ACL2013 参加報告（その３）
— マルチモーダルな意味理解 —

岡崎 直観[†,††]

1　はじめに

近年，言語と物理世界を結び付けて意味解析や言語生成を行おうとする研究が注目を浴びつつある．タスク設定が若干異なるが，グラウンドされた言語獲得（grounded language learning/acquisition），物体・シーン認識（object and scene recognition），画像検索（image retrieval），ヒューマン・ロボット・インタラクション（human robot interaction）などは，言語以外のメディアの情報・環境と言語処理を統合した「マルチモーダルな意味理解」の研究として捉えることができる．本報告では，画像や動画などのマルチモーダルな情報を用いた意味理解，シーン理解，シンボル・グランディングに関して，ACL-2013 で発表された論文を紹介する．なお，ACL-2013 でベストペーパーに選ばれた論文 (Yu and Siskind 2013) の解説については，高瀬氏の参加報告を参照されたい．

2　論文紹介

2.1　視覚的属性に基づく意味表現のモデル

最初に紹介する論文は “Models of Semantic Representation with Visual Attributes” (Silberer, Ferrari, and Lapata 2013) である．この論文を発表した Edinburgh 大のグループは，画像に対する自動キーワード付与 (Feng and Lapata 2008, 2010a)，画像と言語に基づく単語の意味のモデリング (Feng and Lapata 2010b; Silberer and Lapata 2012) に取り組んでおり，本研究は後者の研究の最新版として位置づけられる．従来研究に対する本研究のアイディアは，画像から得られる低レベルな特徴量（色やテクスチャなど）から高レベルな属性値（「口がある」「食べる」など）を予測し，その属性値を単語の意味の特徴空間に用いることである．

図 1 に提案手法の概要を示した．分布的意味論（distributional semantics）では，単語の意味

[†]東北大学大学院情報科学研究科, Graduate School of Information Sciences, Tohoku University
[††]JST 戦略的創造研究推進事業「さきがけ」研究員, PRESTO, JST

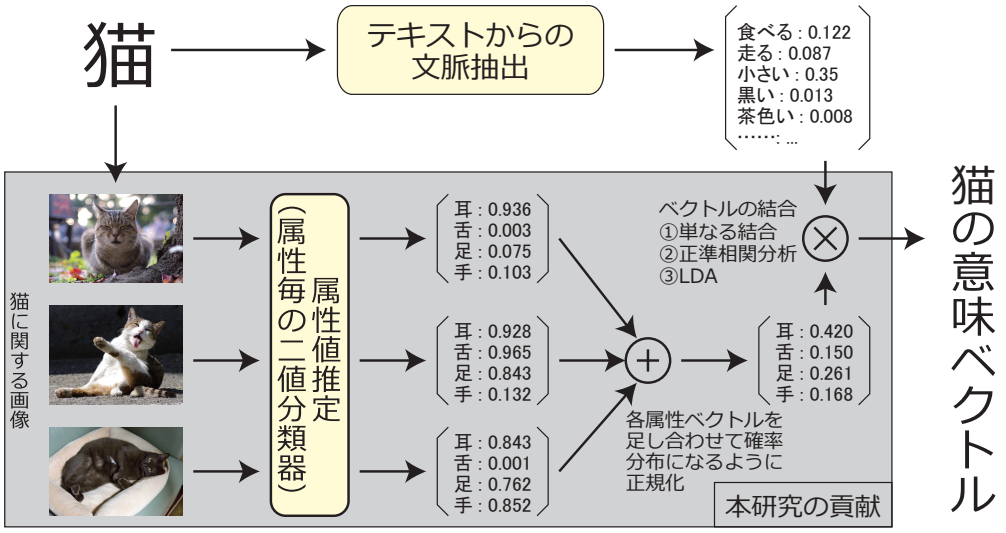

図 1　視覚的属性に基づく単語（猫）の意味表現のモデル化

を表現するために，その実体・概念の物理的・機能的属性を表現したベクトルを用いる．例えば，「猫」に関して「猫は食べる」「猫は走る」「猫は小さい」「猫は黒い」「猫は茶色い」などの属性を手作業で用意したり，大規模なコーパスから自動獲得して，猫の意味を近似するベクトルを得る．このような属性値を実体・概念の画像から自動的に獲得するため，著者らは ImageNet (Deng, Dong, Socher, Li, Li, and Fei-Fei 2009) の 69 万画像に対して 541 種類の属性値の有無を付与し，541 個（属性毎）の二値分類器を SVM で学習した (Farhadi, Endres, Hoiem, and Forsyth 2009)．同一の実体に関する複数の画像に対する属性推定値をベクトル化し，そのベクトルを平均化・正規化することで，実体に関する画像群を属性値ベクトルに変換する．さらに，テキストから得た属性値ベクトル，画像から得た属性値ベクトルを統合する手法として，単なる結合，正準相関分析，LDA の 3 つを提案している．

