社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（18）
2004／11／5

シソーラスを用いた意味フレーム階層ネットワークの効率的構築

金丸 敏幸†

†京都大学大学院人間・環境学研究科 〒606-8501 京都府京都市吉田二本松町
E-mail: †kanamaru@hi.h.kyoto-u.ac.jp

あらまし 本稿では，黒田，井佐原[4]で提唱されたFOCAL (Frame-Oriented Concept Analysis of Language)の分析を行う上で，効率的に階層意味フレームを発見，特定するための手法を提案する．具体的には，集めたデータに対し，シソーラスを用いて同じカテゴリーに属する語をまとめていく操作を行う．上位語をシソーラスの分類に従ってまとめていくことによって，最終的に階層的ネットワークを得る．本手法を用いることによって，意味フレームを発見するまでの労力を低減できるだけでなく，意味フレーム同士の関係を予測することも可能となることを明らかにする．

キーワード シソーラス，階層的意味フレーム，言い換え

Efficient Hierarchical Semantic Frame Network Construction using Thesaurus

Toshiyuki KANAMARU†

† Graduate School of Human and Environmental Studies, Kyoto University　Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto, 606-8051 Japan
E-mail: †kanamaru@hi.h.kyoto-u.ac.jp

Abstract This paper describes a method to construct a hierarchical semantic frame network which is a target of description and investigation for FOCAL (Frame-Oriented Concept Analysis of Language), proposed by Kuroda & Isahara[4]. First, Collected words are paraphrased by definition of thesaurus. Then paraphrased words are united by hypernym to get hierarchical semantic frame finally. The experimental result showed that this method made it possible to not only reduce efforts for detection of semantic frame, but estimate the relations of semantic frames.

Keyword thesaurus, hierarchical semantic frame, paraphrase

1. はじめに

黒田，井佐原[4]で提唱されたFOCAL (Frame-Oriented Concept Analysis of Language)は、ヒトが行っている意味理解のメカニズムを分析する上で、有力な手法となりうる可能性を持つ．しかしながら、現段階では、初期の段階で得られた全てのデータに対し、人手で意味役割や意味タイプといった意味タグを割り振る作業を行う必要がある．この作業にはかなりの労力がかかる上、重複したデータもあるため、必ずしも全てのデータを分析する必要があるわけではない．

また、ある程度の量を分析するまでは、対象とするデータがどのくらいの意味フレームを持つかといった見通しが立たないため、効率がよいとは言えない．

そこで本稿では、FOCALの分析を行う上で、その記述，分析対象となる意味フレームを効率的に発見，推定するための手法を提案する．具体的には、まず，収集したテキストデータに対し、シソーラスを用いて同じカテゴリーに属する語を置き換える操作を行う．次に，重複語をまとめた後，再びそれらを上位語に置き換える．以上の手順を繰り返すことにより，語をシソーラスの分類に従ってまとめていくことによって、最終的に階層的ネットワークを得るというものである．

以下，本稿の流れについて述べる．まず本手法の対象となるFOCALについて概観した後，現在の分析における問題点を指摘する．次に，その問題点を回避し，FOCALの分析を効率よく進めるための手法を紹介する．その後，3.で今回行った実験について説明し，最後にその結果と考察について述べる．

2. 背景

2.1. FOCALとは

最初に，黒田らが提唱したFOCAL (Frame-Oriented Concept Analysis of Language)とは，どういったものであるかを簡単に紹介する．

