Web上の定義文からの言い換え知識獲得

橋本 力 †　　鳥澤 健太郎　　ステイン・デ・サーガ　　風間 淳一　　黒橋 禎夫†
情報通信研究機構 MASTAR プロジェクト言語基盤グループ　† 京都大学情報学研究科

1　はじめに

言葉の意味と表現形式は多対多の関係であることが多く、ゆえに自然言語処理は challenging な領域であると考えられる。言い換え知識獲得は、同じ意味を持つ複数の異なる表現形式を認識／生成するための知識を獲得する技術である。本稿では Web 上の定義文からの言い換え知識獲得法を提案する。「言い換え」は両方向の含意関係が成立する表現対と定義する。同じ概念を定義する文は Web に大量に存在し、それらは言い換え関係にある場合が多く、それゆえ言い換え知識の宝庫と考えられる。(1) は「骨粗鬆症」の定義文対であり、下線 1 、2 で示された言い換え知識（同義句対）が含まれる。

(1) a. 骨粗鬆症とは、①骨の量が減り、②骨がもろくなってしまう病気だ。
b. 骨粗鬆症とは、①骨量が低下し、②骨が折れやすくなる病気だ。

本手法は、Web から定義文を自動獲得し、定義文対から同義句対を自動獲得する。実験から、提案手法により、Web6 億文書から適合率約 94%で同義句対約 30 万対を自動獲得しうることが示された。

2　関連研究

過去の言い換え知識獲得手法は、分布類似度によるもの [3] と単言語パラレルコーパスによるもの [1] に大別される。前者はパラレルでないコーパスからも知識を獲得できるが、十分な統計的情報が得られない低頻度表現に対する精度が概して低いこと、反義関係にある表現対に対しても分布類似度が高くなることが弱点である。

本手法は後者に属するが、このアプローチでは、(ほぼ) 同義関係にある文対から同義関係にある表現対を獲得するが、低頻度表現を含む同義対でも、文対に現れてさえいれば獲得できる。反義表現対を誤って獲得することも稀である。しかし弱点として、同義関係にある文対を大量に収集することが概して困難な点が挙げられる。我々は Web 上の大量の定義文を利用することでこれを克服した。後述するように、我々は F 値 0.91 で約 200 万の定義文（見出し語 867,321 語、定義文対 29,661,812 対）を Web から収集した。[4] も定義文を用いるが岩波国語辞典（63,000 語）と大辞林（233,000 語）のみを情報源とする（定義文対 57,643 対）。Web 全体を情報源としうる本手法と比べて scalability の差は明らかである。実際、[4] では適合率 74.8%で同義句対 500 対を獲得したが、本手法は Web6 億文書から適合率約 94%で同義句対約 30 万対を獲得しうる。

3　提案手法

本手法は、Web からの定義文獲得（3.1 節）と、定義文対からの同義句対獲得（3.2 節）の 2 段階から成る。

3.1　Web からの定義文の自動獲得

[5] は複数の言語パターンと事典から生成した言語モデルを用いて Web から定義文を獲得した。我々は、次のように、教師あり学習を用いて、より高精度に定義文を獲得する。**i)** Web6 億文書から「NP とは」で始まる文を収集する。**ii)** 収集された文の一部に定義文か否かのラベルを手作業で付与する。**iii)** ラベル付き文集合から、文を定義文か否かに分類する分類器を学習する。**iv)** 文集合から分類器によって定義文のみを獲得する。

i) の NP が定義される概念に相当する。ただし、「骨粗鬆症とは一体何ですか。」のように「NP とは」で始まる文全てが定義文とは限らない。そこで分類器により定義文か否かを区別する。学習には SVM(多項式カーネル, d=2) を用いた。学習データは、i) で収集した約 300 万文からランダムサンプリングした 2,911 文から作成した。そのうち 61%が正例だった。素性として、文末と「NP とは」直後の形態素 N グラムと Bag-of-words (ウィンドウサイズ N) を用いた。「NP とは」で始まる非定義文の特徴的表現が文末（「ですか。」）や「NP とは」直後（「一体何」）に現れやすいからである。形態素は表層形、原形、品詞のいずれかで表される。10 分割交差検定の結果、F 値は 0.91 だった。この分類器により、収集した約 300 万文から 1,925,052 の定義文を獲得した。さらに、Wikipedia 記事の第一文を定義文として追加することで、2,141,878 文に増量した。この中には「亥年現象」や「うねり取り」、「カイピリーニャ」、「イヴァリース」等、幅広い分野をカバーする計 867,321 の概念が含まれる。この定義文集合から、同じ概念の定義

