DSIUシステムにおける発想的意味照合

2J－8

－概念ベースに基づく照合法を中心に－

佐藤浩史　藤本和則　松澤和光

日本電信電話(株) コミュニケーション科学基礎研究所

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．しかし，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．その為，予備知識のあまりないユーザでは，たとえ検索エンジン等を使用したとしても，それらの情報を有効には活用できないのが現状である．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]．本稿ではDSIUシステム実現の為の核技術の一つ，発想的意味照合法についての実験および考察を行う．

2．DSIUシステム

2.1　テキストからの知識獲得

DSIUシステムがユーザを支援する例として，ユーザがある製品（例えばディジタルカメラ）の購入を検討する場面を考えてみる．ユーザは自分の好みを自分の言葉で（「とても奇麗に写せて，持ち運びに便利な」等）を入力する．DSIUシステムはこれに対し，インターネット上の各機種の情報を元に「どの機種がよりユーザの好みに合致しているか」を推論し，適切な機種名を提示する．これは，主にその製品の予備知識がないユーザを対象とする．

機種の評価に必要な知識の一つは，機種のスペック（画素数，重量等）とそれがもたらす特性（奇麗さ，携帯性等）の依存関係および影響力である．DSIUシステムでは，これらの知識をインターネット上のテキストから自動獲得する[2]．しかし，テキストから得られる知識は概念の表記がさまざまなので，そのままでは適切な推論が行えない．例えば，「画素数が多い→『美しい』画像」というルールを獲得しても，「『奇麗な』画像を撮りたい」というユーザの要求には適用できない．また，「画素数が多い→最高の『画質』」というルールも実質は同じものであるが，これの適用もできない．これらを適用可能とするためには，表記ではなく意味による柔軟な照合の技術が必要となる．

Abductive Matching with Similarity and Causality

Hiroshi SATO, Kazunori FUJIMOTO and Kazumitsu MATSUZAWA
NTT Communication Science Laboratories
email:hiroshi@cslab.kecl.ntt.co.jp

2.2　発想的意味照合

言葉の意味の類似性／関連性を定式化しようとする研究として，例えば，コーパス上での共起関係に着目したものがある[3]．しかし言葉の意味は単一の方式で測れるものではない．そこで我々は，言葉の意味の類似性／関連性を次の3種類に分類し，それぞれを異なった方式を使い，最終的に統合することで柔軟な意味照合の実現を目指す．

1. 類似照合　（類義／同義）
2. 連想照合　（修飾・被修飾）
3. 因果照合　（原因・結果）

類似照合は（奇麗，美しい）の様にほぼ同じ意味を表す置き換え可能な単語の組を，連想照合は（奇麗，画質）の様に修飾・被修飾の関係にある単語の組を，因果照合は（携帯性，気軽）の様に「携帯性に優れていれば気軽に使える」といった原因と結果の組を，それぞれ対応させる．

連想照合・因果照合に関しては，表層的因果知識ベース[4]等を使うことを現在検討中である．以下では，類似照合に関しての実験・考察を行う．

3．概念ベースを用いた類似照合

3.1　概念ベース

類似照合には単語の類似性判別技術が必要となる．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」[5]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構造上次の制約がある．概念ベースにおける類似度は相対的なものであり，ある単語を基準として，そこから他の単語への類似性の順序をつけるにすぎない．言葉を換えれば，「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」を判断することはできるが，「単語A，Bは互いに似ているか否か」を判断するための類似度閾値を定めることが出来ない．これは，ベクトル空間の基底に単語を使っているため，厳密な意味ではその直交性が確保されていないことに起因する．これにより，単語ごとに固有の類似度の尺度を持つことになり，一意の類似度閾値をもって類似性判別を行うことが出来ないのである．

