



HOKKAIDO UNIVERSITY

Title	WordNet及びWeb検索による常識的知識ベースのための文生成手法
Author(s)	村本, 晃一; Rzepka, Rafal; 荒木, 健治
Citation	ことば工学会, 35, 1-7
Issue Date	2010-08-27
Doc URL	https://hdl.handle.net/2115/63985
Type	journal article
File Information	WordNet.pdf



WordNet 及び Web 検索による 常識的知識ベースのための文生成手法

Using WordNet and Web Search for Generating Sentences

for Common Sense Knowledge Bases

村本 晃一 ジェプカ ラファウ 荒木 健治
Koichi Muramoto Rafal Rzepka Kenji Araki

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

1. はじめに

コンピュータに知的な言語処理を行わせるには、人間の持つ常識や知識といったものが必要不可欠である。そのため、コンピュータに人間の持つ常識や知識を与えるという研究が行われている[1][2]。これらは知識の対象や、質問に返答するなどの利用方法を限定することで、人手によるデータベースの作成や、知識源からの抽出を容易なものとしている。しかし、このような知識の対象や利用方法の限定を行うことは、人間の持つ知識が広い範囲にわたることに相反している。そこで、対象を限定としない常識をデータベース化することができれば、常識を用いた様々な研究に有効に利用することが可能となると考えられる。このような常識的知識をデータベース化したものとして、ConceptNet[3][4]が挙げられる。ConceptNet とは、Open Mind Common Sense[5]を用いた常識的知識の生成プロジェクトであり、単語がノードによって関係づけられているものである。ConceptNet の例を図 1.1 に示す。これらは、MIT のプロジェクトで、現在も研究がなされている。これらを用いた研究は、英語では行われている[6]が、日本語ではあまり盛んには行われていない。これは、Open Mind Common Sense に英語では 1,029,518 文もの常識的知識を表現する文が登録されているのに対し、日本語では 14,499 文しか登録されていないことが原因であると考えられる。この常識的知識を表現する文の量を増やすことができれば、数多くの研究に利用することができると考えられる。

そこで本論文では、この Open Mind Common Sense の文を基に、日本語の概念辞書である日本語 WordNet[7]と Web 検索を用いて多くの常識を含む文を生成する手法を提案する。また、生成した常識的知識の利用方法として、ConceptNet を用いた発話システムを提案する。

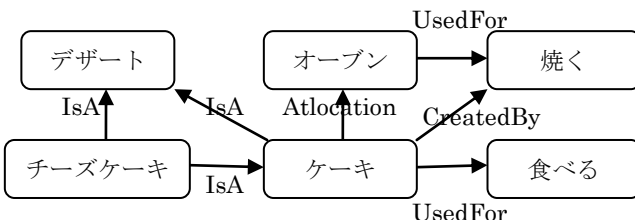


図 1.1 ConceptNet の例

2. 常識的知識を表現する文の生成

2.1 常識的知識

例えば、「犬は歩く。」は常識であるが、「東大寺は日本の寺の具体例だ。」は日本についての教育を受けた者にとってのみ常識となる。本論文では、このような、人間が

生活していく上で獲得する常識に加え、文化的な知識や専門的な知識も含む知識を常識的知識と呼ぶ。この常識的知識を表現する文の生成を行う。

2.2 全体の概要

入力に Open Mind Common Sense の文を、文の生成に日本語の概念辞書である日本語 WordNet を、生成文の正誤判定に Web 検索を用いている。システムの概要図を図 2.1 に示す。

