規表現を適用しない。照合対象項目は、最長一致する正解正規表現または拡張正規表現を持つノードが示す正解文書項目と対応付けられる。

4.1.2 再帰的縦型照合の問題点

再帰的縦型照合では、次のような問題があった。

(1)正解の文字表現の順序関係を考慮していないために、解答のノイズを拾ってしまう。

(2)再帰的に解答木を構築していく過程で他のノードとの情報のやり取りを行っていないために、ノードの構成に矛盾が生じる。

特に、(2)の場合には、一貫した添削メッセージを学習者に提示することができない。図3に、「１３時から１６：００」という解答文書の表現から構築された解答木を、図4には、それに対応する正解木の部分を示す。正解木では「時間」の子ノードが二つあり、いづれか一方が正解である。例にあげた解答文に再帰的縦型照合を行うと、解答木では、「時間」の子ノードの両方が構築されてしまう。このとき、「時間表記１」からは”「１６時」と「まで」に対する表現が不足している”という誤りが、「時間表記２」からは”「１３：００」と「～」に対する表現が不足している”という誤りが抽出される。

4.2 非再帰横型照合方法

前節であげた問題点は、照合対象項目の文字表現を全体として評価していなかったことに起因する。そこで、照合対象項目と正解文書項目との照合時に、文字表現とレイアウトの類似度計算により、文字表現を全体として評価し、対応する正解の文書項目を決定する方法(非再帰横型照合方法)を考案した。

非再帰横型照合での処理は、次のとおりである。

(1)文字表現の類似度による対応付け

(2)レイアウトの類似度による対応付け

(3)文字表現・レイアウトによる対応付けの統合・見直し

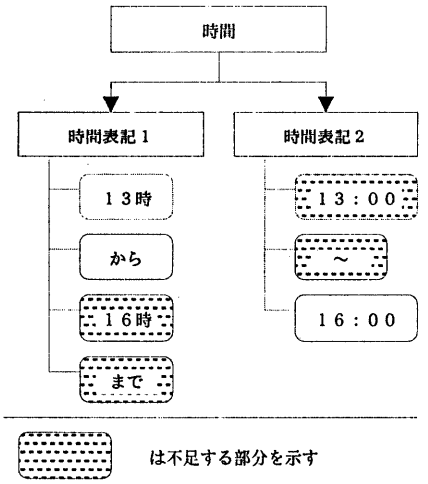

図３　再起的縦型照合による解答木の例

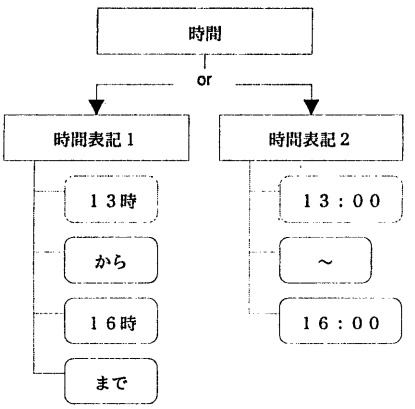

図４　再起的縦型照合の正解木の例

4.2.1　類似度の測定方法

類似度を測る対象において、それを特徴付ける要素の集合をそれぞれR、Sとした時、RとSとの類似度Sim(R,S)はTverskyのcontract model を用いると図５のように定義される[4]。α、β、γは重み係数(α,β,γ>0)であり、fは顕著性を表す関数である。

$$Sim(R,S) \equiv \alpha * f(R \cap S) - \beta * f(R-S) - \gamma * f(R-S)$$

図５　類似度の定義

文字表現およびレイアウトに対する類似度を図５の式にならってそれぞれ$Sim_C(R,S)$、$Sim_L(R,S)$と定義する。$Sim_C(R,S)$と$Sim_L(R,S)$を合わせて評価した類似度$Sim_T(R,S)$を図６のように定義する。関数Fは$Sim_C(R,S)$と$Sim_L(R,S)$を統合して評価する関数である。

$$Sim_T(R,S) \equiv F(Sim_C(R,S), Sim_L(R,S))$$

図６　統合された類似度の定義

4.2.2　文字表現による照合

文字表現を特徴付ける要素としては、各文字表現に含まれる単語とその連接関係などを用いることができる。

矛盾する解答木を構築しないように、対応付けの方法を次のように改める。

(1)正解表現の構成単語間の連接関係を、対応付けの評価に組み込む。

(2)照合対象項目全体を単語単位に分解し、単語レベルで照合対象項目と正解文書項目の文字表現の対応付けを行う。

類似度を測定するために用いる単語とその連接表は、正解表現から作成する。図７に正解表現の例、図８にその正解表現から作成される連接表を示す。

……/……/……

文書項目「件名」の正解表現

販売/促進/会議/開催/に/ついて/（/通知/）

文書項目「本文」の正解表現その１～４

標記/の/件/下記/の/とおり/開催/します/の/で/業務/繰り/合わせ/の/うえ/ご/出席/ください/。

標記/の/件/下記/の/とおり/開催/します/の/で/業務/繰り/合わせ/の/うえ/お/集まり/ください/。

販売/促進/会議/を/下記/の/とおり//開催/します/の/で/業務/繰り/合わせ/の/うえ/ご/出席/ください/。

販売/促進/会議/を/下記/の/とおり/開催/します/の/で/業務/繰り/合わせ/の/うえ/お/集まり/ください/。

・文書項目「……」の正解表現

図７　正解表現の例

連接表のエントリー	連接する単語
.....	.....
販売	促進
促進	会議
会議	開催、を
開催	に、します
に	ついて
ついて	(
(	通知
通知	)
)	(TERMINAL)
標記	の
の	件、で、とおり、うえ
件	下記
下記	の
とおり	開催
します	の
で	業務
業務	繰り
繰り	合わせ
合わせ	の
うえ	ご、お
ご	出席
出席	ください
ください	.
.	(TERMINAL)
お	集まり
集まり	ください
を	下記
.....	.....