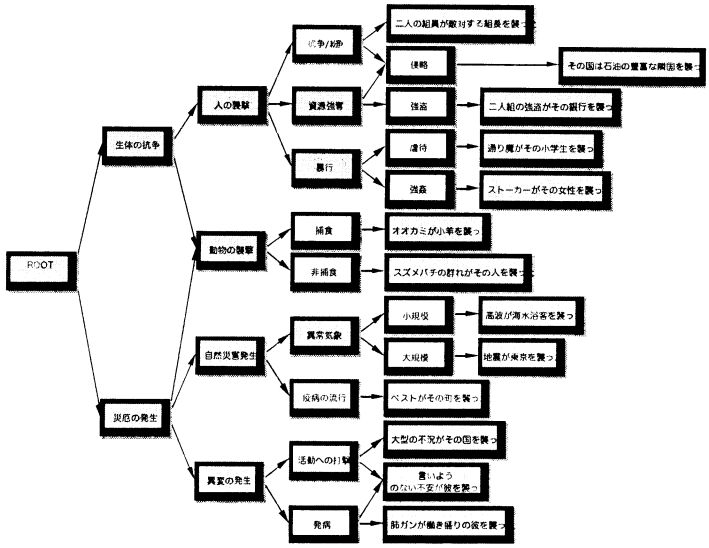

図 1　「X が Y を襲う」の意味フレームのネットワーク(黒田[3]より)

FOCAL は，Berkeley FrameNet (BFN)[1] に強く影響を受けながらも，それとは独自に行われているフレーム指向概念分析と呼ばれる理論的枠組みである．FOCAL の目的は，ヒトの理解には有限個の単位が存在するという仮説の下，その理解のための単位となる「状況」とは正確に何であるかを記述することである．その際に重要となるのが，意味フレームの階層ネットワークである．FOCAL は意味フレーム(semantic frames) の階層ネットワークを意味知識の構造表現に採用するという知見を BFN から取り入れている．

意味フレームは全てが同じレベルの抽象性を持っているわけではなく，その理解のレベルに応じた抽象性を持つとされる．もし，意味フレームの抽象性が低ければ「深い理解」が生じ，抽象性が高ければ「浅い理解」が生じる．この意味フレームの抽象性は，意味フレームを実現する各意味役割が持つ意味素性の数や意味素性同士の具体性，抽象性と関わりを持つ．

従って現段階では，理解のための単位となる「状況」が，実際にはどのような階層ネットワークを持っているか，意味フレームを用いて記述することが当面の目標となる．

2.2. 現在行われている手順

現段階において，FOCAL がどのような手順で分析を行っているのか，実際の例を見てみる．

黒田ら[5]は，動詞「襲う」が持つ意味フレームの階層ネットワーク分析を行うに当たり，初期の段階において，次の手順で分析を行っている．

(1)　まず，対訳コーパス(内山，井佐原[7])から，KeyWord In Context(KWIC)ツールを用いて「襲う」の全用例を抽出する．

(2)　次に，抽出した用例に対し，人手で分析して意味フレームを特定していく．その際，意味フレームを特定するために，次の[(a); (b)1, 2, 3; (c)1, 2, 3; (d)] の情報を指定する．

(a)　文 S が実現しているフレーム名

(b)

1.　S の主語句 s と
2.　s の意味型，並びに
3.　s の意味役割(= FE)

(c)

1.　S の目的語句 o と
2.　o の意味型，並びに
3.　o の意味役割

(d)　S の意味フレーム．

(3)　以上を全ての用例に対して行った後，指定された意味役割を基に，得られた意味フレームを階層ネットワークの形で記述する．

このようにして得られた意味フレームの階層ネットワークの例を図 1 に示す．

2.3. 現在の手法の問題点

この手法には，次のような問題点がある．

まず，得られた例全てを見ているという点が挙げられる．意味フレームは，その性質上，できる限り幅広い事例を分析することが求められる．分析を行わなか

ったデータの中に未知の意味フレームを持つ可能性もあるため，できる限りランダムサンプリングは行わない方が望ましい．そのため，得られたデータの一部をランダムで選び，分析するだけでは不十分である．

しかし，実際の用例の中には，数万の用例を持つ動詞も多数存在する．従って，得られた例全てを見ていくという手法には限界があると言わざるを得ない．

次に，意味フレームの全体像に対して見通しが悪いという点がある．人手で意味フレームをふっていく場合，どのような意味フレームを設定すればよいかという指標が存在しないため，分析を開始してしばらくしなければ，意味フレームを特定しにくい．ある程度の数のデータを見て，初めて傾向を把握でき，意味フレームを設定できるようになるというのが現状である．

