「奇麗だったら美しい！」~ 概念ベースによる意味照合

佐藤浩史　　藤本和則
日本電信電話(株) コミュニケーション科学基礎研究所
{hiroshi, fujimoto}@cslab.kecl.ntt.co.jp

１．はじめに

近年のインターネットの普及により，我々は世界中の最新の情報を瞬時に取得できるようになった．それに伴い，ユーザの意思決定支援をネット上の情報を用いて行うシステムの研究が各方面で盛んに行われている．そこで我々は，統計的決定の見地から，ユーザの要求に対してその判断へのアドバイスを与えるシステムを，DSIU（Decision Support for Internet Users）システムと名付け，実現を目指して研究を進めている[1]．

ネット上の情報を取り扱うためには，テキストからの知識獲得技術が重要となる．その際に手動獲得をベースとしているもの，例えば Ontobroker [2]などがあるが，インターネット上の情報はその量が桁外れに膨大で，かつ更新周期が短いといった特徴がある．そのため，メンテナンスに莫大なコストがかかり，あまり実用的とは言えない．一方，Web-KB project [3]の様に，知識の自動獲得を行うものもある．DSIUシステムにおいても，テキストのレイアウトに着目した自動知識獲得の研究を行っている[4]．

このような自動知識獲得技術を実際にネット上のテキストに適用するためには，テキスト上の単語をその表記ではなく意味で認識することが必要となる．例えば，DSIUシステムが製品（ディジタルカメラ）の選定を支援する為に，製品記事から製品の属性の依存関係知識を獲得することを考えてみよう．製品記事に，「この CCD により，美しい画像が得られます」という表記があったとする．ここから，画像の「美しさ」は「CCD」に依存する，といった知識が獲得できる．この知識を用いて，「美しい」画像が撮りたいユーザに対しては「CCD」が重要な仕様であることが判るわけだが，もしもユーザが「奇麗な」画像を望んでいる場合には，その「美しい」と「奇麗」の表記上の違いにより，知識を適用することが出来ない．この様に，知識を有効に獲得し活用するためには日常語の意味照合技術が必要となる[5]．

本稿では，辞書から構築した日常語データベース（概念ベース[6]）を用いた意味照合について述べる．

２．意味照合

2.1　言葉の意味の類似度

言葉の類似性の定量的評価については，今まで数多くの研究がなされてきた．認知心理学の分野では，古くから様々な実験が行われている[7,8]．これらは興味深い結果を多く出してはいるが，計算機でのシミュレーションを狙ったものではない．そこで，計算機上での類似判別を実現すべく，自然言語処理技術を用いた手法が近年多く提案されている．例えば，シソーラスを用いたもの[9,10]，コーパス上での共起関係を用いたもの[11]，国語辞典の語義文を用いたもの[12,13]等である．我々が構築した「概念ベース」[6]は最後のケースに分類される．

2.2　概念ベース

意味照合には単語の類似性判別技術が必要である．そこで我々は，国語辞典より自動構築した単語の意味のデータベースである「概念ベース」を適用する．これはベクトルモデルによる約 40,000 語の単語概念のデータベースで，ベクトルとして表された単語間の類似度を，その余弦をもって定義することで，任意の 2 単語の類似度を算出することが出来る．

しかし，この概念ベースには次の制約がある．概念ベースにおける類似度は，ある単語を基準として，そこから他の単語への類似性の強弱を見る相対的な指標（すなわち順序尺度）にすぎない．言葉を換えれば，概念ベースでは「単語Ａ，Ｂ，Ｃに対し，ＢとＣのどちらがよりＡに似ているか」を判断するが，「単語Ａ，Ｂは互いに似ているか否か」を判断する必要はないと考えられている．これは，「人間は絶対的な類似性判別は出来ない」という経験の下の思想である．従って，類似度の尺度が統一されている保証はない．例えれば，「長さ」の尺度として，「cm」や「inch」等が混在している状態である．その為，全ての単語にわたって共有できる類似度閾値の設定が出来ないことになる．意味照合は，与えられた閾値の下に「似ているか否か」を判断する必要があるため，このままでは概念ベースを適用できない．しかし概念ベースには，その構造上，絶対的な判別をするための潜在能力があると我々は考えている．

以下，与えられた閾値の下に，単語組に対して「似ているか否か」を判断することを，絶対類似性判別と呼ぶ．

2.3　絶対類似性判別

概念ベースによる絶対類似性判別を実現するためには，類似度の尺度の統一が必要となる．類似度の尺度は，単語ごとの類似度分布にその特徴が現れていると考えられる．そこで，分布の標準化を行い，共通の尺度をもつ類似度を定義する．