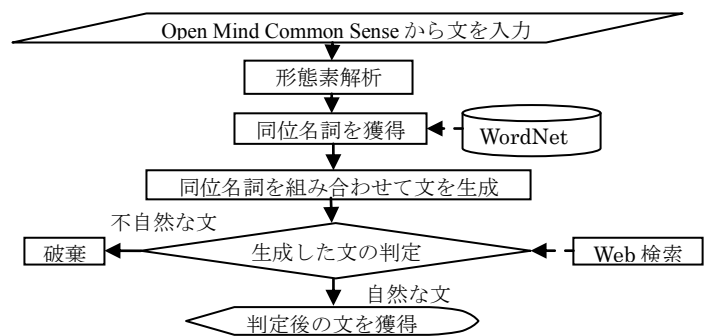


図 2.1 常識的知識を表現する文の生成システム概要図

2.3 日本語 WordNet

WordNet とは、プリンストン大学で開発された Princeton WordNet のことで、単語に synset と呼ばれる概念が付与されている意味辞書である。類似した単語にはそれぞれ同じ synset が付与されている。synset は「ID 番号-品詞」という形で示されており、品詞部分には例えば名詞なら「n」、動詞なら「v」、形容詞なら「a」、副詞なら「r」などのように、アルファベットが割り振られている。またその synset 同士は上位概念や下位概念等数多くの関係によって関連付けられている。今回は、この Princeton WordNet の synset に対して日本語を付与した日本語 WordNet を用いた。日本語 WordNet には、56,741 の synset、92,241 語、synset と単語のペアが 157,398 組登録されている。日本語 WordNet に含まれている単語と関係の例を表 2.1、表 2.2 に示す。

家					
synset	synset に定義されている単語				
03544360-n	家屋	住宅	住居	ハウス	館
08078020-n	家族	世帯	家庭	ファミリー	一家

表 2.1 WordNet に登録されている単語の例「家」

家					
上位概念	建築物	ビル	建造物	ビルディング	建物
下位概念	お屋敷	小屋	平屋	下宿	宿屋
一部	図書室	書斎	ロフト	屋根裏部屋	書房

表 2.2 WordNet に登録されている関係の例「家」

2.4 同位名詞獲得

基となる文を MeCab で形態素解析し、名詞を抽出する。その名詞について WordNet を用いて同位名詞を獲得する。ここでは、同位名詞を「ある名詞の上位概念の下位概念にあたる名詞」と定義する。例えば、「りんご」という名詞の上位概念には「果物」があり、「果物」の下位概念には「サクランボ」や「桃」がある。よって、「りんご」の同位名詞は「サクランボ」や「桃」となる。このように、上位概念の下位概念を参照することで、同位名詞を獲得していく。このように獲得した同位名詞を用いることで常識的知識を表現する文を生成する。同位名詞獲得の例を図 2.2 に示す。ここで、同位名詞に着目した理由について説明する。人間は、例えば「りんごを食べる。」、「桃を食べる。」という文が正しいものであると認識した場合、同じ果物の一種であるサクランボも食べることができるのではないかと推測する。この推測の際の、りんごや桃からサクランボを連想することが、同位名詞を獲得する操作に相当する。これが、同位名詞を基に常識的知識を表現する文を生成する理由である。ただし、非自立名詞と、「人、物、影響、原因、結果、一つ、一部、」という名詞については同位名詞の獲得を行っていない。非自立名詞は単独では意味を持たないため、「人、物、影響、原因、結果、一つ、一部」は前述の推測の対象にはならないと考えられるためである。

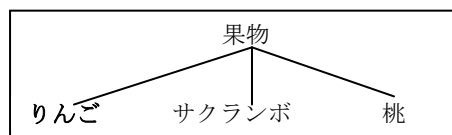


図 2.2 同位名詞獲得の例

2.5 文の生成

2.4 で述べた手法で獲得された同位名詞を組み合わせて、文を生成する。表 2.3 に獲得される文の例を示す。しかし、このようにして獲得した文には、常識的知識とは異なる、常識的には不自然な知識が非常に多く含まれている。表 2.3 の例でいえば、「すべり台を投げる」等がこれに相当する。これらの常識的には不自然な知識をノイズとする。2.7 で詳しく述べるが、生成した文の約 55% がノイズと判断された。これらのノイズを除去するために、Web 検索を利用する。詳しくは 2.6 で述べる。

