2004－DD－42　(1)
2004／1／30

重み付き意味ネットワークを用いた英文要約手法

佐藤　裕介 †　松田　聖 ††

概要 : 知識表現法の一つである意味ネットワークを拡張した、重み付き意味ネットワークを用いて、英文のテキストデータを要約するシステムを構築した。重み付き意味ネットワークにおいて、重みはアークに与えられ、アークが表す両端のノード (単語) 間の関係の強さを表す。その重みを用いて文ごとの貢献度を算出し、貢献度の高い文を抽出し要約文とする。重み付き意味ネットワークを用いることで、単語そのものに対しての前提知識が無くても、意味的な要約を行うことが出来る。また、本システムを用いたシミュレーション結果も示す。

Summarizing English Documents with Weighted Semantic Network

Yusuke SATO†　Satoshi MATSUDA††

Abstract : We propose weighted semantic networks, which extend the semantic networks by adding a weight to an arc. The weight represents the strength of the relation given by the arc. Using the weighted semantic networks, we develop a system that summarizes English documents. Based on the weight, the system gives a degree of contribution to each sentence of the document, and produces the summary of the document by extracting the sentences that have the high degree of contribution. It presents a semantic summary without any domain knowledge about the documents. Some examples are shown to illustrate the effectiveness of the system and of the weighted semantic networks.

はじめに

近年、インターネット等の技術により世界中の情報がたやすく手に入るようになり、我々が短期間に手にし得る情報の量も飛躍的に増えた。このような環境の中で、「大量の情報を処理しきれず手に余る」ということがよくある。この状況を打破するための手段の一つとして、テキスト要約 [1] がある。テキスト要約を利用し、小さくまとめられた要約文を読むことにより、その情報が我々にとって有益、または必要であるかどうかを知ることができる。また、情報検索のソフトウェア等にもテキスト要約機能を追加することで、ユーザにとって更に有益である情報検索が行えることができると考えられる。本研究ではそのようなテキスト要約をよりよくするために、知識表現法の一つである意味ネットワークを拡張した重み付き意味ネットワークを提案し、それらを用いて英文要約を試みる。重み付き意味ネットワークを用いることにより、単語そのものに対する知識が無くても、単語同士の関係のみから要約することが出来る。本論文では、第一章において本システムの基本概念であるテキスト要約と意味ネットワークについて説明する。次に第二章において重み付き意味ネットワークを提案し、第三章において本システムの概要を説明する。最後に第四章で本システムでのシミュレーションを行い、本システムを検証する。

† 日本大学大学院生産工学研究科数理工学専攻　Graduate Department of Mathematical Engineering, Postgraduate of Industrial Technology Nihon University
†† 日本大学生産工学部数理情報工学科　Department of Mathematical information Engineering, College of Industrial Technology Nihon University

1　基本概念

1.1　テキスト要約

テキスト要約を行う事は、自分に必要な情報を得るためには、とても有効な手段である。なぜなら、ユーザは要約文を読むことにより、そのドキュメントが自分にとって重要であるかどうかを知ることが出来るからである。本システムでは、貢献度によって要約をおこなう。貢献度には、文の貢献度と単語の貢献度がある。文の貢献度とはその文が、ドキュメント全体にどれほど貢献しているかの度合いであり、その文に存在する単語の貢献度の平均で与えられる。単語の貢献度は、その単語が文に対してどれほど貢献しているかの度合いである。本システムにおいて、単語の貢献度の決定には重み付き意味ネットワークを用いる。

1.2　意味ネットワーク

重み付き意味ネットワークを説明する前に、その前身である意味ネットワークについて説明する。意味ネットワーク (Semantic Network) とは、1968 年に Quillin によって提案された知識表現手法であり、連想能力をモデル化したものである [2][3]。意味ネットワークは、概念やオブジェクトを表すノード (node) と、ノード間の関係を表すアーク (arc) の二つの要素でネットワーク状に構成されたものである。意味ネットワークの node から node に arc を伝っていくことで、推論を行う事が出来る。図 1 にその基本構造図を示す。