文を対にすることで計 29,661,812 の定義文対を得た。

3.2 定義文対からの同義句対の自動獲得

定義文対からの同義句対獲得の手順は次の通りである。**i)** 定義文を KNP で係り受け解析して用言句を抽出する[1]。ただし用言句は、元の定義文で連続している 2 つ以上の文節（いずれも指示詞は含まない）から成り、末端の文節全てが体言を含むものに限る。**ii)** 定義文間で全ての用言句から用言句対を構成する[2]。(1) の場合、(1 骨の量が減り, 1 骨量が低下し)、(1 骨の量が減り, 2 骨が折れやすくなる)、(2 骨がもろくなってしまう, 1 骨量が低下し)、(2 骨が折れやすくなる, 2 骨が折れやすくなる) 等を構成する。**iii)** SVM（線形カーネル）により、用言句対を同義句対か否かで分類し、さらに分離平面からの距離で用言句対を順位づける。

分類器が使用する素性は、用言句自身が互いに類似しているか（表 1 の **f1-9**）、文脈が類似しているか（表 1 の **f10-17**）、あるいはその両方が成り立つ場合に同義句対になりやすい、という我々の観察結果に基づく。図 1

表 1: 用言句対分類器で使用する素性

f1	用言句対で共有される形態素数 ÷ 用言句対の全形態素数
f2	相手方の用言句に編集距離 1 の形態素が存在する形態素の数 ÷ 用言句対の全形態素数
f3	相手方の用言句に読みが同じ形態素が存在する形態素の数 ÷ 用言句対の全形態素数
f4	短い方の用言句の形態素数 ÷ 長い方の用言句の形態素数
f5	用言句対で主辞形態素の表層形が一致するなら 1、そうでなければ 0
f6	用言句対で主辞形態素の品詞が一致するかどうか。1 or 0
f7	用言句対で主辞形態素の活用型と活用形が一致するかどうか。1 or 0
f8	用言句 p_1 中の過剰形態素数 ÷ p_1の全形態素数（過剰形態素：p_1 と対をなす用言句 p_2 内には無いが、p_2 の文脈には存在する形態素）
f9	**f8** の p_1 と p_2 を入れ替えたもの
f10	用言句対で共有される親文脈係り受け部分木の数 ÷ 用言句対の親文脈係り受け部分木の総数（部分木は形態素の読みで表される。）
f11	部分木が形態素の原形で表される以外、**f10** と同じ。
f12	部分木が形態素の品詞で表される以外、**f10** と同じ。
f13	用言句対で共有される左文脈の形態素数 ÷ 用言句対の左文脈の全形態素数　（形態素は読みで表される。）
f14	形態素が原形で表される以外、**f13** と同じ。
f15	用言句対の両用言と隣接している左文脈形態素数 ÷ 用言句対の左文脈の全形態素数　（形態素は原形で表される。）
f16	用言句対で共有される左文脈 Trigram の数 ÷ 用言句対の全左文脈 Trigram 数　（Trigram は形態素の読みで表される。）
f17	用言句対の抽出元の定義文対の Cosine 類似度

で素性 **f8-12** の補足説明をする[3]。図の場合、p_2 に、p_1 には無いが p_1 の子文脈には出現する形態素「骨」「の」があるので、**f9** は正の値を取る。**f8** は 0 である。また、用言句対 (p_1, p_2) は親文脈部分木 (骨が 折れやすくなる) を共有しており、**f10-12** は正の値を取る。

[1] 1 つの用言に対して複数の用言句が抽出されうる。例えば (1a) の「減り」の場合、「量が減り」と「骨の量が減り」が抽出される。

[2] 文字列上包含関係にある対と、違いが 1 つの固有名詞のみの対 (例えば (Apple で製造した,Xerox で製造した)) は無視する。

[3] 説明のため (1) とは違う文対を例に挙げた。

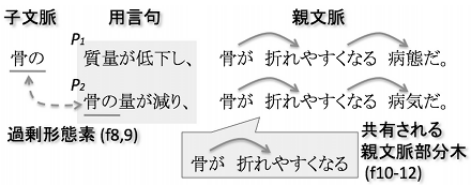


図 1: 素性 **f8-12** の図解

我々は次の素性群も試した；文脈類似語と動詞含意知識を用いた用言句対の類似度；主辞形態素の読み／原形の一致；表層形、読み、原形、品詞のいずれかで表された左／右文脈 N-gram (N=1,2,3) の共有；表層形、読み、原形、品詞のいずれかで表された親／子文脈部分木の共有；両方の用言句に隣接している左／右文脈 N-gram、親／子文脈部分木の共有。結局全部で 78 素性を試したが、個々の素性の性能を測る ablation test を通して、表 1 の素性が最終的に選ばれた。