速球を投げる	人形を投げる。
おもちゃを投げる。	フリスビーを投げる
球を投げる。	すべり台を投げる。

表 2.3 獲得される文の例
【基となる文：ボールを投げる。】

2.6 ノイズの除去

同位名詞を組み合わせることで獲得された文から、ノイズを除去する。ノイズを除去するには、辞書やデータベースを利用する方法があるが、データの量が膨大なため、人手では辞書やデータベースの作成を行うことが困難である。そのため、Web 検索を利用する。Web 検索を利用する利点として、まず、データの豊富さが挙げられる。Web 上には、様々な話題について様々な人が記述したテキストデータが豊富に存在している。それらのデータを一種の知識源として扱うことで、辞書やデータベースの作成を行う手間を省くことができる。また、Web 上の情報は、日々刷新されていくためデータの更新を行う必要もない。Web 検索には Yahoo! JAPAN¹を利用する。ノイズ除去の概要図を図 2.3 に示す。

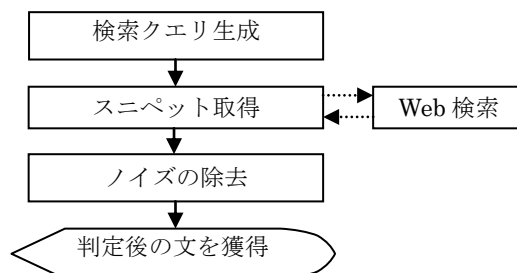


図 2.3 Web 検索による生成文の自然性判定の処理過程

まず、検索クエリを生成する。生成文から名詞・形容詞・動詞を抽出し、名詞に生成文と同様の助詞を付加した上で、「名詞助詞」「動詞(形容詞)」というような検索クエリを生成する。日本語では、助詞によって動作の主格や目的格が大きく変わってくる。例えば、「ボールで遊ぶ」と「ボールが遊ぶ」では、助詞以外の部分は同じでも、意味が全く違うものとなる。このような日本語における助詞の重要性については、これまでも述べられている[9]。そのため、文の自然性を判定するには、助詞が必要である。例えば、「ボールを投げる」という文から「ボールを」「投げる」という検索クエリを生成し Web 検索を行う。このとき、先ほど述べた同位名詞の獲得を行わなかった名詞については、検索クエリに含んでいない。

このように生成した検索クエリを用いて Web 検索を行い、検索結果ページのスニペットを基にノイズの除去を行う。スニペット中の一文に、生成した検索クエリが同時に同順序で出現する回数を算出する。この回数を以後、「スニペット共起回数」と呼ぶ。一文とは、句点、感嘆符、疑問符、ドット(…)で挟まれている部分のことである。スニペット共起回数について閾値を設定し、スニペット共起回数が閾値を下回った場合、ノイズと判定する。スニペット共起回数を算出する際、図 3.4 のように検索クエリとした名詞や動詞、形容詞が連続している場合だけでなく、離れていても一文中に存在していれば、どちらもノイズではないと判定する。これは、検索クエリとした名詞や動詞、形容詞の間に別の単語を含んでいても、一文中に存在していれば、これらの関係は連続している場合と変わらないと考えられるからである。

¹ Yahoo! JAPAN <http://www.yahoo.co.jp>

2.7 ノイズの除去の閾値設定のための予備実験

2.6 で述べたように、ノイズを除去するためには、閾値を設定する必要がある。この閾値設定のために、以下の予備実験を行った。

Open Mind Common Sense の文から名詞一つを含む文をランダムに 50 文選択する。選択する文の例を表 2.4 に示す。それらに対して同位名詞を獲得し、文を生成することで、13,240 文を得た。そのうち、スニペット共起回数が一回以上の文は、2,820 文あった。この 2,820 文から 1,000 文をランダムに選び、著者が評価を行った。評価基準は「0 点：不自然」、「1 点：有り得るが普通ではない・場合によっては有り得る」、「2 点：自然」の 3 種類である。

人は葬式のときに惜しむ。	花は人を和ませる。
電車は人を運ぶ。	鉛筆で書く。
アルコールは燃える。	人は名声が欲しい。
朝食中に、食べる。	葬式は悲しい。
ライオンは吠える。	鳥は飛ぶ。

表 2.4 Open Mind Common Sense から選んだ文の例

基となる文	生成された文
鉛筆で書く。	クレヨンで書く。
	万年筆で書く。
	ペンで書く。
ライオンは吠える。	ジャガーは吠える。
	虎は吠える。
	チーターは吠える。

