語の上位下位概念と文書の主題との関連度を用いた例示部分特定手法

Identifying Example Segments based on Relations between Hyponymy and Themes in a Document

大野 和久[1*]　田村 直之[1]　松本 征二[1]　新堀 英二[1]
Kazuhisa Oono[1]　Naoyuki Tamura[1]　Seiji Matsumoto[1]　Eiji Shinbori

[1] 大日本印刷株式会社 honto ビジネス本部
[1] Dai Nippon Printing Co., Ltd. honto Business Operations

Abstract: Recognizing logical paragraphs and relations of points in documents helps us to comprehend the documents. The logical paragraphs contain various segments such as "elaboration", "contrast" and "example". The authors often write their insistence using abstract terms. Therefore, they express their insistence in concrete cases with the examples, and accelerate to comprehend the documents. In this paper, we identify the example segments based on relations between hyponymy and themes in a document. We consider that sentences which contain concrete terms are divided into two types. The first type expresses themes, and the second type expresses examples. We calculate a rate of theme terms in a sentence, and capture whether the sentence expresses the themes or not. Thus, we find out that the sentence is likely to express the example if the sentence is not likely to express the theme. In our experimental evaluation, we confirmed that our proposed method scored better recall and F-measure than the baseline method.

1　はじめに

文書の読解には，文書を意味段落ごとに区切り，各意味段落での要点どうしを関係づけることが有効と考えられている [1][2]．その理由は，要点の関係性の認識について，認知的負荷が軽減され，情報の整理や文章構造を理解しやすくなるためである．

文書内容に含まれる意味段落のうち，読解を支援する要素の一つとして，例示がある．例示は，一般化された著者の主張を具体化する表現である．著者の主張は抽象的な表現で記述される場合があり，その表現だけでは，読者は著者の主張を理解しにくい場合がある．そこで，読者は，例示を用いた具体的表現を併せて読むことにより，既に持っている知識を想起し，その想起されたイメージと著者の主張を結びつけることができ，著者の主張を理解しやすくなると考えられている [3]．そのため，文書の読解には，著者の主張だけでなく，その主張を説明するための例示も把握することが有効となる．

読解を支援するための文書構造解析技術として，談話構造解析がある．談話構造解析では，意味段落や文章間の関係性を明らかにする．このことにより，話題の推移をとらえたり，各段落および文章の役割を認識することが可能となる．

談話構造解析の既存研究では，例示特定のために，手がかり語を用いて特定する手法や，語の抽象度の遷移を用いて特定する手法がある．手がかり語を用いる手法では，文章中に "例えば" や "〜の「ように」" が含まれる場合，その文章を例示として特定する [4]．また，語の抽象度の遷移を用いる手法では，複数の文章のまとまりがあり，ある文章 A で述べられた事象や状態の具体例が，文章 A に続く文章 B で提示される場合，文章 B を例示として特定する [5]．

ただし，これらの既存手法では，手がかり語が存在しない場合や，例示部分の周りに抽象表現が存在せず，例示部分が独立して出現する場合は，例示部分の特定が困難である．

そこで本研究では，これらの既存手法で特定できなかった例示部分の特定を可能にすることを狙い，語の上位下位概念と文書の主題との関連度を用いて例示部分を特定する手法を提案する．提案手法では，具体的な表現を含む文章が例示になりやすいことに着目した上で，その具体的な表現と主題との関係性を考慮することにより，例示特定についての再現率向上を行うと

*連絡先：大日本印刷株式会社 honto ビジネス本部
〒 162-8001 東京都新宿区市谷加賀町 1-1-1
E-mail: Oono-K6@mail.dnp.co.jp

同時に，精度向上も行う．

ここで本研究では，文書は，章もしくは節全体を示し，文章は，句点で区切られた一つの文を示す．

2 提案手法で特定する例のパターン

本研究では，例示を交えながら著者の主張を論述する文書を対象とする．例えば，評論文，論説文，随筆といった文書である．

これらの文書では，一般的な表現と，具体的な表現を繰り返すことにより，著者の主張を論述している [6]．ここで，一般的な表現は著者の主張を表し，具体的な表現は，読者に対して著者の主張を納得させるための証拠を表す．この具体的な表現が，主張に対する例示となる．このように，文章内容は，二つの表現に分けられる．