確かに，このように最初から人手で分析を行っても，意味フレームが設定できないわけではない．しかし，初期の段階で何の見通しもないまま分析を行うことの効率の悪さは明白である．

以上のことから，この分析における初期段階において，存在する可能性のある意味フレームについて，その範囲を粗く設定することができれば，その後の作業効率と分析の精度を高めることができると考えられる．

2.4. 格フレームと意味フレームの関係

そこでまず，意味フレームの特徴に注目し，コンピュータを用いることで，半自動的に意味フレームの存在を特定する手法について考える．黒田[3]によると，意味フレームは，(状況)理解の単位であり，典型的には<<何が><いつ><どこで><何のために>...<何を><どうする>>のような形で表現できるものである．このような形で表現できるということは，黒田が注で述べているように，意味フレームが外延上，自然言語処理の分野で格フレームと呼ばれているものとほぼ一致することを示している．

格フレームとは，動詞を中心として，その動詞がどのような格要素を持つかという情報を記したものである．自然言語処理の分野では，格フレームは構文解析や意味分析のために用いられている．そのため，精度のよい格フレーム辞書を構築するための手法が，河原，黒橋[2]や宇津呂，宮田，松本[9]などのように，いくつか提案されている．ただし，格フレームはあくまで処理のための辞書のような存在であり，それぞれの格フレームが互いにどのような関係を持つのかといったような分析はあまり行われていない．

そのため，自然言語処理の分野における格フレームの自動構築の手法をそのまま持ち込んだ場合，意味フレームの数を特定することはできても，意味フレーム同士の関係や階層性といったものは，依然明らかでない可能性がある．このような問題を解決するためには，用例に対して，似たような性質を持つものをクラスタリング化するだけでなく，まとめたもの同士の階層性や関係性を明らかにしていく必要がある．

2.5. 階層構造への注目

次に，意味フレームの階層性に注目し，階層構造を持ったシソーラスの利用を考える．階層構造持つシソーラスを用いて上位語へ次々と置き換え，その置き換えの履歴を利用すれば，用例同士の階層性を半自動的に構築することができる．

このように階層構造を持つシソーラスを利用することにより，抽出された用例が持つ各要素同士の関係を明らかにするのが，本手法のねらいである．

3. 実験手順

3.1. 対象とデータ

まず，今回の実験で対象とした意味フレームとデータについて述べる．今回は，日本語の「襲う」について調査を行った．対象として「襲う」を選んだ理由は，黒田らが先に人手により分析を行っており，比較のための階層ネットワークが得やすかったからである．

「襲う」についての用例を抽出した実際の生コーパスデータは，対訳コーパス，Webから収集したテキストデータ約 172MB(個人の日記，インタビュー，コラム等)，新潮文庫の 100 冊 CD-ROM 版，電子百科事典の本文テキスト(小学館スーパーニッポニカ，平凡社世界大百科事典，日本語大辞典)であり，総容量は約 650MB であった．

これらのデータから，KWIC ツールを使い，「(にが)襲(わいうえっ)」を検索語として，用例を抽出した．この段階での全用例数は 1103 例であった．KWIC ツールを使って抽出した用例には，先行文脈，直前語，検索語，後続文脈などいった情報も残っている．今回は，この中の「直前語」を置き換える作業を行った．

3.2. 日本語語彙大系語と置き換えの条件

語の置き換えには，NTT コミュニケーション科学研究所の『日本語語彙大系(以下，大系)[7]』を用いた．大系は，全体で 2700 あまりの意味属性を持つ木構造（最大 12 段）であり，それぞれの語の意味の階層性をできるだけ反映するようにした．

まず，最初の段階で，直前語をその語が含まれるカテゴリーを表す上位語に置き換える．置き換えを行う際には，以前の語を全て保存し，置き換えの履歴とする．

置き換えには，次の二つの条件を設けた．一つ目の条件は，多義性についてである．今回は，多義的な意味を持ち，複数の意味属性に含まれる語については，そのどれにも置き換えることにした．これは，この後の作業で他の意味属性との統合を行うため，幅広く置

