異なる種類の文脈を統合することによる個人シソーラスの構築

柴田 博仁　大村 賢悟　吉岡 健
富士ゼロックス株式会社 研究本部
{hirohito.shibata, kengo.omura, takeshi.yoshioka}@fujixerox.co.jp

1. はじめに

シソーラスは意味を扱う言語処理においては欠かせないものであり，WordNet，EDR 電子化辞書をはじめ，これまで数多くのシソーラスが開発されてきた．一般に，シソーラスは万人にとって妥当な語彙体系を編集したものであるが，個人の語彙体系や連想パターンを模倣するシソーラス（「個人シソーラス」と呼ぶ）を想定すると，その利用価値はさらに広がる．例えば，検索時に個人シソーラスで検索語を展開することにより，個人の興味や関心をふまえた検索が可能となる [2]．Alto という語からコンピュータ科学の研究者は世界初のパーソナルコンピューを連想するであろうし，車好きの人は軽自動車を連想するだろう．Alto という語を個人シソーラスにより展開することで，上記のような語の多義性を排除した個人にカスタマイズされた検索が可能となる．

しかし，一般にシソーラスの構築はコストが高い．先に紹介したシソーラスはどれも，長い年月と膨大な費用をかけて人手で構築されたものである．不特定多数が利用するシソーラスの構築においては高いコストを投資することも妥当であろうが，個人シソーラスにおいてはより安価な構築が必要となる．

本研究では，個人化検索での利用を目的とし，個人が日常的に作成または取得したテキスト文書（論文，特許，研究メモ，メール，スケジュールなどであり「個人文書」と呼ぶ）における語の共起関係を解析することで，個人シソーラスを自動構築することを試みる．なお，本稿で取り扱うシソーラスは，見出し語とその関連語の組であり，見出し語と関連語には関連度が与えられている [4], [7]．シソーラスの向上を目指した提案を行い，初期実験の結果をあわせて報告する．

2. アプローチ

語の共起に基づくシソーラス構築では，どのような共起関係を利用するかということが重要な課題となる．これまでの提案は大まかに 2 種類の共起関係に基づくものに分類できる．一方は語の近接性に基づく共起（例えば，[1], [3], [4]）であり，他方は文法的な共起（例えば，[2]）である．前者では，2 つの語が同一のブロックで利用されているときに共起しているとみなす．後者では，2 つの語が同じ文法構造（例えば，同じ動詞の主語として利用されている 2 つの語）で利用されている場合に共起しているとみなす．

これらの手法において，必ずしも一方が他方より優れているというわけではない．語間のどのような関係の抽出を望むかということに応じて，利用する共起関係を定める必要がある．文法的な共起により文法的に交換可能な語間の関係が抽出されることが期待できるが，われわれが望むのは個人化検索において有効であろうと思われる個人の連想関係に基づく語間の関係の抽出である．また，個人文書には，特定の個人やグループでのみ利用される単語や略語（「個人語」と呼ぶ）が含まれており，非文法的な記述（句読点がない，主語や動詞が暗黙のものとして省略されているなど）も多い．このようなテキストに対しては，構文解析が正しく行われない可能性がある．そこで本研究では，近接的な共起関係に基づいてシソーラス構築を試みる．

一般に，シソーラスを構築するには数十メガから数百メガのテキストが必要であるが，個人の文書群においてはこのような量のテキストは望めない（せいぜい数メガ）．そこで，限られた文書群の中からできる限り個人の直感に合うシソーラスの構築を目的とし，語間の関係を精緻に捉え，またテキスト上での文脈に限定されることなく語の文脈を広く捉えるアプローチを提案する．

異なる粒度の文脈の混在を許す

共起関係に基づいてシソーラスを構築する際，共起関係を決定するブロック（本稿ではこれを抽象化し「文脈」と呼ぶ）のサイズをどの程度にするかは重要な問題である．これまでに，段落 [6]，文書 [6]，辞書での見出しと説明文からなる項目 [4], [7]，固定ブロック長 [5] など，さまざまな文脈の取り方が試みられてきた．