具体的な表現である例示は，手がかり語が存在する場合と存在しない場合に分けられ，合わせて 5 種類のパターンがある．

手がかり語が存在する場合については，2 種類のパターンがある．

1. 例示を示す接続詞
   “例えば” といった接続詞を用いて，例示を述べる場合である．

2. 具体例の列挙を示す語
   具体例を挙げる際に，“〜の「ように」”，“「ある」書籍では”，“〜「など」” といった語を用いる場合である．

一方，手がかり語が存在しない場合では，3 種類のパターンがある．

3. 主題と異なる分野での事象を記述
   文書の主題とは異なる分野での事象を用いて，例示を述べる場合である．例えば，ジャーナリズム論の文書の中で，物理学の理論や美術家の行動を用いて例を述べる場合である．なお，比喩法として用いられる隠喩についても，この場合に含まれると考える．

4. 上位語から下位語へ遷移
   複数の文章があるときに，まず抽象的な表現を述べ，次に，抽象的表現の下位語にあたる語を用いて，例示を述べる場合である．例えば，抽象的表現として，「マスメディアは様々な情報を発信している.」といった文章があり，その文章の後に，「新聞は紙を媒体とし，日々のニュースを発信している.」といった文章がある場合を考える．このとき，マスメディアという上位語から，新聞という下位語へ話題が遷移している．そこで，新聞に関する文章は，マスメディアに関する文章の例示として述べていることになる．

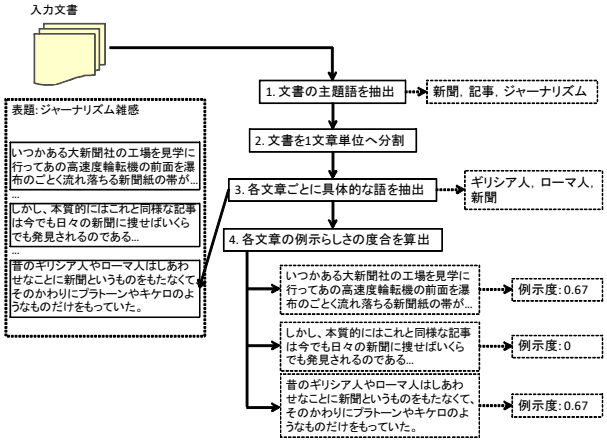

図 1: 提案手法の流れ

5. 著者の体験を記述
   著者の体験を用いて，例示を述べる場合である．例えば，ジャーナリズム論について述べる際に，著者が記者と対談した経験を引合いに出し，その経験を一般化した主張を述べる場合である．

本研究では，この 5 種類のうち，「3. 主題と異なる分野での事象を記述する場合」と，「4. 上位語から下位語へ遷移する場合」の 2 種類について例示部分を特定する手法を提案する．

3 上位下位概念と主題との関連度を用いた例示部分特定

3.1 基本的な考え方

本研究では，まず，主題を表す主題語を文書から抽出した上で，各文章中に含まれる具体的な語の総数に対して，主題語が占める割合を算出する．そして，その総数全体の割合から主題語が占める割合を引くことにより，各文章の例示らしさの度合いを算出し，例示部分を特定する．ここで，主題語とは，主題を表す単語を示す．

この理由として，まず，具体的表現を述べる文章には，具体的事象や事物を示す語が含まれやすく，それらの語が含まれている文章ほど，例示になり得ると考える．具体的な語とは，明確な実体や，個々の事物に即している語であり，例えば，“ローマ人” や “大日本印刷” という語が該当する．

このとき，具体的な語が文章に含まれていても，その語が主題を表していれば，例示として記述されてい

るとは考えにくい．例えば，ジャーナリズム論の文書中に “ローマ人” が出現する場合は例示として考えられるが，ローマ人の歴史に関する文書中に “ローマ人” が出現する場合は，文書の主題に沿っていると考えられ，例示として用いられているとは考えにくい．このように，具体的な語を含む文章は，主題に沿う文章か，もしくは，例示の文章のいずれかに分けられると考え，文章が例示を表しているかどうかを特定するためには，具体的な語が主題に沿っているかどうかを特定する必要があると考える．そこで，文章に含まれる具体的な語の出現傾向が主題に沿っていれば，その文章は例示である可能性が低いととらえ，主題に沿っていなければ，その文章は例示である可能性が高いととらえることにより，例示部分を特定することができると考える．