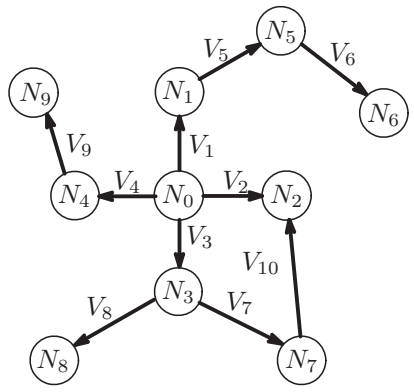

図 1: 意味ネットワークの基本構造図

図 1 において $N_i(i = 0, 1, 2, \ldots, 9)$ は node を表し、要約を行う時は名詞が当てはめられる。また、各 node には重要度が与えられる。重要度とは、最も重要な単語 (最重要単語) から、どれぐらい意味的に離れているかという度合いである。本システムではドキュメント中の最多出現名詞を最重要単語とする。$V_i(i = 0, 1, 2, \ldots, 9)$ は arc を表し、動詞が当てはめられる。この二つの要素により意味ネットワークは構成される。このときの arc の向きは重要度の高い単語から重要度の低い単語へと向く。最も重要な単語 (最重要単語) が決まることにより、全ての arc の向きが決定する。具体例として、N_0 =Taro、N_2 =pen、V_2 =have が与えられたとすると、Taro と pen の間には have という関係が成り立つ事がわかる。

1.2.1　意味ネットワークの抽出

本システムでは、意味ネットワークを抽出することが一番重要である。なぜなら、ここで抽出した意味ネットワークは、重み付き意味ネットワークの作成、重要度の決定等、本システムに深く関与しているからである。具体例として、以下の例文を用いて意味ネットワークの抽出を説明する。

Taro has a pen.

この文を品詞分解すると、Taro と pen が名詞、has が動詞、a が冠詞となる。このとき、名詞である Taro と pen は node に、動詞である has は arc に割り当てられる。その他の品詞は意味ネットワークでは使用しない。つまりこの場合、冠詞である a は意味ネットワークでは使用されない。このようにして、この文より抽出される意味ネットワークは

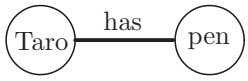

図 2: 摘出例

となる。ここで、着目すべき点は arc に向きが無いことである。現時点において、Taro と pen のどちらの重要度が高いのかが分からないため、向きをつけることはできない。このことは、arc が決してドキュメント中の主述関係を表すわけではないということを表している。図 2 のような枝をすべて集め、連結することで意味ネットワークを構成する。

2　重み付き意味ネットワーク

本論で提案する重み付き意味ネットワークとは、意味ネットワークの arc に対して重みを付加したものである。本システムにおいて重み付き意味ネットワークは、背景知識として用いる。図 3 に重み付き意味ネットワークの基本構造を示す。

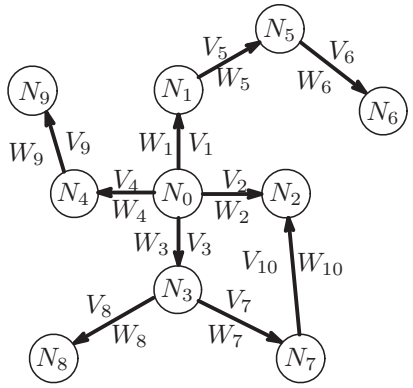

図 3: 重み付き意味ネットワーク基本構造図

図 3 において、W_i は重みを表し、各々の arc に与えられる。テキスト要約で、arc は動詞 V_i を表しているため、重み付き意味ネットワーク上に同じ動詞が割り当てられた arc が多数存在する可能性がある。このとき、同一の動詞が割り当てられた arc の重みはどれも等しくなる。重みは新しくドキュメントを追加するたびに更新される。このとき、新しいドキュメントにおいて使われない動詞を持つ arc の重みは減衰する。このことで、より多くのドキュメントに出現する関係が強化され、少数のドキュメントにしか出現しないような関係は淘汰されていく。