異なる粒度の文脈は，異なる種類の語間の関係をもたらす．文脈の粒度を小さくすることで，"bread and butter" のようにイディオムとして利用されるような語間の関係が抽出され，粒度を大きくすることで語間の意味的な関係が強調される [1]．これまで，文脈の粒度を変更することで，抽出される語間の関係を比較したり [6]，多義性解消のために最適な文脈の粒度を探す [5] という試みはなされてきたが，異なる粒度の文脈を統合的に扱って

語間の類似度を判別する試みは，著者らの知る限り存在しない．

本研究では，異なる粒度の文脈の混在を許し，文脈の粒度に応じて共起の度合を変化させる枠組みを提案する．具体的には，2 つの語が狭い文脈で共起している場合には高い共起度を与え，広い文脈で共起している場合には低い共起度を与える．その目的は，限られた文書群の中で，できるだけ多くの文脈を活用するとともに，文脈の粒度に応じて共起の度合を変化させることで語間の関係を精緻に捉えることである．さらには，さまざまな種類の文脈の混在を許容することにより，語間の多様な関係を抽出することを期待している．この枠組みは，これまでの共起関係の捉え方を拡張したものと考えることができる．

語の社会的文脈の活用

文書は社会的産物である．文書がどのような人にどのような状況で作成され，どのような場所で利用されたか，またどのような経路で流通されたか，これらはどれも文書のもつ社会的文脈の一つである．それは言い換えるなら，文書に出現する語の文脈であるともいえる．例えば，ファイル（文書）は通常，フォルダと呼ばれる階層構造に分類されて管理される．異なるファイルが同一のフォルダに分類された場合，それはユーザが「これらファイルは同種のものである」というメッセージを分類という操作をとおして埋め込んだと考えることができる．また，一人の人間の活動を考えると（ある時期は数学の研究に従事し，その後，言語処理の研究に携わるなどというように）人の活動や興味は長いスパンでみるとゆるやかに変化することが多い．よって，時間的に近い時期に作成された 2 つの文書は，時間的に離れた時期に作成された 2 つの文書よりも関連性が強い可能性が高い．このように，文書はそれが作成，利用される文脈をもち，この文脈をもとに文書間の関連性を予想することは妥当と考えられる．言い換えるなら，文書は社会的文脈に基づいた暗黙の関連性をもっているといえる．

これまでのシソーラス自動構築の試みは，文書がもつ社会的文脈を切り離し，語のテキスト上での関係のみを扱ってきた．文書の（または語の）社会的文脈を考慮してシソーラスを構築することにより，文脈の数を増やすだけでなく，個人のワークスタイルや嗜好性をより強く反映したシソーラスの構築が可能と考える．

本研究では，このような文書のもつ社会的文脈をシソーラス構築に積極的に活用する．文書の社会的文脈としては，著者，共著者，目的，利用場所など，さまざまなものが考えられるが，本稿では文書の置き場所と作成日時にフォーカスをあてる．具体的には，2 つの語が同じフォルダ内の異なる文書で利用されている場合，これらは広い文脈で共起しているとみなす．また，2 つの語が同じ時期に作成された異なる文書で利用されている場合，これらは広い文脈で共起しているとみなす．

ここで議論した文書の社会的文脈は，前節でいうところの異なる粒度の文脈の一種と考えることができる．文脈を文書内での語の利用に限定することなく，文書間の関係にまで拡張したものとみなすことができる．

3. シソーラス構築

前節のアプローチに基づき，ここでは異なる種類の文脈を統合してシソーラスを構築する方法の概要を示す．

語の文脈での重みを算出（語-文脈行列の作成）

まずは個人文書を形態素解析する．この際，個人語の抽出を目的とし，日本語かな漢字変換システムのユーザ辞書の単語を形態素解析の辞書として取り込む．重要な概念を示す語の多くは名詞であるという考えに基づき，名詞を抽出する．なお，極端に出現頻度の少ない語は除外する．

次に，与えられた文脈の種類（文，段落，文書など）に応じて文脈を抽出する（本稿では，文脈の集合と個々の文脈を区別し，前者を「文脈タイプ」と呼ぶ）．ここで，文脈タイプが l 個あるものとし，個々の文脈タイプの文脈集合を C_k として表記する．さらに，C_k 毎に，行が語集合，列が文脈集合に対応し，行列要素を語の文脈での重みとする語-文脈行列 A_k を作成する．