図8　正解表現に現れる単語の連接表

単語の分割は、以下に示す方針に従って、チャート式の形態素解析と同様な処理を行う。

(1)連接表のエントリーにあり、かつ先行する単語に連接する単語であれば分割を行う。

(2) (1)で分割できなかった場合に限って、連接表にエントリーがある単語と一致すれば分割する。

(3) さらに(2)で分割できなった場合は未知語として分割する。

例として「標記の販売促進会議に御出席ください。」を、図8に示した部分の連接表を用いて単語分解した結果を図9に示す。

標記／の／販売／促進／会議／に／ご／出席／下さい／。

二重下線部	連接関係まで満たす単語列
下線部	連接関係を満たさない単語
下線部なし	未知語

図9　単語分割の例

以下に、文字表現の類似度を計算する一方法をあげる。

図5の式におけるRとして正解文書項目に含まれる単語の集合を、Sとしては照合対象項目に含まれる単語の集合を、顕著性を表す関数fとしては集合の要素数を返す関数を用いる。このときαはRとSの共有要素数に対する重み係数となり、βとγはそれぞれRとSの固有の要素に対する重み係数となる。仮にα=2、β=γ=1とすれば、例にあげた照合対象項目と図7に示した正解文書項目との類似度は図10のようになる。

正解表現	件名	本文			
		その1	その2	その3	その4
f(R∩S)	4	4	3	7	6
f(R-S)	5	15	16	12	13
f(S-R)	6	6	7	3	4
Sim(R,S)	-3	-13	-17	-5	-5

図10　文字表現の類似度テーブル

照合対象項目と正解文書項目の文字表現に関して図5の式を用いて類似度を測り、図11に示すような文字表現の類似度テーブルを作成する。

正解表現	文書項目1	…	文書項目n
照合対象項目1	C11	…	C1n
照合対象項目2	C21	…	C2n
.....	C31	…	C3n
.....	…	…	…
照合対象項目m	Cm1	…	Cmn

図11　文字表現の類似度テーブル

類似度テーブルの中より類似度の合計が最大となる組合せを取り出し、それを照合対象項目と文書項目間の文字表現による対応付けとする。

4.2.3　レイアウトによる照合

レイアウトを特徴付ける要素としては行属性と文書内の絶対的位置に関する情報などが用いることができる。文書内の相対的位置関係は他の項目との関連があるため、ここでは用いず、対応関係の見直しの時点で用いる。

レイアウトの類似度を測る一方法としては、照合対象項目と正解文書項目の文書内での正規化された位置(それぞれの文書の大きさで正規化を行った位置)を用いる。文書を縦と横方向にN分割し、

照合対象項目と正解文書項目がN×N分割された領域のどこに位置しているかを測る。複数箇所にまたがっている場合は、その全てを挙げる。

図５の式におけるRとSとして、照合対象項目と正解文書項目の存在する領域の集合を用いる。顕著性を表す関数fとして集合の要素数を返す関数、重み係数はα=2、$\beta=\gamma$=1としてレイアウトの類似度を測る。

文字表現の場合と同様に、照合対象項目と正解文書項目のレイアウトに関して類似度を測り、図１２に示すような類似度テーブルを作成する。

正解表現	文書項目1	…	文書項目n
照合対象項目1	L11	…	L1n
照合対象項目2	L21	…	L2n
……	L31	…	L3n
……	…	…	…
照合対象項目m	Lm1	…	Lmn

図１２　レイアウトの類似度テーブル

4.2.4　照合結果の統合と見直し

文字表現とレイアウトの類似度テーブルの両方を統合して、対応付けを決定する。

統合した対応付けに対して、正解文書項目間の相対的な位置関係を用いて次のような見直し処理を行う。

(1)グループ化

統合した対応付けの結果に対してレイアウトの相対的位置関係の情報を用いてグループ化を行う。

(2)再配置の候補の抽出

対応付けの見直しを行う候補を抽出する。見直しの候補となる条件は次の二つである。

(a)(1)でグループ化された照合対象項目の中で相対的位置関係を満たさないもの

(b) 切離し可能な文字表現部分

(3)置き換え可能な文書項目の探索

(2)の見直しの候補に対してレイアウト上対応付けが可能な文書項目を捜す。複数ある場合は文字表現による類似度のにより候補を決定する。

5　おわりに

解答と正解の文字表現の照合方法を改良することによって、実験モデルにおける問題を解決して、効果的に文字表現やレイアウト情報を用いた、誤り抽出方法を示した。

本システムは、学習者と実時間で応答し添削指導することを目指している。現時点では、照合における類似度測定に時間がかかりすぎている。今後はシステムの応答性の向上を図るために、類似度の計算方法や、対応関係の組合せの絞り込みについてさらに改善していく。また、類似度の測定方法で用いる関数や重み係数についても実際に学生による使用実験を通して改善していく予定である。

謝辞

実験モデルの評価に協力いただいた(学)麻生電子ビジネス専門学校の秘書系学科の先生方に感謝します。

参考文献

[1] 重永他,ビジネス文書作文問題を対象とした教育支援システムの試作,人工知能学会研究会資料,SIG-J-9401-6

[2] 石田他,文書理解システムの試作,自然言語処理シンポジウム,1984

[3]黄瀬他,レイアウトとコンテントの知識を用いた仮説駆動型文書画像理解,情報処理学会論文誌 Vol34 No.8 pp1716～1730,1993

[4]Amos Tversky, Features of Similarity, Psychological ReLiew, Vol.84(1977)