ablation test に用いた学習データは我々が独自に構築したものである。学習データを構築する際、完全なランダムサンプリングだと正例がほとんど集まらないという問題に直面した。そこで、正例の可能性が高い用言句対をあらかじめ自動で収集した。具体的には、我々が試した全 78 素性の値の合計が高いものを収集した。これらの素性値が高いほど用言句対が正例である可能性が高いからである。ただし **f8,9** は、値が高いほど負例である可能性が高いので、値を -1 で重み付けした。結局、全定義文対から 3 万ペアをランダムサンプリングし、その中から素性値合計の最も高い用言句対 3,000 を得た。

次に、用言句対 (p_1, p_2) を同義句対かどうか人手でチェックした。具体的には、p_1、p_2 を文脈に埋め込んだ上で、$p_1 \rightarrow p_2$、$p_2 \rightarrow p_1$ の両方向の含意が成立するかをチェックした。本研究では両方向の含意が成立するものだけを同義句対と見なす。p_1 を埋め込む文脈として、p_2 が得られた元の定義文 s_2 を用いた。一方、p_2 の文脈は p_1 の出所である定義文 s_1 を用いた。つまり、まず定義文対 (s_1, s_2) の間で用言句対 (p_1, p_2) を入れ替え、新たな定義文対 (s_1', s_2') を生成し、次に $s_1 \rightarrow s_1'$、$s_2 \rightarrow s_2'$ の含意関係をチェックすることで $p_1 \rightarrow p_2$、$p_2 \rightarrow p_1$ の両方向の含意をチェックした。図 2 の例では「**骨の量が減り**」と「**骨量が低下し**」が p_1、p_2 だが、$s_1 \rightarrow s_1'$ が成立するので $p_1 \rightarrow p_2$ が成立し、$s_2 \rightarrow s_2'$ が成立するので $p_2 \rightarrow p_1$ が成立する。結局、1,092 の正例と 1,872 の負例を得た。残りは文字化けを含むため除外した。

4 評価実験

提案手法が定義文から同義句対を高精度で獲得できること（4.1 節）と、定義文が同義句対獲得源として優れていること（4.2 節）を示す。人手評価の Fleiss' Kappa 値は 0.68 で、一致率が高いことを示している。

用言句対獲得元の定義文対	用言句を入れ替えた（言い換えられた）定義文対
s_1: 骨粗鬆症とは、**骨の量が減り**、骨がもろくなってしまう病気だ	s'_1: 骨粗鬆症とは、**骨量が低下し**、骨がもろくなってしまう病気だ
s_2: 骨粗鬆症とは、**骨量が低下し**、骨が折れやすくなる病気だ	s'_2: 骨粗鬆症とは、**骨の量が減り**、骨が折れやすくなる病気だ

図 2: 用言句対の双方向含意関係のチェックの例

4.1 提案手法 vs. 既存手法

本節では、パラレルコーパスから同義句対を高精度に獲得することで有名な [1]（以下、BM 法）と [2]（以下、SMT 法）との比較を通して、提案手法が定義文から同義句対を高精度に獲得できることを示す。

BM 法は文対が共有する語を同義語対（正例）と見なし、その文脈（N-gram）を両文から抽出して正例文脈対を抽出する。一方、異なる 2 つの語を負例と見なし、同様に負例文脈対を抽出する。文脈対ごとに、全出現回数のうち正例（負例）文脈対として用いられる回数から正例（負例）文脈対としてのスコアを求め、スコア上位 K の文脈対を用いて同義句対の正例（負例）を獲得する。獲得された正例（負例）により新たな正例（負例）文脈対を得て、さらに正例（負例）を獲得する。上記の処理を最大 T 回繰返す。出力される同義句対にスコアは付与されない。我々は [1] に倣い、N を 1〜3、K を 10 とした。T は論文中に無かったので、我々の予備実験に基づき 3 とした。

SMT 法では、Moses を用いて定義文対から同義句対のフレーズテーブルを構築する。パラメータはデフォルトのままである。同義句対は、Moses が出力する両方向の句翻訳確率の積により順位づけられる。