提案手法の流れについて，図 1 に示す．図 1 では，寺田寅彦による「ジャーナリズム雑感」[1]を例として用いている．

なお，例示であるかどうかの判定については，文書を句点 (“。”，“．”) で区切り，区切られた一つの文章ごとに，例示特定の判定を行う．

3.2 利用する上位下位語の条件

文章から具体的な語を抽出するためには，その語が明確な実態や個々の事物を示している語であるという情報が必要である．そのため，形態素解析結果の名詞だけを利用するといった方法では，その名詞が具体的な語であるかどうかを判断することができない．

そこで本研究では，文章に対して形態素解析処理を行い，形態素が名詞-一般および名詞-固有名詞である場合，その形態素を上位下位概念の情報を含む DB と照合し，その DB に含まれていれば，利用する語の候補とする．そして，各候補の語に対して，DB における語の深さを算出し，ある一定の深さよりも深い場所に位置する語だけを用いる．

語の深さを用いる理由は，DB に含まれる語をそのまま用いると，上位概念の語を利用する場合が発生するためである．例えば，文章から “人気”，“方法” といった語が得られ，これらの語が DB 内に存在する場合，これらの語も抽出対象となる．しかし，具体的な語としては，これらの語は不適と考えられる．

本研究では，形態素解析器として，MeCab[2] を用い，上位下位概念 DB として，日本語 WordNet[7] を利用した．日本語 WordNet では，語の意味概念ごとに上位概念，下位概念が定められており，抽象的な表現である上位概念から具体的な表現である下位概念まで，木構造によって構成されている．

[1]http://www.aozora.gr.jp/cards/000042/files/2492_10275.html
[2]https://code.google.com/p/mecab/

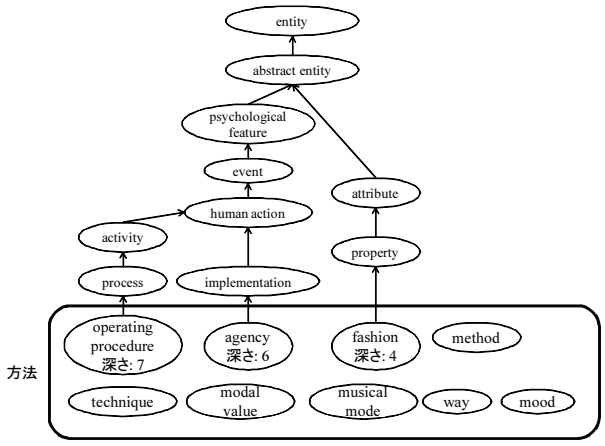

図 2: 深さによる具体的な語の絞込み

語の深さを用いて，利用する上位下位語を絞り込む処理について，図 2 に示す．まず，日本語 WordNet において当該語の上位概念を辿る．このとき，日本語 WordNet では，表記が同じでも，複数の意味を持つ場合，その意味ごとに *Synset* と呼ばれる概念が割り当てられている．そして，その概念ごとに語の上位概念が存在する．例えば，“方法” は，9 個の概念に属しており，その概念ごとに上位概念が存在している．

そこで，各概念ごとに上位概念からさらに上位概念へ辿って行き，上位概念が存在しない概念まで辿る．このときのそれぞれの深さを算出し，最小の深さを当該語の深さとする．この深さについて，深さの値が大きいほど具体的な語であり，小さいほど抽象的な語であると考える．最小の深さを適用する理由は，複数の概念のうち，一つでも深さの値が小さければ，その語は抽象的な意味で用いられる場合があり，すべての概念において深さの値が大きくないと，具体的な語とはいえないと考えるためである．

図 2 の例では，“operating procedure” という概念では，深さが 7 となり，一方，“fashion” という概念では，深さが 4 となった．“方法” では，最小の深さが 4 であったため，“方法” の深さを 4 とした．