2.1　重み

重みとは、arc の関係の強さを表し、重みが大きい arc で表される関係は強い関係である。重みを与えることで、同じ node につながる node 同士であっても、その関係の強さに差が出てくる。例えば図 3 において、N_0 とつながっている node は N_1、N_2、N_3、N_4 と四つある。それぞれの arc に対する重み W_1, W_2, W_3, W_4 に、

$$W_1 = 0.1, W_2 = 0.7, W_3 = 0.3, W_4 = 0.5$$

という値が与えられたとする。このとき N_0 との関係の強い順に $N_2 > N_4 > N_3 > N_1$ となる。重みは様々なドキュメントを追加した結果、算出されるものであるので、その関係がどれくらい重要な関係であったかという指標にもなる。つまり、重みの大きい関係はどのドキュメントにおいても強い関係を示していたということである。

node の重要度の算出

追加されるデータ内の arc に対して重み付けを行うために、まず node の重要度の算出が必要となる。そこで、テキストデータから抽出された意味ネットワークの距離に着目する。意味ネットワークにおける距離とは、任意の node 間にどれほど arc が存在するかという値である。この距離から、node の値である重要度を算出し、重要度を使って重みを計算する。ここで、nodeN_i に対する重要度を I_i とする。また距離における arc の本数は 2 つの node 間の最短経路の本数である。例えば、図 3 において、N_0 と N_2 との距離は、その間に V_2 の 1 本Ａｒｃが存在するので、1 ということになる。前述の N_0 と N_2 との距離で見てみると、$V_3 \rightarrow V_8 \rightarrow V_7$ を通って N_2 まで行く経路が考えられるが、V_2 を通る経路が最短であるため、2 つの node の距離は 1 ということになる。意味ネットワークの任意の単語 N_i の重要度 I_i の算出式は以下のようになる。

$$I_i = I_{max} - d_i$$

ここで、I_{max} は最重要単語の重要度である。最重要単語の重要度は、各名詞の重要度が正の値を取るように、最重要単語から各 node への最短距離の最大値+1 をその値とする。d_i は最重要単語から N_i までの最短距離を表す。また、最重要単語から、どんな arc を伝っても到達しない (関係が無い) 単語は、重要度を 0 とする。このようにして、ドキュメント中の全ての名詞に対し重要度を決定する。

arc の重みの算出

上記のようにして得られた重要度を入力として重みの算出を行う。図 4 のように node N_A、N_B を arc$V$ でつながれている枝が与えられたとする。このとき、N_A と N_B の重要度を I_A、I_B とする。また、図 4 における arc の向きは N_A から N_B であるため、両 node の重要度の関係は $I_A > I_B$ となる。このときの V_α の重み W_α は

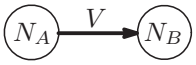

図 4: 意味ネットワーク例

$$W_\alpha = \frac{I_B}{I_A}$$

となる。このとき、ネットワークの他の arc に対しても同じ V_α が割り当てられた場合、それらの重みの平均を $\hat{W_\alpha}$ とする。例えば、動詞 take を arc に持つ枝が、一つのテキスト文で 3 本抽出されたとする。このと

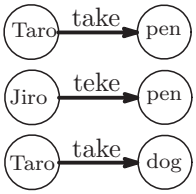

図 5: take が使用されている枝

き、それぞれの名詞に対する重要度が Taro=8、Jiro=5、pen=4、dog=2 と与えられたとする。それぞれの枝に対して、take の重みは

$$\begin{array}{lllll} \text{(1) の枝} & \ldots & \frac{W_{pen}}{W_{Taro}} & = \frac{4}{8} & = 0.5 \\ \text{(2) の枝} & \ldots & \frac{W_{pen}}{W_{Jiro}} & = \frac{4}{5} & = 0.8 \\ \text{(3) の枝} & \ldots & \frac{W_{dog}}{W_{Taro}} & = \frac{2}{8} & = 0.25 \end{array}$$

となる。このとき、動詞 teke に対する重みは、3 つの枝の平均値の 0.517 となる。このようにして得られた重みを新しい重みとして、重み付き意味ネットワークに順次追加する。

重みの学習

新しいドキュメントが追加され、新しい arc が重み付き意味ネットワークに追加されることに伴う重みの調整 (更新) を行う必要がある。このことを重みの学習という。新しい重みの学習は以下の手順で行う。

