3R－7

辞書ベース連想による場面同定に必要な文脈情報量の推定*

角田 達彦† 田中 英彦‡
東京大学 工学部

1 はじめに

自然言語の問題の一つに、語の多義性解消があげられる。近年、辞書を利用した意味解析の研究例はあるが[1]、一般の辞書は複雑な情報をある程度知識のある人に提供することを目的として作られており、また選好のための重みづけも理由づけなく二値の固定結合で済ませている場合が多いため、知識の常識的側面が欠ける、非常に不安定な動作をする、新たな知識獲得の方針がとれないなどの問題点が現れる。

ここでは外界からの知識獲得への自然な拡張を目標に、連想の性質を細分化し、空間的連想として場面同定に着目する。これは連想推論と矛盾発生を用いるアーキテクチャ PDAI&CD [2] の一部であり、連想記憶に絵情報をもとにした英語辞書 OXFORD-DUDEN Pictoral English Dictionary (OPED) をすべて記憶させモジュール化した。これを用いて場面によって意味の変わる多義語の曖昧性が解消される。日常生活のほとんどの場面を実時間で同定するシステムとしては、初めての試みであり、数語の提示で一意に同定することに成功している。本稿ではその実装方法、文脈量の理論的推測 ([3])、および解析結果について述べる。

2 連想記憶部

PDAI&CD[2] の中で使用している連想記憶 WAVE のここでの簡易形 [3] は図 1(a) のように、単語群を提示し対応するカテゴリ (場面) を選択することを目的とする。各単語の活性値 (ここでは {0,1}) を I_j として、各カテゴリの活性値は以下のように求められる。

$$C_i = f(\sum_j W_{ij} I_j) \quad (1)$$

(シグモイド関数とベイズ確率を用い、$f(x) = \frac{1}{1+e^{-x}}$、$W_{ij} = P(C_i \mid I_j)$ である。) Winner-take-all ネットワークにより

$$C_{i*} = \max_i [C_i] \quad (2)$$

によって最尤カテゴリ (場面) が選ばれる。PDAI&CD では図 1(b) のように、同定された場面に対応する語の語義に従い文が解析され、矛盾発生によりフィードバックがかかる。

3 辞書による場面同定

ここでは連想を Kohonen の分類に従い、(a) 空間的連想 (b) 時間的連想 (c) 類似 (d) 反対の 4 種類に分け、(a) の空間的連想を外界 (Real world) からセンサを通じて獲得することを目標とする。良質な日常生活の空間的連想関係のほとんど

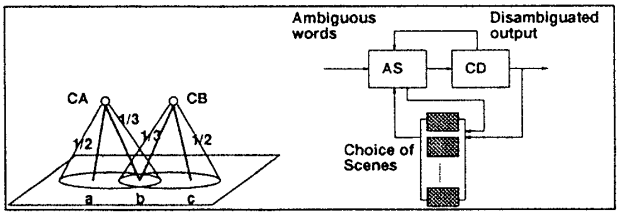

図 1: (a) リンクの重みづけとカテゴリ選択 (b)PDAI&CD のダイアグラム

全てを近似する手段として、OPED を使うことを提案する。OPED は、あらゆる年齢層を対象に、欧米の日常生活の場面が図 2(左) のように絵とそれに対応する英単語 (特に事物名) によって説明されている辞書である。この辞書を使用するにあたり、以下のような仮定、実装を行なった。

- OPED にある日常生活の場面のみを扱い、組合せによりほとんどの場面を近似できると仮定する。
- 構文情報は使わず、OPED の単語のみ扱う。
- 形態素解析は (株) 日本電子化辞書研究所のＥＤＲを使用。

辞書中の 1 場面に複数の単語のセットが対応づけられ (図 2) ている。その単語セットの部分集合が与えられた場合に、その場面が他の場面と区別され同定されるか否かが問題となる。これを次章で解析する。

また多義性に関しては、明示的な多義性を持つ単語のみならず、例えば wall(壁) でも周りに chair や dining table がある文脈 (居間) と、street や car がある文脈 (屋外) では、違う意味合いを持ち、情景に応じた多義性解消が可能になる。

4 解析結果

- 図 3(a) は 1 場面あたりの要素数の分布で、平均 184.2 語。(b) は単語の意味素数、つまり多義性の度合いを示

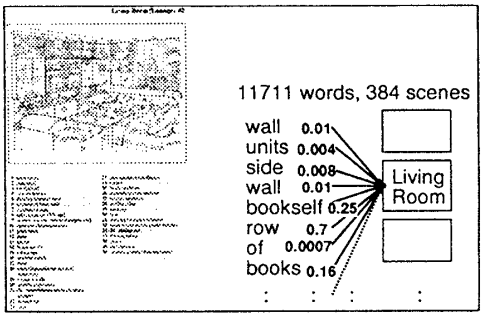

図 2: 居間のシーンと連想記憶 WAVE での重みづけ

*Estimation of Context-information to Identify Scenes by Associative Memory with Pictorial Dictionary
†日本学術振興会特別研究員
‡Tatsuhiko TSUNODA, Hidehiko TANAKA
Faculty of Engineering, University of Tokyo