まず初めに，概念ベースの類似度は比例尺度であると仮

定する．これは，概念ベースがベクトルモデルを採用していることを考えれば，自然な仮定である．この仮定により，各種の統計量を用いることが出来るようになる．次に，単語 w, w' 間の通常の類似度 sim(w,w') に対し，w と全単語の類似度の相加平均を $\overline{sim}(w)$，その分布の標準偏差を v(w) としたとき，w から見た w' との標準化された類似度 s_W を以下で定義する．

$$s_W(w') := (sim(w,w') - \overline{sim}(w)) / v(w)$$

s_W の相加平均は 0，標準偏差は 1 となり，これらの値によらない類似性判別が可能となるが，このままでは対称律が成り立たない．そこで，実数 t に対し次が成立するときに w と w' は (level τ で) 似ていると定義する*．

$$s_W(w') \geq \tau \text{ かつ } s_{W'}(w) \geq \tau$$

この τ が類似度閾値であり，1 つ定めれば全単語共通の類似レベルとなる．なお sim の値域が [0,1] なのに対し，s_W の値域は実数全体となる．

3．実験

類似度の標準化の絶対判別での有効性を示すため，2 種類の比較実験を行う．まず，既存の類似度に敢えて閾値を与えた場合と比較し，標準化の有効性を示す．次に，シソーラスによる類似度での場合と比較し，辞書による概念ベースの優位性を示す．実験に際し，2 つの単語セットを準備する．1 つは，インターネット上からランダムに取得した，72 のディジタルカメラの記事中に出現した自立語 982 語．もう 1 つは，ユーザのディジタルカメラに対する要求文に含まれる頻度の高い自立語 10 語である．この 2 つのセットにより，意味照合の対象が 9,820 組できる．ここから，類似していると思われる組を人手でマークした．その結果，(安い，安価)，(簡単，容易)，(格好，外装) など計 54 組がマークされた．これらを意味照合で類似と判定されるべき正解組とし，各手法の F値を算出した．

$$F := (\beta^2 + 1.0) \cdot P \cdot R / (\beta^2 \cdot P + R)$$

F値は，再現率 R と適合率 P からなる評価値であり，値の大きさがパフォーマンスの高さを表す．また，β は再現率と適合率の重みの比であり，β が大きければその分再現率が重視された評価となる．

3.1　評価 1

標準化された類似度と通常の類似度の比較を行う．類似度を標準化する際には，使用する単語組 9,820 組に対して算出した平均及び標準偏差を用いた．

DSIUシステムでは，知識の信頼度を確率超空間の下に計算を行う[14]．それ故，多少の間違った知識は許容されるが，あまりに多い間違った知識は当然間違った推論結果を引き起こす．従って，意味照合には高い適合率が要求される．その一方，再現率に関しては，知識ソースがネット上のテキストであるため，そのソースの豊富さからあまり高い値は要求されない．これらのことから，F値の β は 1 よりも小さいと結論される．

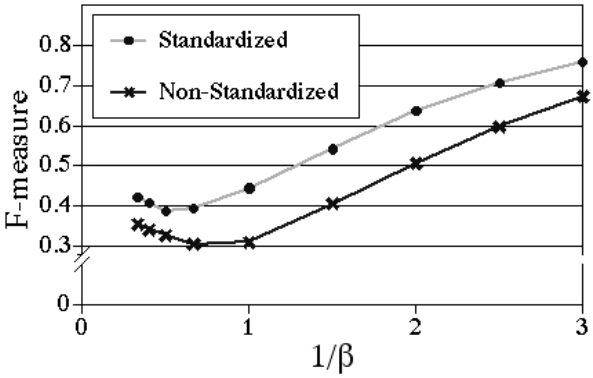

図 1 ：標準化類似度と非標準化類似度の比較評価結果

図 1 により，標準化された類似度による照合は，既存の類似度のものに比べて，β によらず常に高いパフォーマンスを出すことが判る．すなわち，どのようなアプリケーションに対しても良い結果を出すと言える．

	Precision	Recall
Standardized	*0.84*	**0.30**
Non-Standardized	*0.83*	**0.19**

表 1 ：適合率と再現率の例

適合率が 0.8 以上，という制限の下での各手法の再現率を表 1 に示す．適合率 0.8 は，DSIUシステムとして十分に高い値である．標準化された類似度では，従来の類似度のものに対し，約 1.5 倍の再現率が得られている．この事からも，標準化の有効性が言える．

3.2　評価 2

次に，類似性判別の手法自体を評価するために，概念ベースによる類似度と，シソーラスによる類似度の比較を行う．概念ベースの類似度は提案手法により標準化されたものを使用する．

シソーラスは，各単語を複数の意味カテゴリに分類したものである．今回の実験では角川シソーラス[15]を用いた．このシソーラスは，3 階層の木構造をとっており，10 の大分類と，それぞれの子となる 10^1 の中分類，そして孫となる 10^3 の小分類から成る．この小分類に約 50,000 単語が割り振られている．但し，同時に複数の小分類に属する単語も存在する．この階層性および語の多義性を考慮して，類似度を次のように定義する．