表 2.5 生成された文の例

結果を表 2.3 に示す。1,000 文中、著者が不自然と判断した文が 545 文で 54.5%、場合によっては有り得ると判断した文が 211 文で 21.1%、自然と判断した文が 244 文で 24.4%だった。また、それぞれのスニペット共起回数の平均は 7.3 回、11.6 回、16.4 回となった。そこで、スニペット共起回数が 11 回以下と算出された文を「不自然」、11 回以上 16 回以下と算出された文を「場合によっては有り得る」、17 回以上と算出された文を「自然」とするように、閾値を設定した。

著者の評価	文数	スニペット共起回数の平均
0 点	545 文/1,000 文	7.3 回
1 点	211 文/1,000 文	11.6 回
2 点	244 文/1,000 文	16.4 回

表 2.3 閾値の設定のための実験の結果

生成した文のうち、0 点とされた文は 54.5%、1 点とされた文は 21.1%、2 点とされた文は 24.4%となった。著者の評価が 0 点の文でもスニペット共起回数が必ず 0 回というわけではない。これには理由がある。まず、検索クエリとなった単語の前後の単語によって、自然な表現になったり不自然な表現になったりする場合がある。例えば、「昔は赤い。」という文は常識的ではないが、「昔は赤い橋が架かっていた。」のような文が Web 中に存在した場合、システムは正しいと判断する。また、「水は燃える。」という文は、「

水は” “燃える”」が検索クエリとなり、これらはスニペット中で「水は燃えることはない。」のように、否定文として Web 中に存在することがある。このような例に対処する手法は、今後の課題である。

2.8 評価実験

2.7 で設定した閾値を用いてノイズを除去した文について、その文の自然性を被験者が評価する実験を行った。2.6 で用いた 2,820 文のうち、閾値の設定実験に使用した 1,000 文を除く 1,820 文に、設定した閾値によって 0 点、1 点、2 点の評価を与えた。それぞれの点数毎にランダムに 50 文ずつを抽出し、計 150 文を用いて本実験を行う。それぞれについて被験者が 0 点、1 点、2 点の評価を行い、評価した。被験者は全員 20 代学生で、男性 4 名女性 2 名の計 6 名だった。

実験結果を表 2.2 に示す。システムの評価と各被験者の評価は表 5.3 のようになった。システムが 2 点と評価した文の正解率が平均 55.0%、システムが 1 点と評価した文の正解率が平均 28.0%、システムが 0 点と評価した文の正解率が平均 48.6%となった。また、被験者の 3 人以上または 4 人以上の評価が一致した文についての結果を、表 2.3 に示す。3 人以上の評価が一致した文では、システムが 2 点と評価した文の正解率が 62.5%、システムが 1 点と評価した文の正解率が 15.9%、システムが 0 点と評価した文の正解率が 55.6%であった。4 人以上の評価が一致した文では、システムが 2 点と評価した文の正解率が 71.4%、システムが 1 点と評価した文の正解率が 13.2%、システムが 0 点と評価した文の正解率が 67.7%であった。

システムの評価	被験者の評価	被験者平均
2 点	2 点	27.5 文
	1 点	12.5 文
	0 点	10.0 文
	正解率	55.0%
1 点	2 点	21.0 文
	1 点	14.0 文
	0 点	15.0 文
	正解率	28.0%
0 点	2 点	14.7 文
	1 点	11.0 文
	0 点	24.3 文
	正解率	48.6%

表 2.2 人手による評価

2.7 で述べたように、ノイズの除去を行わなかった場合、2 点とされた文は 24.4%であり、ノイズ除去を行ってシステムが 2 点と評価した文の正解率は 55.0%であるので、30.6 ポイントの精度の向上がみられた。しかし、システムが 0 点や 1 点と評価した文にも人手の評価では 2 点となる文が存在している。また、常識的か否かの評価は、被験者によって異なるため、評価が分かれることもあった。3 人以上の評価が一致しなかった文は 150 文中 21 文、4 人以上の評価が一致しなかった文は 150 文中 46 文だった。表 2.2、表 2.3 のどの結果においても、1 点とされた文の正解率が低くなっている。これは、評価を行う際に、2 点と 0 点の差は多くの人で一致していても、2 点と 1 点、1 点と 0 点の差が非常に曖昧なものであるためだと考えられる。

