文書横断文間関係タグ付コーパスの構築

衛藤純司
(有)ランゲージウェア
etoh@titan.ocn.ne.jp

奥村学
東京工業大学精密工学研究所
oku@pi.titech.ac.jp

１ はじめに

自動要約の分野では、最近、単一の文書だけではなく、複数の文書にわたる要約の研究が行われるようになってきた。一つのテーマをめぐって書かれた一連の文書においては、同じ事柄を述べた文ないし文章が繰り返し現れることがある。その現れ方も、言い換える、簡略に述べる、詳細に述べる、実例を挙げる、視点を変える、などさまざまである。複数の文書を要約するということは、このような相互に関連のある一連の文ないし文章から適当に取捨選択して再構成するということにほかならない。

Radev.D.R.は、複数のテキスト相互を関連づける理論として CST 理論（cross-document structure-theory）を構想し、２４の関係概念を提案している。[1]

私たちも、一つのテーマをめぐって書かれた一連の日本語文書において、文や段落の間にどのような関係があるか、また、全体として見た記事同士の間にどのような関係があるかを分析し、タグ付けを行った。

２ タグ付コーパスの概要

今回タグ付けの対象としたのは、新聞の報道記事である。全部で８５のテーマにわたっていて、それぞれのテーマには、最小３、最大１９、平均して１４の記事が含まれている。これはテキスト自動要約の評価型プロジェクト TSC(Text Summarization Challenge)で使用されたもので, すでに自動要約や１テキスト内のテキスト構造(文間関係)に関するタグが付与されている[2]。　また、本研究と並行して、植田らによって「照応, 省略, 共参照」に関するタグ付けが行なわれている[4]。

私たちは、一つのテーマに含まれる複数の記事を相互に比較して、次項に述べるような関連があると認められる文・段落・記事にタグをつけた。データの形式は次の通りである。

文レベル：<Sx：関係のタイプ>
・・・・・・・
</Sx：関係のタイプ>

段落レベル：<Py：関係のタイプ>
・・・・・・・
</Py：関係のタイプ>

文書レベル：<D z ：関係のタイプ>
・・・・・・・
</D z ：関係のタイプ>

S (Sentence)、P (Paragraph)、D (Document) はそれぞれ文レベル、段落レベル、文書レベルを表す記号であり、ｘ、ｙ、ｚは互いに関連する文・段落・文書のペアを指し示す数字である。

３　関係のタイプ

まず、私たちが用いた関係のタイプの一覧を示す。大ざっぱに分けて、事柄の同一性に基づく関係と、差異性に基づく関係に分けられる。さらに、後者は非時間的な差異と時間的な差異に分けられる。参考までに、Radev が提案した２４の関係概念のうち対応するものを加える。

同一性に基づく関係		同等(Identity, Equivalence)
		簡略(Summary)
		詳細(Refinement)
		例示
		並列(Parallel)
		参照(Citation, Attribution, Indirect speech, Agreement, Judgement, Change of perspective)
		補足
差異性に基づく関係	非時間的	対照(Contradiction, Contrast)
		追加(Elaboration)
		背景(Historical background)
	時間的	推移(Follow-up, Fulfillment)
		更新(Update)
		継起
		因果

Radev の関係概念のうち次の諸概念は私たちの関係概念と対応しない。
Translation/Subsumption/Cross-reference/ Modality/Generalization/Definition

同等：同じ事柄を述べたもの。まったく同一の表現である場合もあれば、異なる言い回しで同じ事柄が表現されている場合もある。

例１：「アウストラロピテクス・ガルヒ」（「ガルヒ」は現地語で「驚き」の意）と名付けた。

このため、新種の猿人化石として、現地語で「驚き」を意味する言葉から「ガルヒ猿人」と名づけられた。

簡略：同じ事柄を要約したり、一部を省略したりしたもの。時系列的に前の記事を後の記事が簡略化するという関係である。

例２：母親のドリーは、成長した羊の乳腺（せん）から取り出した細胞のＤＮＡが入った核を、未受精卵に移植する技術を使って一九九六年七月に誕生した。

ドリーは、妊娠中の羊から採った乳腺（にゅうせん）細胞をもとにつくり出された。

詳細：同じ事柄をより詳しく述べたもの。時系列的に前の記事を後の記事が詳細化するという関係である。

例３：百六十八人には合格通知とおわびの文書を郵送し、電話でも連絡。

同大学は二十一日、追加合格者に合格通知を郵便で発送。
同日午後から夜にかけて手分けして受験生の自宅などに電話をかけ、ミスがあったことを謝罪したうえで、入学の意思を聞いた。