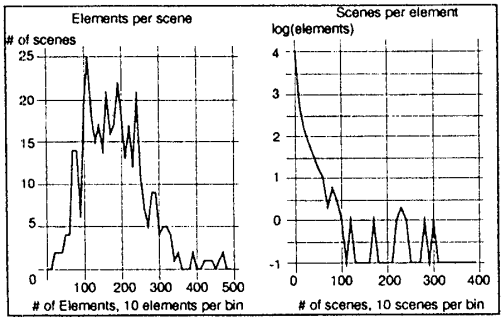

図 3: (a) シーンあたりの要素数の分布 (b) 一要素に関係するシーン数の分布

表 1: 連想記憶上の辞書の解析結果

Total # of scenes	384 scenes
Registered # of words	27,500 words
Total # of words	11,711 words
Average # of words / scene	184.2 words
Max # of words in one scene	478 words
Required # of words to identify scenes at 90% ratio	5 words
Theoretical estimation of required # of words to identify scenes at 90% ratio	2 words

す。100 場面以上に現れる単語は 'a'、'the' など (10 種類) で、場面同定には不要だが、公平さのため用いている。

- 場面の一部を入力したとき、その単語数に対する場面同定確率を図 4(a) に示す。理論的シミュレーション (参考文献 [3] を参照) も表示してある。その基本的パラメータ値に関しては、表 1のように、OPED を解析した結果を用いている。どちらも全解探索は事実上無理のため、Monte-Carlo による 1,000 回の試行の結果である。また場面同定に必要な単語数の分布を図 4(b) に示す。
- 図 4により、完全セットに近ければ近いほど、想起確率は上昇することがわかる。約 90%の認識率を達成するには、5 語示すだけでよい。すなわち文脈として 4 語示すだけで十分であることがわかる。これは場面あたりの平均単語数約 184 に比べて極めて小さい。
- 理論的予想と解析結果の違いの原因は、(1) 'a'、'the' などの無意味語 (2) 乱数近似の不当性があげられる。これを裏付けるものとして、2 つの場面間で重なる単語数の分布を図 5に示す。OPED-2 というのは (1) を削除した結果であり、OPED との違いは顕著に現れるものの、Theory との違いを埋めるにはほど遠い。つまり、どちらの原因も本質的であるが、自然言語の連想性の偏りの存在により、理論的予測をするだけでは不十分で、実際の解析が重要であることが帰結される。
- 辞書と連想記憶の解析結果と理論的予測を表 1にまとめる。辞書の登録語数 27,500 と解析時の全単語数 11,711 の違いは、形態素解析の違いによる。前者では 'research laboratory' などの複合語は一つの独立した単語として与えているが、後者では複数の単語として分けるためである。通常の文では分かれて出現することが多いため、後者の方法を用いている。

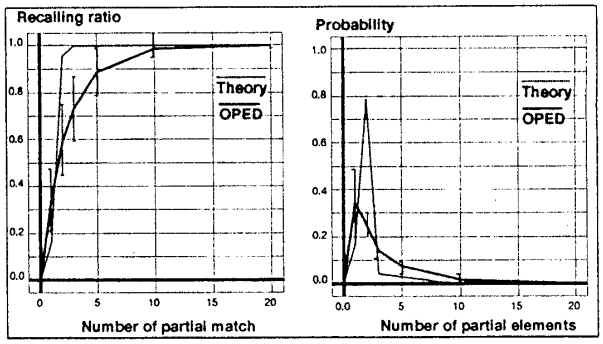

図 4: (a) 部分要素数に対する想起確率 (b) シーン同定に必要な要素数の分布

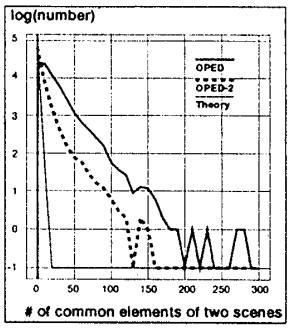

図 5: 二つのシーンに共通に含まれる要素数の分布

5 おわりに

自然言語の多義性解消のための一要素技術として日常生活のほとんどの場面を扱った辞書 OPED を連想記憶に実装し、空間的連想記憶による場面同定を実時間で行なうモジュールを初めて構築し、解析を行なった。約 4 語を文脈として保持するだけで 90% の同定率が得られている。このモジュールは将来の画像理解技術と自然言語処理の統合のための接点として用いることが想定されている。今後は、新たな場面獲得の具体的方法、そして上記の残りの 3 種類の連想関係の取得方法が課題としてあげられる。

この研究の一部は、文部省科学研究費の助成による。また EDR 電子化辞書の使用許可に対して (株) 日本電子化辞書研究所に、そして日頃討論させて頂いている電子技術総合研究所の新情報計画室の方々に感謝の意を表する。

参考文献

[1] N.M.Ide and J.Veronis. Extracting knowledge bases from machine-readable dictionaries : Have we wasted our time? In *KB & KS 93*, pp. 257–266, 12 1993.

[2] T.Tsunoda and H.Tanaka. Semantic ambiguity resolution by parallel distributed associative inference and contradiction detection. In *Proceedings of IJCNN, Nagoya-93, vol.1*, pp. 163–166, 10 1993.

[3] 角田達彦, 田中英彦. 連想推論における逐次学習方式の定式化とその評価 – 曖昧性解消に必要な文脈情報の定量化. 情報処理学会第 4 7回全国大会, Vol. 2, pp. 35–36, 1993.