そして，この深さに対して，あらかじめ閾値を設定しておき，閾値を超えていれば，具体的な語として用いる．

3.3 主題語の抽出

文書の主題とは，文書の中で著者が特に主張したい内容を表している表現である．先行研究では，表題に含まれる語を主題語としてとらえたり [8]，文書内の出現頻度が高い語を主題語としてとらえる考え [9] がある．そこで，本研究においては，以下の二つの条件のいず

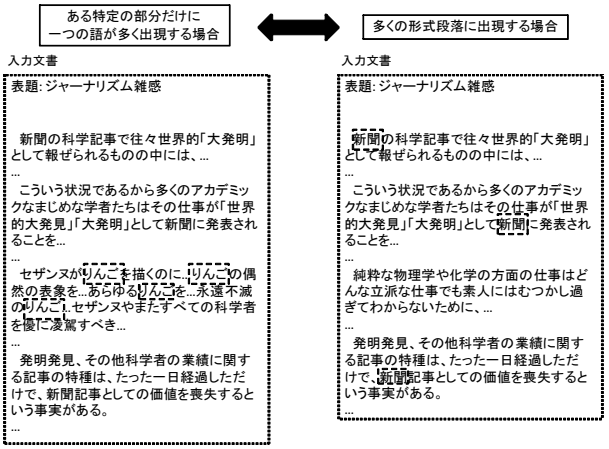

図 3: 出現段落数を用いた主題語抽出

れかに該当する語を主題語の候補として抽出し，各条件での候補を合わせた結果を主題語とする．

一つ目の条件は，表題に含まれる語である．まず，MeCab によって表題を形態素解析する．そして，3.2 節と同様に，形態素が名詞-一般および名詞-固有名詞である場合に限り，その形態素を日本語 WordNet と照合し，深さが閾値を超える語を，表題に含まれる主題語とする．

二つ目の条件は，出現する形式段落の数が多い語である．この理由は，著者の主張を示す主題語は，文書中で繰り返して記述されることが一般的だからであり，語の網羅性が主題語抽出のための指標として有効であると考えるためである．なお，このときの語は，3.2 節で述べた処理により，具体的な語だけに絞られている．

ある一箇所において特定の語が頻出する場合と，文書中に繰り返し出現する場合の違いについて，図 3 に示す．“りんご” という語は，ある形式段落において頻出するが，特定の形式段落にだけ出現するため，主題を表しているとは考えにくい．一方，“新聞” という語は，多くの形式段落において繰り返し出現しており，文書の主題を表していると考えられる．

文書内から全部で m 個の具体的な語 $w_1, w_2, \cdots, w_m$ が得られるとき，i 番目 $(i = 1, 2, \cdots, m)$ の語の主題らしさを表す度合 s_i を算出する式を，(1) 式に示す．

$$s_i = \frac{p_i}{N} t \tag{1}$$

ここで，p_i は語 w_i が含まれる形式段落の数，N は文書内の形式段落の総数，t は表題に含まれる語への重みを示す．

(1) 式では，文書内の形式段落の総数に対する，語が含まれる形式段落の数の割合を算出している．このとき，表題に含まれる語は，著者の主張を特に表していると考え，重み t を与えている．主題語として決定するために，s_i が，あらかじめ設定した閾値を超えていれば，その語を主題語として抽出した．ただし，抽出する主題語の最大数は 3 とした．

なお，合わせた結果が主題語の最大数を超えていれば，表題に含まれる主題語が，より著者の主張を表していると考え，表題に含まれる主題語を優先して用いることとした．

3.4 主題語の上位語の抽出

3.3 節で抽出した主題語に加え，主題語の上位概念も，主題を表す語として用いる．この理由として，ある主題について記述する場合，その主題語の上位概念も，文書の主題に含まれると考えられるためである．例えば，“新聞” を主題語とする文書の中で，“マスメディア” という語が現れる場合を考える．この場合，“新聞” の上位概念となる “マスメディア” は，“新聞” よりも抽象的な表現となる．そのため，文書内容が抽象的な表現と具体的な表現に大別されることを考えると，この文書では，“マスメディア” は抽象的な表現である著者の主張を表していると考えられ，例示のような具体的な表現として扱われるとは考えにくい．そこで，抽出した主題語だけでなく，主題語の上位概念も主題として用いることとする．

