忘却をシミュレートする知識辞書システム

奥野正也, 佐川雄二, 田中敏光（名城大理工学研究科）

1. はじめに

我々は，頭の中でイメージを作り，話したり物事を考えて生活している．特に言語の使用には単語を使用しているが，普段の生活の中で，頭でイメージしたものと一致する単語がなかなか出てこないことがある．

そうした，忘れてしまい，思い出せずにいる単語が検索できるシステムが構築できれば，人の知的処理を助けてくれると同時に，人の思考に近いシステムとなり，人とコンピュータとのコミュニケーションにも役立てるのではないかと考えた．

そこで忘れた単語を思い出す手助けをするシステムを構築を目的として，そのために必要な知識辞書の検討と開発を行った．

2. 従来の単語による知識表現の手法

単語による知識表現の方法にはシソーラスによる方法と属性空間を用いる方法がある．以上の二つの方法はどちらも手動でシステムの構築を行っている．このため，システムが十分な知識を構築するまで時間が掛かる．また，システムの構築には，システム設計者の主観が大きく反映されてしまう為，すべてのユーザに対してシステムが使えるか，設定が正しいのかといった問題がある．

3. 知識辞書作成システム

3.1 単語間距離と時間

単語を思い出すきっかけとして，その単語と関連する他の単語がある．ただ，なかなか思い出せない単語の場合は，関連する単語は思い出せてもその単語自体が思い出せないと思われる．単語間の関係の強さを単語間距離で表すことがあるが，その観点から見れば，思い出せない単語は，他の単語との単語間距離が大きくなってしまっているのである．

一般に単語間距離は学習された後は固定的であるが，上記のようなことを実現するためには，時間とともにあまり使われない関係を忘れていく機構が必要である．すなわち時間とともに単語間距離が大きくなっていくような知識辞書である．ここでの忘却はあくまでシミュレーションであり，本当に忘れてしまうわけではない．

3.2 知識辞書の構成

本研究では，辞書の要素は名詞に限定する，したがって辞書は，各名詞をノードとし，それぞれの名詞と名詞の関係をリンクとする図１のようなネットワークである．リンクには，単語間の距離値，頻度，型の三つの情報が付与される．

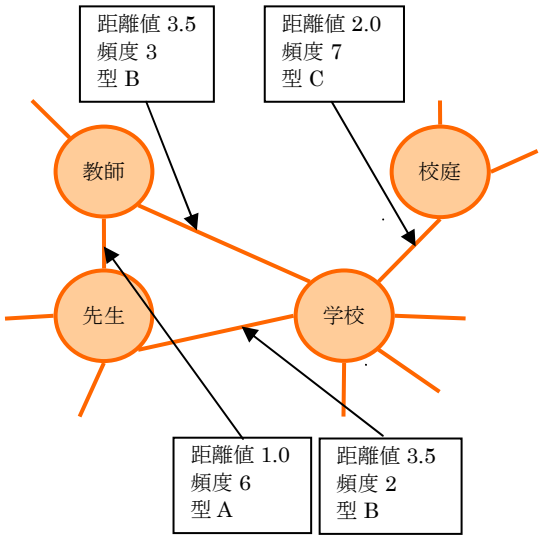

図１　名詞のネットワーク

3.3 単語間の距離

単語同士の結びつきである単語間の距離の度合いを単語間の距離値とおく．単語間の距離値は，単語同士の結びつきが強ければ，単語間の距離値が０に近い値になり，結びつきが弱ければ大きな整数値となる．