以下，与えられた閾値の元に，単独の単語組に対して「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．

3.2 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．

単語 w, w' に対し，通常の類似度を sim(w,w') とする．w と全単語の類似度の相加平均を $\overline{sim}(w)$，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 p に対し次が成立するときに w と w' は（level p で）似ていると定義する．

$$s_w(w') \geq p \quad かつ \quad s_{w'}(w) \geq p$$

この p 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．

3.3 実験

インターネット上のディジタルカメラの記事に出現した語982語と，ユーザの要求に含まれる頻度が高い語10語の9,820組に対し，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．これを正解とし，概念ベースによる類似照合の精度評価を行う．似ている／似ていないの類似度閾値を変数 p とし，正解54組と類似度が p を上回った組に対し，適合率・再現率を算出した．図1が従来の類似度での結果，図2が今回提案した類似度での結果である．なお，x 軸が類似度閾値 p，y 軸が適合率・再現率であり，類似度の値域の違いにより，x 軸のスケールは図1と図2では異なる．

3.4 考察

DSIUシステムでのルール獲得の為に有効な照合を行うには，再現率が 0.3 以上は必要と考えられる．その際の適合率を見てみると，従来方式では 0.4 であり，これでは獲得するルールの大半が間違ったルールとなってしまう．これは推論には致命的である．しかし，新方式では 0.8 と高い値をとっている．DSIUシステムではルールの妥当性を確率超空間で算出する手法[6]をとるため，若干の間違いは許容される．従って，この方式による類似照合はDSIUシステムにとって十分有効であると言える．

この結果から単語固有の分布が類似性判別に悪影響を与えていたことが判る．我々はさらに，標準化された分布において，閾値のレンジを統一することを検討中である．

4．まとめ

DSIUシステムにおいて必要となる，テキスト中の単語の意味による照合技術，発想的意味照合の提案を行った．また，この発想的意味照合の実現に向けて，概念ベースでの絶対類似性判別方式を提案し，実験を行った．結果，この方式の有効性を確認できた．

今後は，概念ベース自体の精度の向上，および，他のデータベースを用いた連想照合・因果照合の検討を行っていく予定である．

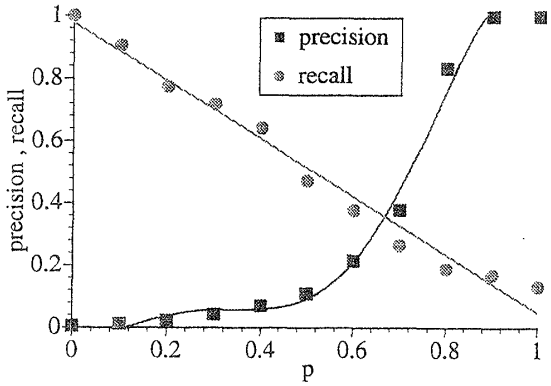

図1　従来の類似度による結果

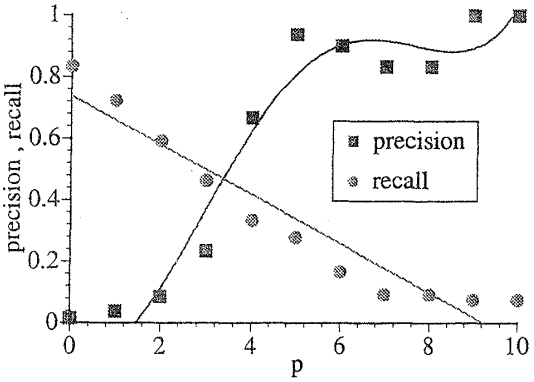

図2　提案方式の類似度による結果

参考文献

1. K.Fujimoto, K,Matsuzawa: Intelligent systems using web-pages as knowledge base for statistical decision making, New Generation Computing, Vol.17, No.4 (1999)
2. 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，AI学会研究会資料 SIG-KBS-9803-9, pp.49-54 (1999)
3. D.Hindle: Noun classification from predicate-argument structures, Proc. of 28th Annual Meeting of ACL, pp.268-275 (1990)
4. 佐藤，笠原，松澤: テキスト上の表層的因果知識の獲得とその応用，信学技報，TL98-23 (1999)
5. 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)
6. 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)