実験では，与えられた単語に関連する単語を順位付きで出力するタスクにおいて，単語の意味モデルで自動的に推定した結果と，人間が作成した正解を比較することで，意味表現モデルの良し悪しを評価している．実験結果では，テキストと画像の両方の属性値を用いることで，単語の類似性の推定精度が向上すること，色やテクスチャなどの画像の低レベルな素性ベクトル

を用いるよりも，画像を高レベルな属性値に変換した方が，よい性能を示すこと等が報告されている．

2.2 動作記述を動画にグランディング

次に紹介する論文は，“Grounding Action Descriptions in Videos” (Regneri, Rohrbach, Wetzel, Thater, Schiele, and Pinkal 2013) である．この論文は *Transactions of the Association for Computational Linguistics (TACL)* のジャーナル論文として出版された[1]．論文の著者は，計算意味論を専門にする Saarland 大学と，画像処理を専門にするマックス・プランク研究所のメンバーで構成されている．

本研究では，動画中の画像情報を動詞の意味モデルと統合するため，テキスト記述と動画を対応付けた TACOS コーパス（Saarbrucken Corpus of Textually Annotated Cooking Scenes）を構築した．TACOS コーパスの構築には，MPII Cooking Composite Activities (Rohrbach, Regneri, Andriluka, Amin, Pinkal, and Schiele 2012) に収録されている動画が用いられている．この MPII コーパスには，「きゅうりを切る」などの料理に関する 41 種類のタスクが，212 件（1 動画あたり平均 4.5 分）の動画として収録されている．それぞれの動画には，動画中に出てくる動作（「切る」など）と，その動作に関与する物体（「にんじん」「手」など）が，その動作の開始時間と終了時間と共にテキストで付与されている．本研究では，MPII 中の 127 件の動画に対して，その動作の内容を文で記述したものを追加し，TACOS コーパスとした．

さらに，TACOS コーパスに含まれる動作記述のペアのうち，900 ペアに対して，3 人の被験者が記述内容の類似度を 5 段階でスコア付けし，動作記述の類似性に関する評価データを構築した．この評価データを用い，テキスト情報のみで記述の類似度を測る 2 手法，動画情報のみで動作の類似度を測る 2 手法，及びその組み合わせ手法の性能を評価した．実験結果により，動画情報を用いた手法は動作の類似性を測定するのに効果的であること[2]，テキストと動画の両方の情報を用いることで，類似度の測定の性能が向上することが確認された．

2.3 知覚と構文解析の同時学習

最後に紹介する論文，“Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World” (Krishnamurthy and Kollar 2013) も TACL の論文である．論文の著者らは，その所属などから MIT の Tom Mitchell 氏のグループに近いメンバーと推測される．

本論文は，図 2 の左上に示した環境 d と，右上に示したクエリ z を与えたとき，そのクエリ全体が指す物体（群）γ を返すというタスクを通して，言語表現と環境中のシーンの対応付け

[1] ACL 2013 の参加者からは，通常の ACL のセッションよりも TACL セッションの論文の方が面白かったという声が聞かれた．

[2] 動画情報を用いて動作の類似度を測る手法では，今のところ動きに関する特徴量（勾配など）のみを用いており，物体の色や形などの特徴を捉える特徴量を用いていないため．