システムの評価	被験者の評価	3人以上	4人以上
2点	2点	25文	25文
	1点	5文	5文
	0点	10文	5文
	正解率	62.5%	71.4%
1点	2点	22文	20文
	1点	7文	5文
	0点	15文	13文
	正解率	15.9%	13.2%
0点	2点	12文	8文
	1点	8文	2文
	0点	25文	21文
	正解率	55.6%	67.7%

表 2.3 被験者の 3 人または 4 人以上の評価が一致した文

また、システムが 0 点と評価し、被験者 4 人以上が 2 点とした文は 8 文存在した。例えば、「夕暮は赤い。」は、「夕暮れは赤い。」であれば Web 中に存在している。このように、表記の揺れによって同じ意味でもシステムの判断が異なる場合がある。「ヒトはねだる。」、「夕刻は赤い。」についても同様のことが言える。システムが 2 点と評価し、被験者 4 人以上が 0 点とした文は 5 文存在した。このうち、「人は運転の際に眠くなる。」という文から生成された「人は止めの際に眠くなる。」という文は、「止めの」「眠くなる」が検索クエリとなるため、「鼻水止めの薬は眠くなる」といった文を抽出し、自然であると判断する。つまり、検索クエリが短いため、前後に別の単語が付与される可能性がある。「先は赤い。」、「人は感じが欲しい。」、「界は冷たい。」にも同様のことが言える。「人は絶望が欲しい。」は、被験者 4 人が不自然とした文であるが、小説や、歌詞中に登場することがあるため、Web 中に存在しシステムが自然だと判断した。

システムと被験者の評価が一致しなかった文が一致しなかった原因の内訳を図 2.4 に示す。



図 2.4 システムと被験者の評価が一致しなかった文の原因

2.9 常識的知識を表現する文の生成に関するまとめ

2 章では、日本語 WordNet を用いて Open Mind Common Sense に含まれる文の名詞について同位名詞を獲得することで、常識的知識を表現する文を生成し、Web 検索を用いて常識的ではない文を除去する方法について述べた。評価実験の結果、システムが自然であると評価した文の正解率は 55.0% で、ランダムで選択した場合と比べ、30.4 ポイントの精度の向上が確認された。しかし、これはまだ十分な精度ではなく、今後精度を向上させていく必要がある。今後の課題をいくつか挙げる。まず、今回は名詞を一つ含む文のみに着目したが、名詞が複数個含まれる文について

も同じ処理が有効かを検証する必要がある。動詞や形容詞などの、名詞以外の品詞についても同様である。また、今回は常識的ではない文をノイズとして除去する方法を述べたが、このノイズも常識的知識を表現する文ととらえることができる。例えば、「水は燃える。」はノイズであるが、システムがこの文を 0 点と評価することができれば、否定文「水は燃えない。」に変換するだけで常識的知識を含む文となる。また、ノイズ除去の精度を向上させる方法をいくつか挙げる。まず、検索クエリを設定する際に、「名詞 + 助詞」を検索クエリとしたが、助詞は基となる文の助詞をそのまま使用している。しかし、助詞を変更することにより精度を向上させることが考えられる。例えば「鳥は飛ぶ。」という文において、「鳥は」だけでなく、「鳥が」を検索クエリとすることで、より多くの文を収集することができると考えられる。「は」を「に」に変える等、意味が大きく変わる助詞の変更もあるので、それぞれの助詞の関係を考慮する必要がある。次に、スニペットを用いたノイズの除去について、2.7 で述べたように、前後の単語によっては自然さが変化する文もある。一文中に共起するかどうかのみではなく、前後の関係を考慮することで精度の向上が見込める。さらに、現在は検索クエリ同士の距離を考慮していないので、それを考慮することによっても、精度を向上させることが可能であると考えられる。また、閾値を決定する実験を著者のみでなく、多くの被験者で行うことも精度向上には重要である。