ここで，文脈における語の重みを単純に TF-IDF で算出することには問題がある．個人文書は，個人の興味や関心に応じて作成・収集されたものであり，記述内容の偏在性が高い．よって，個人の関心のある語が高頻度で出現することは必然であり，IDF の利用により高頻度の語の重要性が低いと判断されるのは不適切である．IDF を利用せず TF のみで重みを算出するか，IDF の算出において個人文書群以外のコーパスを利用するなどの工夫が必要である．次節で述べる実験では TF のみを利用して重み付けを行っている．

最後に，各行列 A_k の各列ベクトルのノルムが 1 になるよう正規化する．

類似度の算出

最初に，各文脈タイプ C_k での文脈をどれくらい重視するかを示す指標として，C_k に係数 α_k が与えられているものとする．異なる種類の文脈を統合して語間の類似度を定めるにあたり，ここでは「連接方式」と「結合方式」の 2 つの方法を提案する．

連接方式では，複数の語−文脈行列 A_k を比重 α_k でスカラー倍し，これらを横に連接した横長の行列 $A = (\alpha_k A_1 \ \alpha_2 A_2 \ \ldots \ \alpha_l A_l)$ を作成する（A の列の数は $m =$

Σm_k). 語間の類似度は, 行列 A の対応する行ベクトルの余弦により算出する.

結合方式では, 個々の語-文脈行列 A_k で算出した語間の類似度 s_k を各文脈の比重 a_k で統合する. 類似度の統合では加重平均を利用する.

ここで述べた連接方式, 結合方式により算出された語間の類似度に基づき, 個々の語について類似度がしきい値以上の語を収集したものがシソーラスとなる.

4. 実験

シソーラスの構築

提案方式の有用性を探るため, 著者の 1 人を被験者として初期実験を実施した. 被験者はユーザインタフェースの研究者であり, 7 年間にわたり情報管理ツール「情報箱[1]」によりさまざまなテキスト情報を蓄積している. 情報箱は Windows のエクスプローラとメモ帳を組み合わせたような GUI を備え, 主にテキスト情報の管理を可能とする. 被験者が蓄積しているテキスト情報の種類としては, 自分が書いた論文, 特許, 報告書, 研究メモ, 読んだ論文や本の情報, 重要なメール, 日常業務のノウハウ, スケジュール, 日誌などである. 本研究の枠組みが情報箱の機能やデータ形式に限定されることはないが, 比較的広範にわたる個人の文書群をもとにシソーラスを構築したいという理由により, 本実験においては社内で手軽に利用できる情報箱のデータをシソーラス構築の対象として利用した.

構築に利用した文書数は 16,471 件 (7.6MB) であった. 形態素解析では「茶筌」を利用し, 個人語の抽出を目的として, かな漢字変換システム IME のユーザ辞書に登録された 1,013 語を茶筌の辞書に取り込んだ. 語の抽出では出現頻度 10 以上の名詞 5,201 語を抽出した.

前節で紹介したシソーラス構築方法に従って, 各文脈タイプ C_k の比重 a_k を変動させることで, 複数のシソーラスを構築した. 個々のシソーラスの構築時間は通常の PC (CPU は 2.4GHz, メモリは 524MB, OS は Windows 2000) で 2 時間程度であった.

評価方法

シソーラス構築のアルゴリズムを比較するため, 以下の手順に従いサンプル語と被験者の主観的評価に基づく関連語からなる評価データを作成した.

1. シソーラスに掲載される 5,201 語の中からランダムに 80 語を選定し, サンプル語とした.
2. 80 のサンプル語に対し, 各アルゴリズムで上位 30 の語を収集した. さらに, ランダムに選んだ 20 語を加え, 語の順番をランダムに入れ替え, 各サンプル語に対する刺激語のリストを作成した. 各サンプル語に対する刺激語は, 平均で 98.9 語であった.
3. 被験者には, 最初, サンプル語を提示しサンプル語に対する自由連想を行わせた. 次に刺激語のリストを提示しサンプル語との関連性について 3 段階 (関連なし (0), 弱い関連あり (1), 強い関連あり (2)) で評定を行わせた.
4. 被験者が自由連想した語の中でシソーラスに掲載されているものと, 刺激語に対する評定で 1 または 2 とマークしたものを関連語とした. 各サンプル語に対する関連語数は, 平均で 23.1 であった.