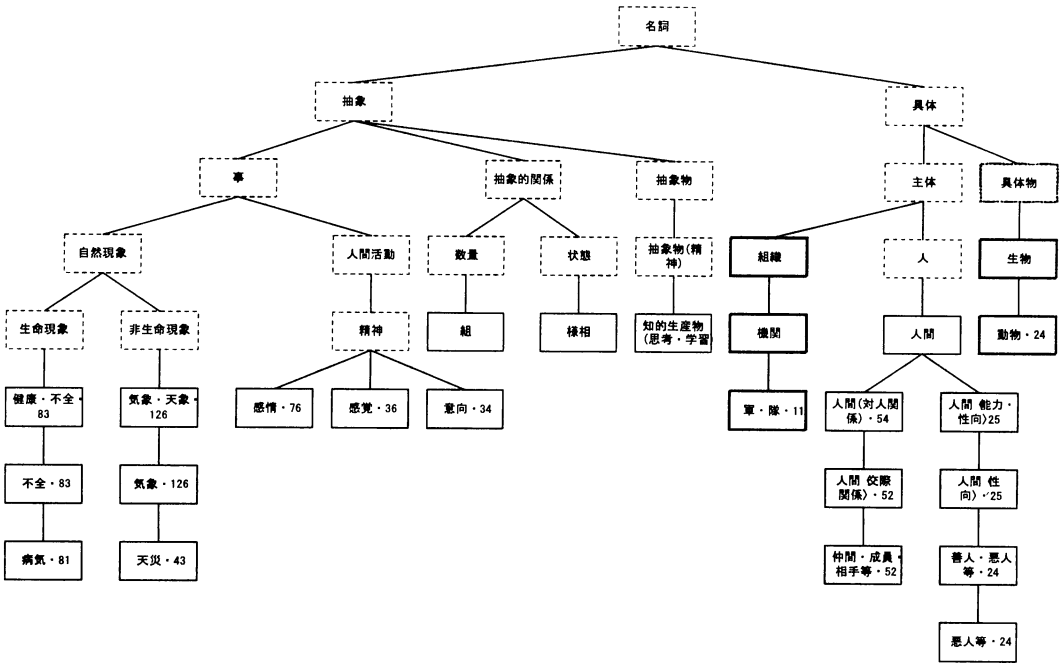

図 2　自動構築によって得られた「襲い手」の階層意味ネットワーク

き換えをしておいても，本来の意味からはずれた用法は，統合の段階で無視できるようになると考えたためである．

二つ目は大系にない語の取り扱いについてである．大系は約 10 万語の一般名詞を登録してあるが，これで全ての名詞をカバーできているわけではない．いくつかの語については，大系に存在しないなどの理由により，置き換える対象となる上位語を見つけられず，置き換えが行われないこともある．今回，大系の固有名詞辞書は使用しなかったため，固有名詞や複合名詞については，置き換えが行われず，分析の対象からはずれてしまっている．しかしながら，この点については，十分な数の用例を用いることによって，他の用例でカバーするができると考えた．

直前語を上位語に置き換えた後は，上位語をさらに置き換えていく作業を繰り返す．最終的に 12 回の置き換え作業を行い，全ての語を最上位のカテゴリーである「名詞」に置き換える．置き換えが終了した後，記録してあった置き換え語の履歴をもとに，上位から順番に同じ上位カテゴリーにまとめた語を並び替え，統合していく．実際には，全ての語が最後に「名詞」カテゴリーに置き換えられているため，一つ前がどのカテゴリーに含まれていたかを調べる．大系では，「名詞」の下は「具体」と「抽象」の二つに分けられているため，全ての用例は，まずこの二つのカテゴリーに大別されることになる．以下，この作業を繰り返す．

最後に，それぞれのカテゴリーに含まれる用例数をカウントした．

3.3. 意味フレームの認定

ところで，シソーラスを用いて階層構造を構築する場合，どこまでを意味フレームとして認めるかという問題がある．大系を用いた場合，原理的には大系に載っている全ての名詞は，最終的に名詞に置き換えられることになる．