3. 常識的知識の有用性

2 章では常識的知識を表現する文の生成について述べた。本章では、常識的知識の利用方法とその有用性について述べる。常識的知識の

3.1 ConceptNet を用いた連想発話手法

常識的知識の有用性の検証のために ConceptNet を用いた連想発話手法を提案する。概要図を図 3.1 に示す。1 章でも述べたように、ConceptNet は単語がノードによって関連付けられているものである。そこで、この関係を用いて連想を行い、発話を生成する。

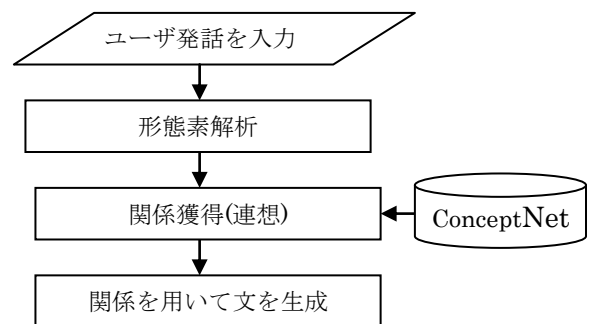


図 3.1 連想発話手法の概要図

3.2 関係の獲得

入力文の名詞を ConceptNet に入力し、関係する単語と関係を獲得する。例を表 3.1 に示す。例えば、「犬を飼っている。」という入力文から「IsA(犬, 動物)」, 「SymbolOf(犬, 忠誠心)」, 「CapableOf(犬, 吠える)」等の関係を獲得する。これらから各関係毎に著者が作成したテンプレートに基づき、文を生成する。生成される文の

例を表 3.2 に示す。例えば、「IsA(犬, 動物)」から「犬は動物だよな。」, 「SymbolOf(犬, 忠誠心)」から「犬は忠誠心の象徴だね。」を生成する。

IsA(犬, 動物)	SymbolOf(犬, 忠誠心)
CapableOf(犬, 吠える)	AtLocation(犬, 家)
HasProperty(犬, 機敏)	Desires(犬, 骨)

表 3.1 獲得される関係【基となる文：犬を飼っている。】

犬は動物だよな。
犬は忠誠心の象徴だね。
犬は吠えることができるんだね。
犬は家で見かけるよな。
犬は機敏だよな。
犬は骨を欲しがるとだよな。

表 3.2 生成される文【基となる文：犬を飼っている。】

3.3 評価実験

3.2 で提案した手法について、ConceptNet 内にある関係のうち、どの関係が発話に適しているか、また逆に適していないかを判断するために評価実験を行った。ConceptNet 内にある名詞をランダムに選択し、入力文の名詞とする。入力文は「名詞助詞動詞」という形で著者が作成した。助詞と動詞については Web から自動構築された大規模格フレームシステム[10]を用いて、名詞に対して使用される頻度が最も高いものを選択した。大規模格フレームシステムとは、用言とそれに関係する名詞を用言の用法ごとに整理したものである。これは、Web 上の約 16 億文の日本語テキストから自動的に構築されており、約 4 万用言からなるものである。作成した入力文をシステムに入力し、連想を 1 度行うことで出力文を得る。結果として、14 の関係について各 2 文ずつ、計 28 文が獲得された。14 の関係とその意味を表 3.3 に、作成した入力文と、それに対するシステムの出力文の例を表 3.4 に示す。

これらの入力文と出力文について、「文法的な自然さ」、「意味的な自然さ」、「システムの語彙の豊富さ」、「システムの知識量」の 4 項目を、5 を最良とする 5 段階で被験者が評価する実験を行った。被験者は 5 名で、全員 20 代男子理系大学生である。被験者 5 人の平均の結果を表 3.5 に示す。

関係	意味	関係	意味
AtLocation	A は B で見かける。	Desires	A は B が欲しい
IsA	A は B である。	Causes	A は B の原因だ。
CapableOf	A は B ができる。	UsedFor	A は B に使う。
InstanceOf	A は B の一種だ。	HasSubevent	A の時 B が起きる。
HasProperty	A は B という特徴がある。	SymbolOf	A は B の象徴だ。
PartOf	A は B の一部だ。	CausesDesire	A は B をしたくなる
MadeOf	A は B で出来ている。	DefinedAs	A は B という意味だ。