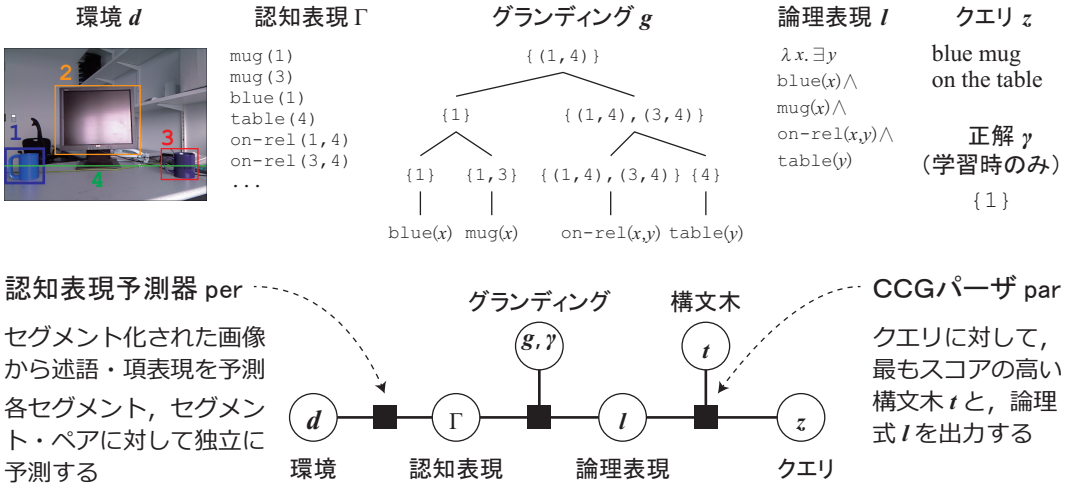

図 2　Logical Semantics with Perception (LSP) の概念図

を学習することを目標としている[3]．著者らはこのタスクのように，言語表現と環境中のシーンの意味解析を同時に行うことを Logical Semantics with Perception (LSP) と呼んでいる．本研究の特徴として，クエリの意味解析（クエリ文から論理式 l への変換），および環境の意味解析（環境から認知表現 Γ への変換）の部分に関する訓練事例を与えず，環境 d，クエリ z，クエリが指す物体 γ のみを教師信号として学習する点が挙げられる．言い換えれば，本研究は意味解析（CCG パーサのモデルと認知表現予測器のモデル）の結果を隠れ変数とした弱教師有り学習として定式化されている．これは人間が環境から言語を獲得していく状況を模擬しており，挑戦的な問題設定である．評価実験では，シーン認識と地図認識の 2 つのデータセットを用い，既存の最高精度の手法 (Matuszek, FitzGerald, Zettlemoyer, Bo, and Fox 2012) よりも高い性能が得られることを実証した．また，意味表現の正解を与える完全な教師あり学習と比較して，本研究の弱教師あり学習の手法でも遜色ない性能が得られることを示した．

3　おわりに

本報告では，マルチモーダルな意味理解に関する研究の最新動向を紹介した．シンボル・グランディングを目標に掲げた研究は古くからあるが，近年の画像処理の進歩により，急に盛り上がりを見せているように感じる．実際，今回紹介した論文とベストペーパに選ばれた論文 (Yu and Siskind 2013) では，画像処理を専門とする研究者が著者として名を連ねており，研究の完

[3]図 2 は元論文の Figure 2 を参考に作成した．図中の写真は，著者らのウェブサイト上で公開されているデータから引用した． http://rtw.ml.cmu.edu/tacl2013_lsp/

成度も高い．これらの研究は，言葉を理解するコンピュータを作るという，言語処理が長年取り組んできた究極のテーマと関連が深く，今後の動向から目が離せない．

謝辞

第 5 回最先端 NLP 勉強会[4]にて NAACL-2013, ACL-2013, TACL の最新動向を紹介して頂いた皆様，電子情報通信学会・言語理解とコミュニケーション研究会（NLC）やその後の懇親会において議論をして頂いた皆様に感謝いたします．

参考文献

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). “ImageNet: A Large-Scale Hierarchical Image Database.” In *CVPR 2009*.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). “Describing Objects by their Attributes.” In *CVPR 2009*.

Feng, Y. and Lapata, M. (2008). “Automatic Image Annotation Using Auxiliary Text Information.” In *ACL-HLT 2008*, pp. 272–280 Columbus, Ohio.

Feng, Y. and Lapata, M. (2010a). “Topic Models for Image Annotation and Text Illustration.” In *NAACL HLT 2010*, pp. 831–839 Los Angeles, California. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010b). “Visual Information in Semantic Representation.” In *NAACL HLT 2010*, pp. 91–99 Los Angeles, California.

Krishnamurthy, J. and Kollar, T. (2013). “Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World.” *Transactions of the Association for Computational Linguistics (TACL)*, **1**, pp. 193–206.

Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012). “A Joint Model of Language and Perception for Grounded Attribute Learning.” In *ICML 2012*.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). “Grounding Action Descriptions in Videos.” *Transactions of the Association for Computational Linguistics (TACL)*, **1**, pp. 25–36.

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., and Schiele, B. (2012). “Script Data for Attribute-based Recognition of Composite Activities.” In *ECCV 2012*.

[4] http://www.logos.t.u-tokyo.ac.jp/snlp5/

Silberer, C., Ferrari, V., and Lapata, M. (2013). “Models of Semantic Representation with Visual Attributes.” In *ACL 2013*, pp. 572–582 Sofia, Bulgaria.

Silberer, C. and Lapata, M. (2012). “Grounded Models of Semantic Representation.” In *EMNLP-CoNLL 2012*, pp. 1423–1433 Jeju Island, Korea.

Yu, H. and Siskind, J. M. (2013). “Grounded Language Learning from Video Described with Sentences.” In *ACL 2013*, pp. 53–63 Sofia, Bulgaria.

略歴

岡崎 直観（正会員）：2007 年東京大学大学院情報理工学系研究科・電子情報学専攻博士課程修了．同年，東京大学大学院情報理工学系研究科・特別研究員．2011 年より，東北大学大学院情報科学研究科准教授．自然言語処理，テキストマイニングの研究に従事．情報理工学博士．情報処理学会，人工知能学会，ACL 各会員．

（2013 年 8 月 1 日依頼）
（2013 年 9 月 20 日受付）