1 古い arc に対し一定の減衰率により重みを減衰させる。

2 新しい arc と同じ arc を古いものから探す。

3 同じ arc が見つかったら、それら二つの重みの平均を新しい重みとする。

4 新しい arc が無くなるまで、2,3 の手順を繰り返す。

動詞 V_i に対しいて、追加する重みを W_i^{new}、現在の重みを W_i^{old}、減衰率を ψ とすると更新後の重み $\hat{W_i}$ は

$$\hat{W_i} = \frac{(W_i^{\text{old}} * \psi) + W_i^{\text{new}}}{2}$$

と定義される。具体例として上記の動詞 take において、追加する重み $W_{take}^{\text{new}} = 0.6$、今までの重み $W_{take}^{\text{old}} = 0.4$、減衰率 $\psi = 0.9$ とすると、新しい重み $\hat{W}_{take}$ は

$$\hat{W}_{take} = \frac{(0.4 * 0.9) + 0.6}{2} = 0.48$$

となる。このように、随時、重みを学習していく。次章において本システムでの要約の流れを示す。

3 要約手法

本システムにおける要約は以下の流れにおいて行う。以下に要約手法の大まかな流れを示す。

1 形態素解析

2 最重要名詞の決定

3 意味ネットワークの抽出

4 各名詞の貢献度を決定

5 各文の貢献度を算出

6 文を削除して要約

まず、入力されたデータに対し形態素解析 (品詞分解) を行う。形態素解析と並行して最多出現名詞を求め、最重要単語とする。次に、テキストデータより意味ネットワークを抽出する。抽出された意味ネットワークと、重み付き意味ネットワークを用いて、全名詞の貢献度を決定する。単語の貢献度は、重み付き意味ネットワークの重みを利用して決定される。次節で、単語の貢献度の決定法を詳しく説明する。そのようにして得られた名詞の貢献度を一文ごとのに代入し、その平均をその文のドキュメントに対する貢献度とする。文の貢献度の低い文を取り除き、文の貢献度の高い文を残すことで、要約文とする。

3.1 単語の貢献度の決定

単語の貢献度は、重み付き意味ネットワークの重みを用いて各単語ごとに決定される。抽出された意味ネットワークを最重要単語から arc を通過するごとに、重みによって単語の貢献度を減少させ、node にたどり着いた時の値を、その単語の貢献度とする。このとき、テキストデータより抽出される arc には重みが存在するものと存在しないものがある。重みの存在する arc を通過する場合、最重要単語に近い node に割り当てられた名詞の貢献度を C_α、他端の node に割り当てられた名詞の貢献度を C_β とし、それらをつなぐ arc の重みを W とすると

$$C_\beta = W * C_\alpha$$

となる。つまり、最重要単語から任意の単語 N_i までの間の枝に重みがすべて存在する場合、N_i に割り当てられた名詞の貢献度 C_i は

$$C_i = C_{max} * \omega_i$$

となる。ここで、ω_i は、最重要単語から N_i までに通過した arc の重みの総積である。しかし、重みのない arc が存在する時は、その arc を通過する時は、重要度の算出法と同じ方法を用いる。また、最重要単語とまったく関係の無い単語の貢献度は 0 とする。

3.2 文の貢献度の算出

このようにして決定された単語ごとの貢献度を使い文の貢献度を算出する。文中に存在する名詞の貢献度度の平均をその文の貢献度とする。平均を取ることで、文の長さ、単語数の多さ等に影響されずに、その文の貢献度を算出することが出来る。

3.3 要約

文の貢献度が一定値以上の文を抽出し、要約文とする。この閾値を様々に設定することで、目的に合わせた要約文を作ることが出来る。

4 検証

最初に重みの学習を行う。次に複数のテキストデータを用いて要約をすることにより、本システムの有効性を測る。

4.1 学習の入力データ

学習の入力データとして、ジャンルの同じテキストデータを用いる。用いるジャンルは「物語」とする。「物語」のテキストデータ 10 個を入力データとして使用した。以下に、学習された重みの一部を示す。

was	=	0.905
give	=	0.765
bring	=	0.531
touch	=	0.354
opened	=	0.172

上のような学習した重みを用いて以下の三つのテキストデータで本システムを検証する

- 学習で使用したテキストデータ
- 入力データと同ジャンルのテキストデータ
- 他ジャンルのテキストデータ

学習で使用したテキストデータは、単語情報が、すべて重み付き意味ネットワークに組み込まれているため、重み付き意味ネットワークが最大限に活かすことが出来ると考えられる。同ジャンルのテキストデータでは、単語情報が全て重み付き意味ネットワークに組み込まれてはいないが、組み込まれている単語情報は同ジャンルからの単語情報であるので、有効に使えると考えられる。他ジャンルのテキストデータにおいては、そのテキストデータに関係のない単語情報が与えられている状態での要約で有効性を測る。