表 3.3 出力された 14 の関係とその意味

入力文	出力文(関係)
花が咲いていました。	花は山で見かけるんだよね。(AtLocation)
紅茶を飲みました。	紅茶は飲み物だよな。(IsA)
雨が降っていました。	雨だと憂鬱かもね。(Causes)

表 3.4 入力文と出力文の例

	文法	意味	語彙	知識
AtLocation	4.4	3.3	3.3	3.6
IsA	4.8	4.2	3.1	3.4
CapableOf	3.9	3.4	3.3	3.4
InstanceOf	4.5	4.05	3.7	3.7
HasProperty	4.4	3.1	3.4	3.7
PartOf	4.3	3.1	3.2	3.0
MadeOf	4.4	3.3	3.0	3.5
Desires	4.4	3.7	4.2	4.5
Causes	4.3	4.2	3.9	4.2
UsedFor	3.6	3.5	3.1	3.3
HasSubevent	3.9	2.7	2.8	3.0
SymbolOf	4.6	3.9	4.2	4.3
CausesDesire	3.2	3.1	3.5	2.9
DefinedAs	3.6	3.7	4.1	3.9

表 3.5 1 度の連想についての実験結果

「文法的な自然さ」については、著者が作成したテンプレートに基づいて出力文を作成しているため、全ての関係において 3 ポイント以上となった。「意味的な自然さ」については、ほとんどの関係で 3 ポイント以上であり、1 度の連想ではどの関係を用いてもある程度の精度を得ることが可能である。唯一 2 ポイント台だった組として、「入力『食事をしました。』出力『食事時に気持ち悪くなるよね。(HasSubevent)』」があるが、これは個人差がある文であるため、評価が低くなったと考えられる。

3.4 連想の連鎖

3.2 では入力文に対して連想を 1 度行ったが、連想を連鎖させ 2 度行い発話を生成する手法も述べる。連想の連鎖の例を表 3.6、表 3.7 に示す。例えば、「犬を飼っている。」という入力文に対して「IsA(犬, 動物)」を連想した後更に、「動物」を ConceptNet に入力し、「Desires(動物, 水)」を獲得する。つまり、「犬を飼っている。」という入力文から「IsA(犬, 動物)Desires(動物, 水)」という関係を獲得する。この関係から著者が作成したテンプレートに基づき、「犬は動物だよな。だから、水を欲しがるとだよな。」という文を生成する。

IsA(犬, 動物), Desires(動物, 水)
Desires(犬, 骨), HasProperty(骨, 硬い)
AtLocation(犬, 家), IsA(家, 建物)

表 3.6 連想の連鎖により獲得される関係【基となる文：犬を飼っている。】

犬は動物だよ。だから、水を欲しがるとね。
犬は骨を欲しがるとね。それで、骨は本当に硬いよ。
犬は家にいるんだよ。そっで、建物だよ。

表 3.7 連想の連鎖により生成される文
【基となる文：犬を飼っている。】

3.5 連想の連鎖に関する評価実験

3.4 で提案した連想の連鎖を用いた発話生成手法について、3.3 と同様の評価実験を行った。入力文は 3.3 と同様の方法で 25 文用意し、出力文は 3.4 で提案した手法に基づき生成する。入力文と出力文の例を表 3.8 に示す。被験者は 5 人で、20 代男子理系大学生 3 名、20 代男子理系社会人 1 名、20 代女子文系大学生 1 名である。被験者 5 人の平均の結果を表 3.9 に示す。

入力文	関係
	出力文
みかんを食べました。	InstanceOf(みかん, 果物) HasProperty(果物, 栄養 ある)
みかんは果物の一種だね。ってことは、栄養があるよ。	

北海道に行きました。	HasProperty(北海道, 自然 豊か) HasProperty(自然, 美しい)
北海道は自然が豊かだよ。だから、美しいよ。	

緑茶を飲みました。	IsA(緑茶, 飲み物) UsedFor(飲み物, 喉 潤す)
緑茶は飲み物だよ。だから、喉を潤すのに使うも。	

表 3.8 入力文と出力文、関係の例

被験者 5 人の平均	文法	意味	語彙	知識
0～1 ポイント	0 組	0 組	0 組	0 組
1～2 ポイント	0 組	5 組	0 組	1 組
2～3 ポイント	7 組	7 組	10 組	6 組
3～4 ポイント	8 組	8 組	13 組	15 組
4～5 ポイント	10 組	5 組	2 組	3 組