例示：同じ事柄を、具体的な例を挙げて説明する場合。

例４：トルコ当局は２８日朝から生存者の救出作業を本格化した。

両都市の救助隊員は災害救助犬を使って、がれきの中に生き埋めになった住民を終日、捜索している。

並列：同じタイプの事柄をただ単に並べたもの。

例５：スウェーデン・アカデミーは８日、１９９８年のノーベル文学賞をポルトガルのジョゼ・サラマーゴ氏（７５）に贈ると発表した。

スウェーデン王立アカデミーは１４日、１９９８年のノーベル経済学賞をアマーチャ・セン英ケンブリッジ大教授（6）に授与すると発表した。

参照：同じ事柄を別の視点から述べるような場合。別の視点とは、例えば、過去を振り返る視点であったり、予測する視点であったり、理由づけをする視点であったり、さらには別の人物の視点であってもよい。別の人物の視点の特別な場合として、引用の関係もこれに含める。

例６：理由を問われると「私は気が小さいので慰留されると戸惑ったかもしれない」と述べた。

「慰留されれば戸惑ったかも」と語った１日の会見の言葉に、悩みの深さが表れている。

補足：ある事柄に関することを補足的に説明する。一連の記事の流れから見ると本筋ではないが、その事柄を何らかの意味で補うような場合である。

例７：パイロットらの賃金制度改定をめぐる全日空の労使対立で、同社の乗員組合（石飛明夫組合長、１３８０人）は6日午前０時から国際線の一部路線で無期限ストライキに突入した。

ストは、主力機「ボーイング７４７―０００」の就航する欧米路線などが対象。

対照：同じ事柄について違ったこと、対立すること、矛盾することを述べる場合である。

例８：その“視力”は東京から約１００キロ離れた富士山頂の５円玉の穴が見分けられるほどという。

東京から富士山頂のテニスボールを見分ける能力を持つ。

追加：ある事柄について、前の文にはなかった新しい情報を付け加える場合である。

例９：外務省領事移住部などによると、台湾に住む日本人永住者と長期滞在者は昨年１０月現在で約１万２３２６人。

台湾には約１万３０００人の在留邦人がいるが、２１日午前１１時半（日本時間同午後０時半）現在、邦人の被害は確認されていない。

背景：ある事柄の歴史的・非歴史的な背景を述べる。

例 10：タリバンは現在、アフガン全土の９割を制圧するが、政権を承認しているのはパキスタンなど３カ国だけ。

オマル師は１９９４年秋に創設した神学生組織を、４年間でアフガン全土の９割を掌握する勢力に発展させた。

推移：ある事柄が時間の経過とともに変化する場合である。例えば、ある事件が発生し、警察が捜査し、容疑者が逮捕され、起訴され、判決が下る、というような、一連の経過が述べられるような場合である。

例 11：昨年末から今月６日朝にかけて、２０代の女性３人が、神奈川県内で睡眠導入剤のトリアゾラム（商品名・ハルシオン）を飲まされ、うち２人が変死体で見つかっていたことが６日、分かった。

伝言ダイヤルに登録していた若い女性３人が神奈川県内で男に薬物を飲まされ、２人が昏睡（こんすい）状態に陥って凍死し、女子学生が財布を奪われた事件で、同県警捜査本部は７日午後、同県大井町出身で住所不定、無職、星野克美容疑者（２３）を女子学生に対する昏睡強盗の疑いで逮捕した。

事柄の変化には、特に数量に関わるものがあるが、それには二通りあって、そのうち、株価の変動や年平均気温の変動のように、その都度の数値が意味を持ち、数値の変動がそのまま経済や気候の変動を表すものを、この「推移」に含める。

例 12：１２日の日経平均株価は一時、前日終値より２００円以上値上がりしたが、外国人投資家の売りも依然多く、前日終値比６円９９銭高の１万４３７６円６２銭で取引を終えた。

３日の日経平均株価終値は前日終値比１１５円３８銭安の１万４２６１円２４銭と４日ぶりに反落した。

更新：数量の変化のもう一つのもの、すなわち災害の被害者数のように、数値の変動が被害者の数そのものの変動を表すのではなく、それまで判明していなかった正確な数値が判明しただけで、最後の数値だけが意味を持っているものがある。これを「更新」とする。