4.2 学習で使用したデータ

重み付き意味ネットワークを作成時に使用したテキストデータとして、「Cinderella」を用いて要約をおこなう。要約文として以下の文が得られた。

要約文

- poor Cinderella had to rush about upstairs and downstairs.
- Cinderella fixed their hair in fancy waves and curls.
- said Cinderella and Cinderella climbed into the coach.
- now Cinderella was enjoying the ball so much that Cinderella forgot her fairy godmothers warning until it was almost midnight and the clock began to strike.
- prince said.
- at last the prince came to Cinderella house.

考察

上記を見ても分かるように、この要約文は妥当であるといえる。また、このテキストデータは入力時で使用しているため、このテキストデータに出現する全ての動詞に対して重みが存在している事も、このような良好な結果が得られた理由である。

4.3 同ジャンルのテキストデータ

同ジャンルで、重み付き意味ネットワークの作成時に使用しなかったテキストデータの例として「The Little Match Girl」を要約する。要約文として以下の文が得られた。

要約文

- a poor little girl drew one out.
- the little girl had already stretched out her feet to warm them too ; but the small flame went out , the stove vanished : a poor little girl had only the remains of the burnt out match in her hand.
- a poor little girl lighted another match.
- now there a poor little girl was sitting under the most magnificent Christmas tree : it was still larger , and more decorated than the one which a poor little girl had seen through the glass door in the rich merchant house.
- the little maiden stretched out her hands towards them when the match went out.

考察

この場合においても良好な結果が得られたといえる。このことで、そのドキュメントに対する単語情報が不完全であっても要約が出来るといえる。また、そのドキュメント自体に対する知識がまったく無くても、重み付き意味ネットワークが完全に全ての動詞に対して重みが存在するならば、良好な要約が行えると考えられる。

4.4 他ジャンルのテキストデータ

重み付き意味ネットワークの作成に使用したジャンル (物語) とは異なったジャンルのテキストデータとして「JFK's Inaugural Address」を要約する。要約文として以下の文が得られた。

要約文

- Americans dare not forget today that Americans are the heirs of that first revolution.
- Americans offer a special pledge.
- Americans dare not tempt them with weakness.
- but a call to bear the burden of a long twilight struggle.

考察

検証テキストデータの背景知識も無く、ジャンルに関する背景知識も無いデータを要約したが、要約文を見てみると妥当であることが言える。このことにより、ジャンルに関係なく重みを有効的に使用することが出来ると考えられる。

5 おわりに

本システムの構築により、前提知識が無い状態でもテキスト要約が行える事が立証された。また、ジャンルに関係なく要約できたことによって、重み付き意味ネットワークに十分な情報があるとき、任意のジャンルの任意のテキストデータに対しても要約が行えると考えられる。今後の研究課題として、様々なジャンルのテキストデータを重み付き意味ネットワークの入力とした場合の検証が挙げられる。これは、本論においては 1 ジャンルの知識しかない状態での検証であったが、これを複数のジャンルの知識を統合した状態での検証するということである。そのような状況であっても本システムの有効性を示すことが出来ると、総合的な重み付き意味ネットワークを作成すれば、任意のジャンルの任意のテキストデータに対しても有効的な要約が行えると考えられるからである。

参考文献

[1] R. グリシュマン 著、山梨正明、田野村忠温 共訳 (1989),『計算言語学－コンピュータの自然言語理解－』, サイエンス社

[2] A.Barr/E.A.Feigenbaum 編、田中幸吉、淵一博 監訳 (1987),『人工知能ハンドブック』, サイエンス社

[3] Daniel G BOBROW/ALLAN.COLLINS 編、田中幸吉、淵一博 監訳 (1978),『人工知能の基礎』, 近代科学社