アルゴリズムの評価においては, 再現率と適合率を組み合わせた F 測度を利用した. 再現率を R, 適合率を P とするとき, F 測度は R と P の重みつき調和平均として表される. 実験では, R と P を同じ重み (ともに 1/2) にし, 80 のサンプル語について, 各アルゴリズムで算出された上位 25 の語を対象として F 測度を算出する.

結果と考察

文脈タイプとして, 文 (C_1), 段落 (C_2), 文書 (C_3), フォルダ階層に基づく文書グループ (C_4), 文書の作成日時に基づく文書グループ (C_5) の 5 つを考える. 文脈タイプの比重を各々 a_1, a_2, a_3, a_4, a_5 とし, その組み合せを a_1-a_2-a_3-a_4-a_5 のように表記する.

実験の目的は F 測度が最高となる a_1 から a_5 の組を探ることにより, 本アプローチの利用によるシソーラスの性能向上を確認することである. なお, C_4 では同一のフォルダに属す文書を 1 つの文脈とし, C_5 では同じ月に作成された文書を 1 つの文脈とした.

全てのパラメータを同時に変動させると組み合せの数は膨大になる. そこで, 最も効果的な文脈タイプを固定し, 他の文脈タイプによる効果をそれに付加するという形で最適なパラメータの組み合せを探す. 単独の文脈タイプを利用して F 測度を測定した結果, 文脈タイプとして C_1 を用いた場合は 0.361, C_2 では 0.382, C_3 では 0.389, C_4 では 0.122, C_5 では 0.113 であった (文脈タイプが 1 種類の場合は, 連接方式, 結合方式とも同じ値を示す). 単独の文脈タイプとして文書 (C_3) を利用する場合, 最も F 測度が高くなった. そこで, 以降では, a_3 を 1 に固定し, その他のパラメータの値を変動させる. そして, 単独の文脈タイプを利用する従来手法の中で最良の結果を示した 0-0-1-0-0 (文脈タイプの比重として文書のみが 1 であり, 残りは 0 の場合) と比較する.

最初に, 文書より文脈の粒度として小さい文 (C_1) と段落 (C_2) の文脈の統合について考える. 連接方式においては, a_1, a_2 ともに 0 から 2 まで 0.2 刻みでパラメータを変動させた. その結果, 0.2-0.2-1-0-0 において F 測度は最高値 0.448 を示した. 従来手法 (C_3 単独の場合) に対して 0.059 の向上 (性能向上率にして 15.2% の向

[1] http://www.fujixerox.co.jp/soft/johobako/

上）を示した．また，結合方式において 0 から 6 まで 0.2 刻みでパラメータを変動させた結果，2.8-0.4-1-0-0 においてF測度は最高値0.476を示した．従来手法に対して 0.087 の向上（性能向上率にして 22.4%の向上）を示した．なお，C_1, C_2, C_3単独の文脈を利用する従来手法と，提案手法の上記の 2 つの値との比較を図 1 に示す．以上の結果から，パラメータを適切に選ぶことで，異なる種類の文脈タイプを統合的に扱う本研究のアプローチによりシソーラスの向上が可能であることがわかる．