例 13：トルコからの報道によると、同国南部で２７日午後５時（日本時間同１１時）ごろ、マグニチュード（M）６・３の地震が発生し、崩壊した家屋の下敷きになるなどして少なくとも１０７人が死亡、約８００人が負傷した。

トルコ南部で２７日に夕発生した地震の犠牲者数は２８日夜までに死者１１２人、負傷者１５１７人に達した。

継起：同じタイプの事柄が時間を置いて生起する場合。上記「ノーベル賞」の例は、日付は異なるが、同じ行事の中にある同じ位置づけの出来事なので「並列」とするのに対して、次の例は日付が異なれば別の地震なので「継起」とする。

例 14：トルコからの報道によると、同国南部で２７日午後５時（日本時間同１１時）ごろ、マグニチュード（M）６・３の地震が発生し、崩壊した家屋の下敷きになるなどして少なくとも１０７人が死亡、約８００人が負傷した。

トルコからの報道によると、同国南部で４日午前５時（日本時間同１１時）ごろ、マグニチュード（M）５・１の地震が発生し、１０３９人が負傷した。

因果：ある事柄を原因として、後続の記事でその結果が述べられるような場合。

例 15:英国北アイルランドのカトリック武装組織アイルランド共和軍（ＩＲＡ）は十七日、武装解除の検討を約束する声明を発表した。

英・北アイルランドの和平交渉は、カトリック過激派、アイルランド共和軍（ＩＲＡ）が実質上、武装解除の検討に応じる声明を十七日に発表したことで、昨年四月の和平合意の履行に向けて一歩前進した。

４　関係タイプの頻度分布

文レベルの関係タイプの数を頻度順に掲げる。ただし、相互に関係のある文のペアを１件としたが、同等や推移や並列などでは３文以上が同じ関係を持つことがあって、その場合は全体を１件として数えた。

関係のタイプ	頻度
同等	６３８
簡略	１８１
推移	１３４
詳細	１３１
追加	５９
並列	３７
参照	３５
対照	２４
背景	２３
更新	１３
補足	１３
例示	６
継起	５
因果	５

５　要約の方略に関する考察

上記の頻度分布を見ても分かるように、同一性に基づく関係（同等、簡略、詳細）と、時間的な差異性に基づく関係（推移）が、主要な関係である。そこで、要約の方略として（１）「同等」ないし「簡略」の関係にある文から一つを選ぶ、（２）「推移」の関係にある文を連ねる、という２つの方法を組み合わせる、ということが基本になるだろうと考えられる。試しにこの方法で要約してみたものを次に掲げる。元の記事は「米国国防総省のネットワークにクラッカーが侵入した」記事で、記事数は７記事、総文数は４８文である。

> ジョン・ハムレ米国防副長官は二十五日、機密情報を含む国防総省のコンピューター・ネットワークが最近二週間、ハッカーによる組織的で大規模な侵入を受け、司法省が捜査を開始したことを明らかにした。
>
> 米国防総省のコンピューターネットワークに対する過去最大規模のハッカー侵入事件で、米連邦捜査局（ＦＢＩ）は、犯行グループのメンバーと見られる高校１年生２人を突き止めた。
>
> 米司法省は十八日、米国防総省のコンピューター網がハッカーによる侵入を受けた事件で、捜査協力を依頼していたイスラエルの警察当局が同日、主犯格の十八歳のイスラエル人少年を逮捕したと発表した。

６　おわりに

この種の分類体系はこれで必要十分だというものはない。試行錯誤を積み重ねて、誰にも受け入れられる安定したものに仕上げてゆくべきである。今後は、複数の作業者に実際にタグ付けをしてもらい、各作業者による揺れを分析して、より普遍性をもつものに改良していく予定である。

参考文献

[1]Radev,D.R.： A Common Theory of Information Fusion from Multiple Text Sources, Step One:Cross-Document Structure, ACL2000 1st SIGDial Workshop on Discourse and Dialogue, pp.74-83, 2000.

[2] 横山憲司, 難波英嗣, 奥村学: Support Vector Machine を用いた談話構造解析, 情報処理学会研究報告, NL-15, 2003.

[3]奥村学：テキスト自動要約, 情報処理, 特集「自然言語による情報アクセス技術」, Vol.45, No.6, pp.574-579, 2004.

[4] 植田禎子, 荻野孝野, 飯田龍, 乾健太郎, 奥村学：照応, 省略, 共参照タグ付コーパスの構築, 言語処理学会第１１回年次大会発表論文集, 2005.