教師無しの手法である BM、SMT の両手法と比較するため、提案手法の教師無し版も用意した。以後、教師ありの提案手法を *Sup*、教師無しの方を *Uns* とする。*Sup* と *Uns* の違いは、前者が分離平面からの距離で同義句対を順位づけるのに対し、後者は素性値の合計によって順位づける点である。

全定義文対からランダムサンプリングした 10 万対から上記 4 手法により同義句対（用言句対のみ）を獲得した。同義句対のスコア上位 5,000 から 200 をランダムサンプリングして著者以外の 3 名で評価した。BM 法は同義句対を順位づけないため、出力全体からランダムサンプリングした。評価は、3.2 節の学習データ構築の時と同様、同義句対 (p_1, p_2) を文脈に埋め込んで上で行った。ただし、文脈として獲得元の定義文ではなく、p_1 を含む文 s_1、p_2 を含む文 s_2 を Web から取得して文脈とした点が異なる。s_1、s_2 のいずれか一方でも Web から取得できない (p_1, p_2) は評価対象外とした。s_1、s_2 はともに p_1、p_2 の獲得元の定義文とは異なる。2 人以上が両方向の含意が成立すると判定した場合に正解とした。

図 3(a) に各手法の適合率曲線を挙げる。このグラフから、*Sup* の適合率が他の手法を上回り、上位 1,000 付近で約 94%を示すことがわかる。これが定義文 10 万対から得られた結果であることから、全定義文対 29,661,812 対（つまり Web6 億文書）から適合率約 94%で約 30 万の同義句対を獲得しうると推定できる。

表 2: 獲得した同義句対の数

	定義文対				Web 文対			
	Sup	Uns	BM	SMT	Sup	Uns	BM	SMT
trivial あり	1,381,424		24,049	9,562	277,172		5,101	4,586
trivial 無し	1,377,573		23,490	7,256	274,720		4,399	2,342

表 3: 提案手法 *Sup* で獲得した同義句対の例

順位	同義句対
70	企業の財政状況を表す ⇔ 企業の財政状態を示す
112	インフォメーションを得る ⇔ ニュースを得る
656	きまりのことです ⇔ ルールのことです
1,553	角質を取り除く ⇔ 角質をはがす
2,243	胎児の発育に必要だ ⇔ 胎児の発育成長に必要不可欠だ
2,855	視力を矯正する ⇔ 視力矯正を行う
2,931	チャラにしてもらう ⇔ 帳消しにしてもらう
3,667	ハードディスク上に蓄積される ⇔ ハードディスクドライブに保存される
4,870	有害物質を排泄する ⇔ 有害毒素を排出する
5,501	１つのＣＰＵの内部に２つのプロセッサコアを搭載する ⇔ １つのパッケージに２つのプロセッサコアを集積する
10,675	外貨を売買する ⇔ 通貨を交換する
112,819	派遣先企業の社員になる ⇔ 派遣先に直接雇用される

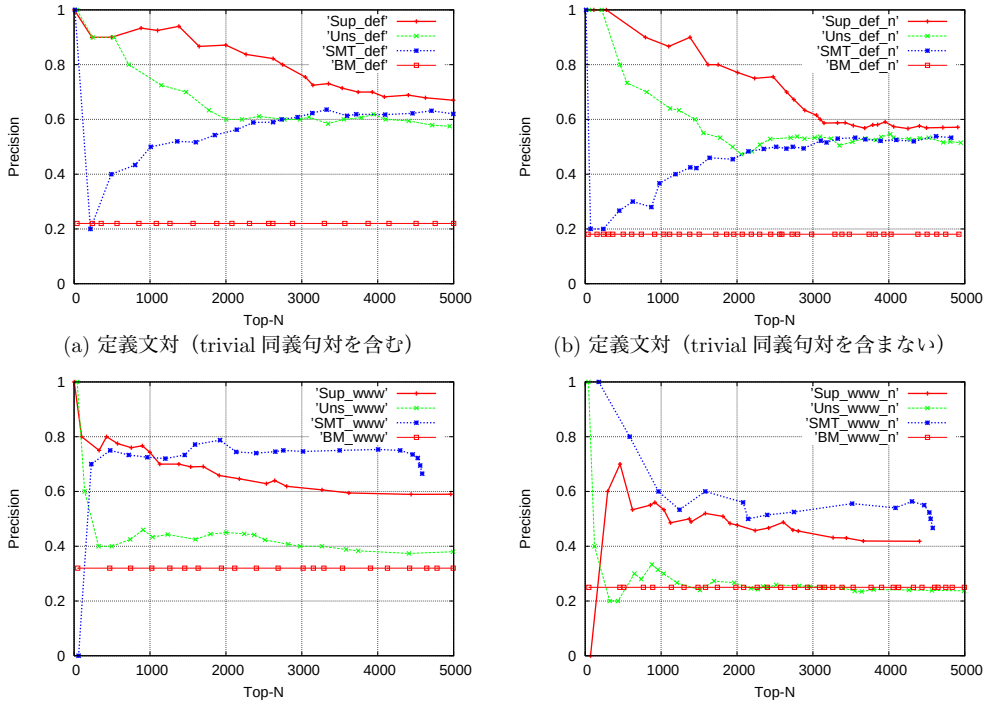
さらに、読みが同じか、内容語が全て同じ同義句対を trivial なものと見なし、それらを除外した結果を図 3(b) に挙げる。この場合も *Sup* の適合率が他を上回ること、上位 1,000 付近まで適合率約 90% を保つことがわかる。以上の結果は、提案手法が定義文対から同義句対を高精度で獲得できることを示している。