しかし，置き換えによって得られた全ての名詞によるフレームが存在するかどうかという問題が残る．意味フレームについての実在性を考慮した場合，フレームの構造については以下のような点を考慮する必要がある．

(4)　フレームには通常，非常に多くの情報が格納されているが，これはフレーム形成の初期段階において，乏しい内容しかなかったフレームの中核的要素に，一定の条件の下で「環境情報」がつけ加わった結果であると考えるべきである．

(5)　このようなフレームの変化の問題は，知識の成長の問題として捉えるべきである．

(6)　つまり，フレームの変化は，状況に関する知識の成長につれて起こると考えられる．

(7)　従って，意味フレームに関係する

「ありとあらゆること」がフレーム構造の構成要素である必要はないし，その可能性も低い．

例えば今回の場合，最終的に得られる「<名詞>が襲う」という意味フレームについて考えると，このような意味フレームを認めることにあまり意味はないと考えられる．そこで，どこまでを意味のあるフレームとして認めるかということを決めておく必要がある．

今回の置き換えにおいては，次の二つの意味フレームの定義を考慮した．

(8) X がフレーム構造をもつとは，X と非 X の区別を可能にする 特定の内部構造をもつことである．

(9) F(X) は，より大きなフレームに埋め込まれたり，他のフレーム群と結合したりすることで，拡張される．

そこで，(8) については，下位構造で実際の用例が存在することを条件とした．また (9) については，(8) の条件を満たすフレーム同士が統合される場合は，意味のあるフレームとして認めることで条件を満たそうと試みた．

4. 実験結果および考察

4.1. 実験結果

3. の実験によって得られた結果を図 2 に示す．この図では，二つ以上の下位項目が統合されたもののみを示している．また，項目が波線で囲まれているものは，階層ネットワークのためには必要であるが，先に述べた下位に用例を持たないフレーム同士の結合によって得られた，実際には意味をなさないフレームであると判断されたものを示している．

今回の手法では，以下に示す上位語にまとめられる語が「襲い手」として得られたことになる．

(10) 人間，仲間・成員・相手等，悪人等，病気，天災，気象，意向，感覚，感情，様相，組，知的生産物(思考・学習)

これらを図 1 で示した，黒田ら[5]が人手で分析した意味フレームと比べてみる．黒田らが最初の段階で得た意味フレームは以下の通りである．

(11) グループ間抗争/紛争，軍事侵略，資源強奪，虐待，強姦，動物の攻撃(捕食的)，動物の攻撃(非捕食的)，大規模な異常気象，小規模な異常気象，疫病の流行，活動への打撃，発病

黒田らの得た意味フレームを構成している「襲い手」の種類と，本手法で得られた「襲い手」の種類を比較すると，黒田らが分析した意味フレームとほぼ同様のものとなっており，なおかつ，それらのネットワークの分化も実現しているとみなすことができる．また，黒田らの分析にはなかった「感覚や感情」が襲うというフレームを読み取ることができ，かつ，そのフレームは，「抽象」下の「事」で分岐していることから，「病気」や「気象・天災」が襲うというフレームに近い物であることが予想される．

前者について，黒田らがこれらの「襲い手」ならびに，それらが実現する意味フレームを発見できなかった原因は，調査対象であるコーパスが新聞であることによる調査領域の狭さに起因するものと考えられる．

しかし，後者については，単にコーパスを調べるよりも効率的にフレームを発見でき，また他のフレームとの関係を浮かび上がらせるという点において本手法の有効性を示していると思われる．事実，中本ら[6]は心理実験などの結果から，「発病」フレームには，さらに細かい下位フレームが実在するとの判断を行っている．つまり，この手法を用い，どの階層で統合されるかを分析することによって，意味フレームの関係を決定する手がかりになると見なすことができる．