本研究では，日本語 WordNet において，主題語が属する概念の各上位概念に属する語を，主題語の上位概念として用いた．このとき，主題語と直接結びつく上位概念の語だけを用いた．

3.5 例示度の算出

本研究では，文章中に含まれる主題語の頻度と，主題語以外の具体的な語の頻度とを比較し，例示らしさの度合を算出することにより，例示部分を特定する．ここで，例示らしさの度合を例示度とする．

3.1 節で述べたように，具体的な表現を含む文章は，主題に沿う文章か，例示の文章の二つに分けられると考える．そこで，文章中に出現する具体的な語の総数のうち，まず，主題語とその上位語の割合を求め，主題に沿う文章であるかどうかの度合いを算出する．そして，総数全体の割合から主題の割合を引くことにより，その文章の例示度を算出する．文書内から全部で n 個の文章 $u_1, u_2, \cdots, u_n$ が得られるとき，j 番目 $(j = 1, 2, \cdots, n)$ の文章の例示度 e_j を算出する式を (2) 式に示す．

$$e_j = 1 - \frac{f_j + g_j}{M} \tag{2}$$

ここで，f_j は文章 u_j に含まれる主題語の出現総数，g_j は文章 u_j に含まれる主題語の上位語の出現総数，M は文章 u_j に含まれる具体的な語の出現総数を示す．

(2) 式により，例示度が高いほど，例示を表現している文章であると考え，例示度が低いほど，主題を表現している文章であると考える．

4 評価実験

4.1 実験方法

提案手法による例示特定結果と，既存手法による例示特定結果において，それぞれの再現率，精度，F 値を比較し，提案手法の評価を行う．

例示特定結果の評価として，文章単位での例示特定結果に加え，部分単位での例示特定結果についても評価を行う．

その理由として，提案手法および既存手法では，各文章が例示であるかどうかを判断するが，実際の例示は，一つの文章だけでなく，複数文章のまとまりによって一つの例示を表現している記述があるためである．そこで，実験対象の各文書内容を人手によって，例示を表現している複数文章のまとまりと，主題を表現している複数文章のまとまりに分けた．例示を表現している複数文章のまとまりを例示部分とする．

なお，例示部分を特定できたかどうかの判断としては，人手によって判定した例示部分に含まれる文章のうち，少なくとも一つの文章を特定できていれば，例示部分を特定できているとみなす．

提案手法については，例示度の閾値を $0.5, 0.6, 0.7, 0.8$ の 4 種類に設定し，各場合において，算出した例示度が閾値を超えていれば，例示とみなすようにした．

既存手法としては，手がかり語による例示特定を行った．このときの手がかり語としては，既存研究 [4][5] を参考に，“例えば”，“たとえば”，“例”，および，連体詞の “ある” を用いた．連体詞の “ある” は，“ある書籍” といった記述を指す．

表 1: 実験対象文書

文書番号	作品名	著者
1	ジャーナリズム雑感	寺田寅彦
2	科学的新聞記者	桐生悠々
3	漫画と科学	寺田寅彦
4	教育映画について	寺田寅彦
5	流言蜚語	寺田寅彦
6	形態について	豊島与志雄
7	芸術と社会	津田左右吉

実験対象の文書として，青空文庫[3]より 7 作品を用いた．具体的な作品名および著者を表 1 に示す．これらの作品を用いた理由は，これらの作品が評論や随筆のジャンルに該当し，例示を交えながら著者の主張を記述しているためである．

これらの作品に対して，人手によって，各文書内の各文章が例示文章であるかどうかを判定した．判定基準としては，2 章で述べた，5 種類の例のパターンをもとに判定した．

各作品の例示文章の割合を表 2 に示し，例示部分の割合を表 3 に示す．例示文章の割合の平均は 0.41，例示部分の割合の平均は 0.54 となり，文書内容の半数が例示であることを示している．この結果は，2 章で述べた，抽象的な表現と具体的な表現が繰り返されて記述されるといった論述形式を表している．