表 2 左に各手法の同義句対獲得数を挙げる。*Sup* と *Uns* の獲得数が他より多いことがわかる。*Sup* が適合率でも獲得数でも他を上回ることは注目に値する。

表 3 に *Sup* により獲得した同義句対の例を挙げる。例にある通り、獲得した同義句対のほとんどは定義文に特化したものではなく、広く再利用可能なものだった。

4.2 定義文対 vs. Web 文対

単純な手法で Web から収集した意味的に類似している文対（Web 文対）と定義文対を比較することで、後者が同義句対獲得源として優れていることを示す。Web 文対は次の手順で集めた。まず、Web6 億文書からランダムサンプリングした 180 万文（サンプル文）の内容語名詞により Web 検索する。次に、検索結果のスニペッ



(a) 定義文対（trivial 同義句対を含む）　(b) 定義文対（trivial 同義句対を含まない）

(c) Web 文対（trivial 同義句対を含む）　(d) Web 文対（trivial 同義句対を含まない）

図 3: 各手法の適合率曲線

ト群を文（スニペット文）に分解する。元のサンプル文と最も類似するスニペット文をペアにする。類似度は共有している内容語名詞の数で表す。この結果から 10 万対をランダムサンプリングして Web 文対とした。

4.1 節と同じ 4 手法で Web 文対から同義句対を獲得した。 *Sup*、*Uns* で使用する素性は 3.2 節で述べた 78 素性から Web 文対用に選択し直した。具体的には、Web 文対と同じ方法で集めた 2,741 文対から学習データを作り、それを用いて ablation test により素性を選択した。

図 3(c) に各手法の適合率曲線を挙げる。4.1 節と同様 trivial 対を除いた場合も評価した。その結果を図 3(d) に挙げる。表 2 右に獲得数を挙げる。

提案手法の適合率が定義文対の場合と比べて低いこと、Web 文対を対象とした場合どの手法も適合率 90%には至らないことがわかる。自動獲得結果をそのまま QA 等のタスクに用いる場合、適合率 90%は欲しいと考えるが、その基準を満たすのは *Sup* と定義文対の組合せのみである。また、どの手法の場合も、Web 文対より定義文対の方が獲得数が多い。以上の結果は、定義文対が同義句対獲得源として優れていることを示している。

5 結論

我々は Web 上の定義文からの同義句対獲得法を提案した。実験結果から、定義文は同義句対の宝庫であること、提案手法がそれらから高精度に同義句対を獲得すること、提案手法により Web6 億文書から適合率約 94%で約 30 万の同義句対を獲得しうることが示された。

今後、定義文対 29,661,812 対から獲得した同義句対を ALAGIN フォーラム (`www.alagin.jp`) から配信する。

また、同じ概念についての複数の定義文以外にも、同じ UNIX コマンドの複数の解説文や、同じ料理についての複数のレシピ文、同じ研究課題の関連研究についての複数の記述など、**同じ対象に対して同じ機能を果たす複数の文**を収集して、本手法を適用する予定である。

参考文献

[1] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the ACL joint with the 10th Meeting of the European Chapter of the ACL (ACL/EACL 2001)*, pp. 50–57, 2001.

[2] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 177–180, 2007.

[3] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[4] 村田真樹, 金丸敏幸, 井佐原均. 複数の辞書の定義文の照合に基づく同義表現の自動獲得. Vol. 11, No. 5, pp. 135–149, 2004.

[5] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.