* or でも定義はできるが，精度が低下してしまう

$$c(w,w') / (c(w) \cdot c(w')) + (\Sigma_S b(s,w) \cdot b(s,w')) / (10^2 \cdot c(w) \cdot c(w'))$$

w, w' を単語，s を中分類，c(w) を w が属する小分類数，c(w,w') を w と w' が共に属する小分類数，b(s,w) を w が属し，かつ s の子である小分類数とする．第1項が両単語が同じ小分類に属した場合であり，第2項が同じ中分類下の小分類に属した場合である．

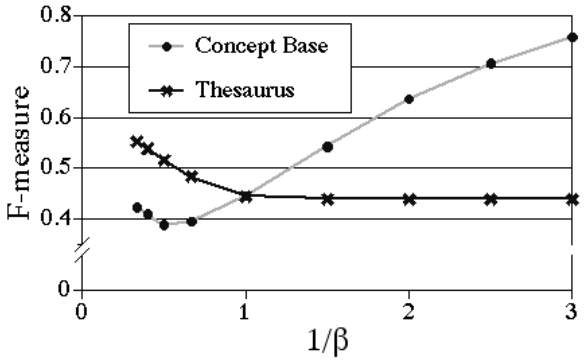

図 2 ：概念ベースとシソーラスの比較評価結果

図 2 から，1/β が 1 より大きい場合は，概念ベースがシソーラスより優れていることが判る．さらに，シソーラスは再現率は高い値を出すが，適合率が低いこともわかる．このことから，適合率を重視する DSIUシステムにはシソーラスよりも概念ベースが適しているといえる．また我々はさらに意味照合の性能を向上させるため，概念ベースをシソーラスと組み合わせて使う方法を実験中である．

４．まとめ

テキストから知識を効率良く獲得するためには，意味照合の技術が必要となる．我々が辞書から自動構築した概念ベースは単語間の類似度を算出できるが，その値は相対的であるためそのままでは意味照合を行うことが出来なかった．本稿では類似度の分布に着目し，類似度の標準化の提案を行い，実験でその有効性を確認した．

この照合技術を組み込んだ DSIUシステムに関しては，以下のページで情報を得ることが出来る．

http://www.kecl.ntt.co.jp/DSIU/

今後は，より柔軟な照合法の検討と，精度向上の研究を行っていく予定である．

参考文献

1. 藤本, 賀沢, 佐藤, 阿部, 松澤, DSIUシステム:Decision Support for Internet Users, *人工知能学会論文誌*, Vol. 15, No. 1, pp. 61-64 (2000)
2. Fensel, D., Decker, S., Erdmann, M., and Studer, R., Ontobroker: The very high idea. In *Proc of the Eleventh International Flairs Conference, FLAIRS-98*
3. Craven, M., DiPasquo, P., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S., Learning to extract symbolic knowledge from World Wide Web. In *AAAI-98*. The AAAI Press. (1998)
4. 賀沢, 藤本, 松澤, Webテキストを知識ベースとして用いる推論システムの提案: テキストからの知識獲得方式を中心に，*人工知能学会研究会資料 SIG-KBS-9803-9*, pp.49-54 (1999)
5. 佐藤, 藤本, 松澤, DSIUシステムにおける発想的意味照合 –概念ベースに基づく照合法を中心に–, 第59回情処全大予稿集, Vol.2, pp. 65-66 (1999)
6. 笠原, 松澤, 石川, 国語辞書を利用した日常語の類似性判別, *情処論文誌*, Vol.38, No.7 (1997)
7. Tversky, A. Features of similarity. In *Psychological Review*, 84, pp.327-352 (1977)
8. Suzuki, H., Ohnishi, H., and Shigemasu, K., Goal-directed processes in similarity judgements. In *Proc. 14th Annual Conference of the Cognitive Science Society*. pp. 343-348 (1992.)
9. Morris, J. and Hirst, G., Lexical cohesion computed by thesaural relations as an indicator of the structure of text. In *Computational Linguistics*, 17, pp. 21-48 (1991)
10. 大井, 隅田, 飯田, 単語間の意味的類似度に基づく文書検索手法, *言語処理学会第 2 回年次大会予稿集*, pp. 109-112 (1996)
11. Hindle, D., Noun classification from predicate-argument structures, In *Proc. of 28th Annual Meeting of ACL*, pp.268-275 (1990)
12. 北川, 清木, 人見, 意味の数学モデルとその実現形式について, *信学技報 DE93-4*, pp.25-32 (1993)
13. 小嶋, 古郡, 単語の意味的な類似度の計算, *信学技報 AI92-100*, pp.81-88 (1993)
14. 藤本, 松澤, Webテキストから獲得した制約型確率知識を扱う超空間推論法，*情処研報 ICS-116*, pp.1-6 (1999)
15. 大野, 浜西, *角川類語辞典*, 角川書店 (1985)