3.2 節における，具体的な語を絞るための深さの閾値は，表 1 の各作品の文書内容に含まれる具体的な語に対して深さを算出し，その算出結果から，4 とした．また，3.3 節の，主題語の抽出における，表題に含まれる語への重みは，$t = 2$ とし，主題語決定のための閾値は，0.4 とした．

表 2: 例示文章の割合

文書番号	全文章数	例	例以外	例の割合
1	148	59	89	0.4
2	69	9	60	0.13
3	80	20	60	0.25
4	75	29	46	0.39
5	46	23	23	0.5
6	55	32	23	0.59
7	44	26	18	0.59

表 3: 例示部分の割合

文書番号	全部分数	例	例以外	例の割合
1	56	30	26	0.54
2	21	9	12	0.43
3	28	15	13	0.54
4	26	15	11	0.58
5	11	6	5	0.55
6	20	12	8	0.6
7	7	4	3	0.57

[3] http://www.aozora.gr.jp/

4.2 実験結果

文章単位での例示特定結果について，提案手法および既存手法の再現率，精度，F 値を表 4 に示す．また，部分単位での例示特定結果について，表 5 に示す．提案手法の結果については，7 作品に対する例示特定結果の平均値を示している．

これらの結果から，精度においては，既存手法の方が上回っているが，再現率においては提案手法の方が，いずれの閾値においても上回っており，既存手法では特定できていなかった例示を特定できていることがわかる．

例示度の閾値を変化したことによる結果の違いについては，例示文章単位では閾値が 0.6 での F 値が最も高くなり，例示部分単位では閾値が 0.5 での F 値が最も高くなった．

例示文章単位では，誤って例示特定する場合が多く見受けられ，その誤り数が精度に影響を与えている．このときの精度低下の原因として，主題語が省略されていることを考える．主張を表すすべての文章には，必ずしも主題語が含まれているとは限らず，省略される場合がある．その場合，主題語以外の具体的な語が一つでも出現していれば，主題に沿う文章であるにもかかわらず，例示度を高く算出した．結果として，主題語が省略されている文章が頻出し，誤って例示特定した文章が多くなり，精度に影響を与えたと考える．

例示度の算出結果について，具体的な結果例を図 4 に示す．文章 1 は，例示として正しく特定することができた文章である．文章 1 が現れる文書の主題語は “記事”，“新聞”，“ジャーナリズム” であり，主題語の上位語には “マスメディア”，“ニュース” といった語が含まれる．そして，この文章には “セザンヌ”，“りんご”，“カンバス” が具体的な語として出現する．これらの語は主題に含まれない語であるため，例示度は 1 と算出される．

文章 2 は，例示度を考慮することにより，著者の主張が含まれることを正しく算出することができた文章である．文章 2 は，文章 1 と同じ文書内に現れるが，具体的な語として，“セザンヌ”，“新聞”，“科学”，“記者” が出現する．このうち，“新聞” が主題語となり，3 回出現しているため，例示度は 0.57 と算出される．文章 1 に比べ，文章 2 は著者の主張も含まれていると考えられるため，例示度を用いて，その区別を行うことができたと考える．

ただし，誤った例示特定を行う場合もあった．文章 3 は，具体的な語としては適さない語を利用していることにより，具体的な語の割合が高くなり，誤って例示と判断した場合である．この文章では，“類型”，“自身” といった語を利用しており，それらの語の頻度が高くなることにより，誤って例示と判断している．3.2 節