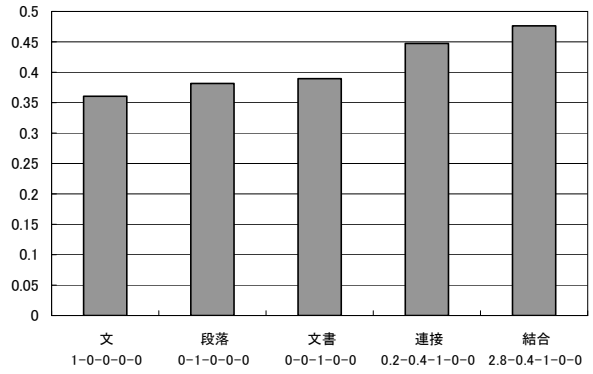

図 1. シソーラス構築方式での F 測度の比較

次に，社会的文脈（C_4, C_5）の活用の効果を調べる．連接方式においては，a_4, a_5 ともに 0, 1, 5, 10, 20, 50, 500, 1000 と不連続な刻みでパラメータを変動させた．結果として，0-0-1-1-1 の場合において 0-0-1-0-0 の場合に比べ 0.001 だけ F 測度の向上が観察されたが，それ以外は 0-0-1-0-0 の場合よりも低かった．結合方式においては，0 から 1 まで 0.1 刻みでパラメータを変動させた．その結果，0-0-1-0-0 の場合が最高であり，それを超えるパラメータの組み合せが観察されることはなかった．以上の結果から，今回の実験では社会的文脈の活用によりシソーラスを向上させることはできなかった．その理由として活用した社会的文脈における文脈の数が極端に少なかったことがあげられる（C_3 の文脈数 16,471 に対して，C_4 が 70，C_5 が 72）．情報箱は通常のファイルシステムに比べて豊富な検索機能を備えるため，フォルダによる階層分類があまり行われなかった（被験者は 1 つのフォルダに平均で 253.3 個の文書を格納していた）．また，文書の作成日時による分類に関しても，月毎の分類では分類の仕方が粗すぎたといえる．

今回の実験では，従来手法に比べて，連接方式で 15.2%，結合方式で 22.4% の性能向上を示した．連接方式よりも結合方式の方が高い性能向上を示したが，このことは必ずしも結合方式の優位性を示すものではない．結合方式は連接方式に比べてパラメータの取り方に敏感であり，うまくパラメータを設定した場合には高い性能を示すが，パラメータの選び方に応じては連接方式に比べて性能を落とすこともある．

5. 結論

本稿では，個人シソーラス構築に向けた 2 つのアプローチを提案した．一方は異なる種類の文脈を統合的に扱って語間の類似度を定めるアプローチである．文脈として文書のみを利用する従来の最良の手法に対して，提案方式は性能向上率にして（結合方式にて）22.4%の向上を示した．他方は，文書のもつ社会的文脈をシソーラス構築に活用するアプローチであり，今回の実験で効果を確認することはできなかった．

本稿では，パラメータの決定方法に対する明確なガイドラインを与えることはできなかった．パラメータを網羅的に変動させるいわば力まかせの探索を行ったのは，「異なる種類の文脈を統合することにより，さらには統合のパラメータを適切に設定することにより，シソーラスの性能向上が可能である」ということを期待したためである．また，1 人の被験者に対する実験から多くのことを主張する危うさを危惧したためでもある．今後，本実験で効果を確認できなかった社会的文脈に対する取扱いを改良していく予定である．また，現在，本稿での結果の一般化を目的とし，複数人に対して同種の評価実験を行っている．各文脈タイプに対する比重の与え方の指針を与えること，連結方式と結合方式の特徴をより明確にすることが今後の課題である．

参考文献

[1] K. W. Church and P. Hanks: Word association norms, mutual information, and lexicography, *Computational Linguistics*, 16 (1), 1990.

[2] R. S. Flournoy, etc.: Personalization and users' semantic expectations, *Proc. of SIGIR Workshop*, 1998.

[3] D. Hindle: Noun classification from predicated-argument structures, *Proc. of ACL*, 1990.

[4] 笠原要，松澤和光，石川勉: 国語辞書を利用した日常語の類似性判別，情報処理学会論文誌，38 (7), 1997.

[5] H. Schutze and J. Pedersen: A cooccurrence-based thesaurus and two applications to information retrieval, *Information Processing and Management*, 33 (3), 1997.

[6] 渡部勇，三末和男: 単語の連想関係によるテキストマイニング，情報処理学会研究会 情報学基礎，FI55-8, 1999.

[7] 渡部広一，河岡司: 常識的判断のための概念間の関連度評価モデル，自然言語処理, 8 (2), 2001.