一方で，細かい粒度のフレーム分析はまだまだ不十分である．例えば，黒田らのカード分析課題でその違いが現れた，「動物への攻撃(捕食的)」と「動物への攻撃(非捕食的)」の違いといったものは，この手法では発見することが困難と思われる．なぜなら，大系では，これらの動物の間に区別は設けられていないからである．このように，大系では一つのカテゴリーにまとめられていても，実際にヒトが文を理解する際には，非常に細かい違いを明確に区別していることがある．そのため，意味フレームの全てについてシソーラスを用いて，自動構築することはまだまだ困難であることが予測される．しかし，現段階では，この手法のみで意味フレームの構築を完了するのではなく，この手法によってある程度意味フレームの階層ネットワークを特定した上で，分析を進めていけば目的は達成されるものと思われる．

今回の実験によって，未発見の意味フレームを発見出来た点を考慮すると，単にコーパスから用例を抽出するよりも，本手法が効率よく分析を行えると結論づけることができる．

5. おわりに

5.1. 今後の課題

今後は，これら自動的に得られた分類を基にして，機械学習の手法を用いて，意味フレームの特定に際して，より重要な意味役割を自動的に分類できることを目標とする．これにより，自動的に得られる意味フレームの特定精度をあげることを目指す．

また，多義語については，今回，なるべく重複を許すような形で置き換えを行ったが，これらについても，河原・黒橋[2]が格フレームの研究で行っているような，

類似度を利用した手法を用いることによって，有効でない分類を極力減らす方向を目指すといった課題が考えられる．

さらに理論的な問題として，格フレームと意味フレームの関係について，さらなる検討を行い，自然言語処理の分野でこれまでに得られた知見を最大限活用できる方法についても模索する必要がある．

5.2. まとめ

本稿では，意味理解のメカニズムを分析する上で、有力な手法である FOCAL の枠組みにおいて，効率的に階層意味フレームネットワークを特定する手法について提案した．現時点では，抽出された用例全てを人手によって分析しているため分析に時間がかかり，そのことで全ての意味フレームを特定出来ない可能性がある．しかし，分析の初期段階において本手法を用いることにより，意味フレームの範囲が特定できる可能性が示された．これにより，FOCAL の意味フレーム分析の速度を改善することが可能となり，効率的に分析を進めていくことの可能性が示された．

今後は，実験により明らかとなった提案手法の問題点を改善するとともに，意味フレームと格フレームとの関係についてもさらなる検討を行い，自然言語処理で得られた知見をできる限り FOCAL の分析に取り入れていくことを課題とする．

文　献

[1] C. J. Fillmore, C. R. Johnson, and M. R. L. Pentruck, Background to FrameNet, International Journal of Lexicography, Vol.16, No.3, pp.235-250, September 2003.

[2] 河原大輔，黒橋禎夫，“用言と直前の各要素の組を単位とする格フレームの自動構築，”自然言語処理，Vol.9，no.1，pp.1－16，Jan.2002.

[3] 黒田航，“意味フレームに基づく概念分析の射程: Berkely FrameNet and Beyond，”日本認知言語学会第 5 回大会発表予稿集，pp.134-137，2004.

[4] 黒田航，井佐原均，“日本語の意味タグ体系を定義する試み: FrameNet の視点から，”自然言語処理学会第 10 回大会発表論文集，pp.148-51，2004.

[5] 黒田航，中本敬子，野澤元，“状況理解の単位としての意味フレームの実在性に関する研究，”日本認知科学会第 21 回大会発表論文集，pp.190-191，2004.

[6] 中本敬子，野澤元，黒田航，“動詞「襲う」の多義性:カード分類課題と意味素性評定課題による検討，”日本認知心理学会第 2 回大会発表論文集，Vol.38，2004.

[7] NTT コミュニケーション科学研究所（編），岩波書店，日本語語彙体系 CD-ROM 版，東京，1999.

[8] 内山将夫，井佐原均，“日英新聞記事および文を対応付けるための高信頼性尺度，”自然言語処理，No.10，Vol.4，pp.201-220，July.2003.

[9] 宇津呂武仁，宮田高志，松本裕治，“最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価，”情報処理学会 自然言語処理研究会，97-NL-119，pp.69-76，May.1997.