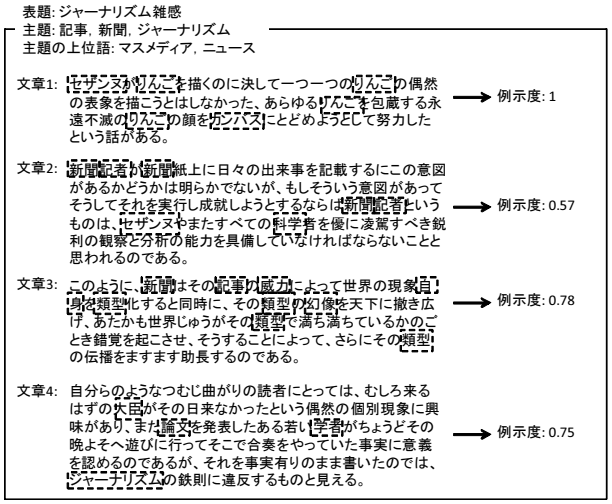

図 4: 例示度算出結果の具体例

による処理をもとに，利用する具体的な語の種類を絞り込んでいるが，深さだけでは抽象的な表現の語を除去しきれていないと考える．

また，文章 4 は，主題語に比べて具体的な語が頻出することにより，誤って例示と判断した場合である．この文章には，“論文”，“学者” といった具体的な語が出現する一方で，“ジャーナリズム” といった主題語も出現する．人手による判断では，この文章を著者の主張と判断したが，主題語以外の語の頻度に基づく提案手法では，この文章は例示である可能性が高いと判断する結果となった．出現頻度による重みだけでは，具体的な表現と一般的な表現を区別しきれていないと考える．

表 4: 文章単位の再現率，精度，F 値比較

	再現率	精度	F 値
既存手法	0.16	0.86	0.25
提案手法 (閾値: 0.5)	0.94	0.5	0.63
提案手法 (閾値: 0.6)	0.89	0.55	0.66
提案手法 (閾値: 0.7)	0.77	0.57	0.64
提案手法 (閾値: 0.8)	0.7	0.57	0.62

表 5: 部分単位の再現率，精度，F 値比較

	再現率	精度	F 値
既存手法	0.25	0.96	0.38
提案手法 (閾値: 0.5)	0.95	0.71	0.8
提案手法 (閾値: 0.6)	0.89	0.72	0.79
提案手法 (閾値: 0.7)	0.76	0.72	0.73
提案手法 (閾値: 0.8)	0.66	0.72	0.68

5 おわりに

本研究では，文書中の例示部分を特定する手法として，語の上位下位概念と文書の主題との関連度を用いて例示部分を特定する手法を提案した．この手法では，具体的な表現が含まれる文章は例示になりやすいことに着目し，既存手法では取得できていなかった，文章中に手がかり語が存在しない場合や，例示部分が独立して出現し，語の抽象度の遷移を取得できない場合において，例示特定についての再現率向上を行った．また，具体的な表現が含まれるからといって必ずしも例示にはならない場合があることも考慮し，具体的な語と主題との関係性を用いて例示度を算出することにより，精度低下の防止も行った．

評価実験の結果，手がかり語を用いた例示特定と比較して，再現率およびF値について，提案手法が上回っていることを確認した．

今後の課題として，精度向上のための2点の課題をあげる．1点目は，具体的な語としては適さない語を除去することである．

2点目は，人手による判断としては著者の主張として考えられる文章について，主題語に比べて具体的な語が頻出する場合でも，主題に沿う文章であると判別できるようにすることである．

参考文献

[1] Akio Suzuki. Differences in Reading Strategies Employed by Students Constructing Graphic Organizers and Students Producing Summaries in EFL Reading. *JALT Journal*, Vol. 28, pp. 177–196, 2006.

[2] 岡孝明, 武田英明. 技術論文のチャート化による論文理解の支援についての分析. 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 99, No. 447, pp. 27–34, nov 1999.

[3] 塚本真紀. 具体例の生成が文章理解による学習の転移に及ぼす影響. 尾道大学芸術文化学部紀要, Vol. 4, pp. 30–36, 2005.

[4] Daniel Marcu. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, Vol. 26, pp. 395–448, 2000.

[5] 梅澤俊之, 原田実. センタリング理論と対象知識に基づく談話構造解析システム DIA. 自然言語処理, Vol. 18, No. 1, pp. 31–56, jan 2011.

[6] 出口汪. 図解「出口式」論理力ノート : カリスマ講師が教える仕事で成功する思考法. PHP 研究所, 2006.

[7] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese SemCor: A Sense-tagged Corpus of Japanese. *In Proceedings of The 6th International Conference of the Global WordNet Association (GWC-2012)*, 2012.

[8] 野本忠司, 松本裕治. テキスト構造を利用した主題の推定について. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 96, No. 65, pp. 47–54, jul 1996.

[9] Gerard Salton and Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pp. 513–523, 1988.