3.4 単語間の距離値の変化

単語間の距離値は以下の図のように変化する．

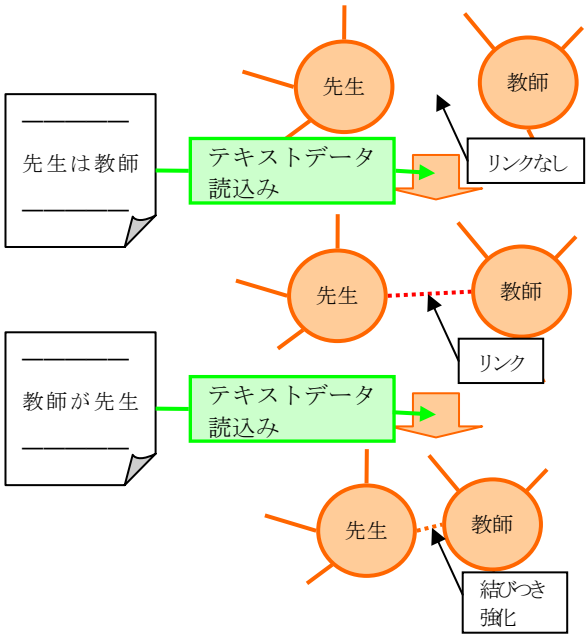

図２　テキスト読込みによる単語間の距離値変化

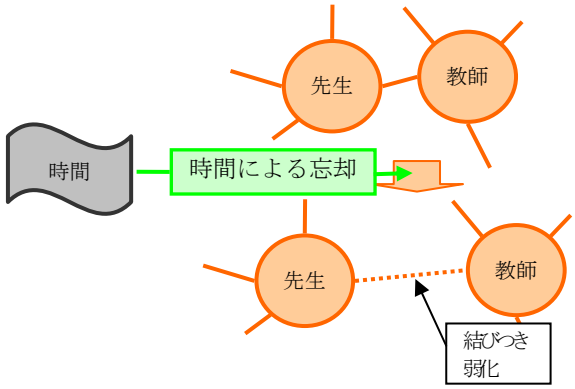

図３　時間による単語間の距離値変化

3.5 テキスト読み込みでの単語間の距離値計算

テキストデータを形態素解析[3]し，同じ文中に共起する名詞の組み合わせをそれぞれ抜き出す．

新たな名詞の組み合わせが出現した場合，単語間の距離値は１，頻度は１に設定する．

既出の名詞の組み合わせが出現した場合，既存の単語間の距離値から４分の１の値を入れる．さらに頻度を＋１に更新する．

3.6 忘却による単語間距離計算

忘却は単語間の距離値に以下の式の計算を加算することによってシミュレートする．

200＊一日／頻度

以上のように単語間の距離値は頻度によって一日での増加率が変わる．

3.7 知識の事前の一般知識取得

テキストデータからの知識獲得には，ドメインが偏ってしまうことや十分な情報が短期間の時間に知識辞書に与えられないことがある．そのためユーザが知識辞書へのテキストデータを与えていく前に，広くドメインを網羅している情報を与える必要がある．

この情報はシソーラスの単語同士のつながりを単語間のリンクにそのまま利用する事や，広辞苑などの辞書をテキストデータとして知識所書に与える事で構築する．ただし，それによって得た情報は，いつでも知識の中でも重要な情報として常に利用できるようにしなければならない．また，ユーザによって構築される知識データ領域とうまく調和させなければならない．これについては次項の型の利用にて言及する．

3.8 型情報の利用

単語間の距離値の計算による知識辞書の構築を行った場合，常に出現する単語同士の組とすでに出現されなくなった単語同士の組や，ある時期にのみ出現する組などに分けられるようになる．さらに事前に取得する一般知識も存在するため，型情報を用いて単語間のリンク情報の分類分けを行う．型によって単語間の距離値の変域の制限や単語間の距離値の変化率を変化させる．

4. 実験

Visual C++で本システムを実装し，毎日新聞 2003 年の１月１日から１月 31 日までの一ヶ月分をテキストデータとして読み込んだ．実験で得られた知識辞書から「小泉」とつながりのある，同じ頻度の単語「総務」と「自民」の単語間の距離推移を以下に示す．尚，これらは事前知識に獲得したものでもなく，型情報は利用されていない．

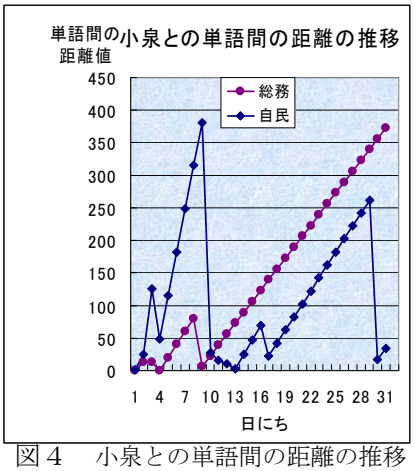

図４　小泉との単語間の距離の推移

5. おわりに

本システムによりシステムを使用しているユーザが読み込ませたテキストによってユーザにあった知識辞書の構成が可能になった．また時間によっても知識構成が変化していくのでユーザ自身の知識変化にも対応できるようになり，目的とする忘れた単語を思い出す手助けをするシステムの基礎が出来上がった．

参考文献

[1] 冨浦 洋一，田中 省作，日高 達，“共起データに基づく名詞の多次元空間への配置，”人工知能学会論文誌，Vol. 19(2004)，No.1 pp.1-9.

[2] 笠原 要，稲子 希望，加藤 恒昭，“単語の属性空間の表現方法，”人工知能学会論文誌 Vol.17(2002)，No. 5 pp.539－547.

[3] 松本，北内，山下：形態素解析システム「茶筅」，奈良先端科学技術大学院大学（1997）