表 3.9 連想の連鎖についての実験結果

「文法的な自然さ」は 1 度の連想よりも 3 ポイント未満の組が増えている。これは、連想が連鎖するために、テンプレートの不備が増えてしまったことに起因している。「意味的な自然さ」は 25 組中 13 組が 3 ポイント以上だった。「意味的な自然さ」が 3 ポイント未満だった組の例を表 3.10 に、3 ポイント以上だった文を表 3.11 に示す。3 ポイント未満だった原因として、動詞を考慮していない点が考えられる。提案手法では、名詞からのみ連想を行っているため、入力文の動詞と出力文が無関係になってしまうことがある。これを解決するには、例えば Web 検索を利用するなどが考えられるが、今後の課題である。

入力文	出力文
鮫に襲われました。	鮫はジョーズの映画で見かけるんだよ。そっで、ビデオ店で見かけるもんだよ。
サインペンで書きました。	サインペンはペンだよ。だから書くべき内容を欲しがるとね。

表 3.10 「意味的な自然さ」が 3 ポイント未満だった組

入力文	出力文
読書をしました。	読書時に目が痛くなるよ。そういえば、目って顔の一部だよ。
コックになりました。	コックは調理することができるんだ。それって、にんじんを切る？

表 3.11 「意味的な自然さ」が 3 ポイント以上だった組

3.6 連想発話手法についてのまとめ

入力文の名詞から連想を行い、出力文を生成する手法を提案した。ConceptNet におけるどの関係を用いても発話を生成できることが確認された。また、連想の連鎖を行うことで、より知識を持っているような発話を生成できる可能性があることを示した。今後、全く無関係な連想を除去することができれば、更に自然な発話を行えると考えられる。

4. おわりに

本論文では、WordNet と Web 検索を用いた常識的知識を表現する文の生成手法を提案した。システムが自然であると評価した文の正解率は 55.0% で、ランダムで選択した場合と比べ、30.4 ポイントの精度の向上が確認された。しかし、これはまだ十分な精度ではなく、今後精度を向上させていく必要がある。また、常識的知識の利用法として、ConceptNet を用いた連想発話手法を提案した。連想発話においては、ConceptNet におけるどの関係を用いてもある程度の精度の発話を生成できることが確認された。更に、連想の連鎖を行うことでより知識を持っていると思われるような発話を行うことができる可能性を示した。今後の課題として、常識的知識を表現する文を生成する際の精度、また連想発話の精度を上げていくことが挙げられる。

参考文献

[1]大江奈緒子, 渡部広一, 河岡司, ” 地理に関する常識知識を用いたシステムの構築”, 電子情報通信学会技術研究報告, Vol. 104, No. 728, pp. 19-24, 2005

[2]秋葉友良, 藤井敦, 伊藤克且, ” 質問応答における常識的な解の選択と期待効用に基づく回答群の決定”, 情報処理学会, NL-163, pp. 131-138, 2004

[3]ConceptNet4
<http://conceptnet.media.mit.edu/>

[4] Havasi, C., Speer, R. & Alonso, J. “ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge.” Proceedings of Recent Advances in Natural Languages Processing, 2007

[5]Open Mind Common Sense
<http://openmind.media.mit.edu/>

[6] Liu, H., Lieberman, H. and Selker, T. “A Model of Textual Affect Sensing using Real-World Knowledge”, International Conference on Intelligent User Interfaces, pp. 125-132, 2003

[7] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko

Kanzaki “Enhancing the Japanese WordNet” in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP, 2009

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
<http://mecab.sourceforge.net/>

[9]北原保雄, “朝倉日本語講座 5 文法 I”, 朝倉書店, 2003

[10]河原大輔, 黒橋禎夫, “高性能計算環境を用いた Web からの大規模格フレーム構築”, 情報処理学会自然言語処理研究報告, 2006-NL-171(12), pp. 67-73(2006).