4R-7

自然言語テキストからの概念記述言語CDL構造の半自動抽出

山口 清弘† 石塚 満†
† 東京大学情報理工学系研究科 創造情報学専攻

1 はじめに

電子化された膨大なテキストデータから情報を取り出し構造化することで，情報検索や情報抽出に役立てようとする研究が広く行われている．横井らは，文全体を一つの意味構造に翻訳することを目的として，概念記述言語である CDL (Concept Description Language) を設計・開発した [1]．CDL では文中のすべてのエンティティに対して，それがどのエンティティとどのような意味で関係するのかを特定することにより，文の意味を表現する．これは意味計算 (*Semantic computing*) 基盤になると共に，言語横断の情報検索や情報抽出，機械翻訳を実現する基盤となる．

自然言語テキストを完全な CDL 表現へと翻訳するためには，多くの自然言語処理タスクを解かなければならない．特に，CDL のエンティティを同定するために語義曖昧性解消 (*Word sense disambiguation; WSD*)，エンティティ間の関係を同定するために意味役割付与 (*Semantic role labelling; SRL*) の 2 つのタスクを解くことが不可欠となる．実用的には，現状の自然言語処理の精度から考えて，例えば日本語入力システム (Input method editor) などのように，可能性の高い候補を利用者に提示したうえで，そこから正しい候補を選択してもらうというアプローチを取ることになるだろう．そのためには，複数のタスクから得られた結果を候補の一覧として一つにまとめる必要がでてくる．タスクごとに独立した既存手法を組み合わせるというアプローチを取った場合，その手法は自明ではない．また，あるタスクの結果を別のタスクで利用するという形をとると，各タスクでのエラーが徐々に蓄積・拡大されていってしまう (Error propagation) 問題もある．本研究では Markov logic[2] と呼ばれる手法を用いて，この点を解決する．

2 関連研究と本研究の位置づけ

Markov logic は一階述語論理と Markov networks を組み合わせたもので，複数のモデルを同時に扱うためのフレームワークとして近年注目を集めている．意味役割付与と語義曖昧性解消はタスクとして相互に関連性を持つと考えられることから，Markov logic を用いてこの 2 つを同時に行う手法がいくつか提案されている．Meza-Ruiz らは 2009 年に，英語テキストに対して意味役割付与と述語語義曖昧性解消を同時に行う手法を提案している [3]．この手法は 2010 年に Che らによって，述語以外に対しても語義曖昧性解消を行うように改良されている [4]．日本語テキストの場合，2011 年に吉川ら [5] が日本語テキストの述語項構造解析を Markov logic を用いて行っているが，ここで用いられているのは深層格ではなくて表層格である．言語横断的なタスクなどの場合は，広く自然言語一般に通じる深層格の方が有利であるが，日本語テキストにおいて，深層格の意味役割付与と語義曖昧性解消を同時に行う手法は，まだ提案されていない．

3 Markov logic

Markov logic では一階述語論理式 ϕ と実数値の重み ω の組 (ϕ, ω) を考え，この組の集合 L を *Markov Logic Network (MLN)* と呼ぶ．論理式のすべての変数を定数で置き換えるような変数束縛の有限集合 C が与えられたとき，Markov network $M_{L,C}$ を得ることができる [2]．$M_{L,C}$ によって規定される各可能世界 x の確率分布は，式 1 によって与えられる．ただし Z は正規化定数で，$f(c, \phi)$ は c で ϕ の変数を束縛した基底論理式が真となるなら 1，そうでなければ 0 を返す素性函数である．

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{(\phi,\omega)\in L} \omega \sum_{c\in C} f(c, \phi) \right) \quad (1)$$

Metkov network における重みの学習は条件付確率尤度の計算，推論は最大事後確率 (MAP) 推定問題となる．

4 提案手法

本研究では Markov logic を用いて潜在意味解析および語義曖昧性解消を行い，実際に日本語テキストから CDL 表現を抽出できることを検証し，その性能を実験的に評価する．CDL でタグづけされた大規模なコー

Semi-automatic Extraction of Concept Description Language (CDL) Structures from Japanese Texts
†Kiyohiro YAMAGUCHI †Mitsuru ISHIZUKA
†Graduate School of Information Science and Technology, University of Tokyo

パスはまだないので，深層格でタグづけされた日本語コーパスとして EDR 電子化辞書 [6] を利用する．我々の定義した潜在述語および観測述語を表 1 に挙げる．

潜在述語	説明
isArgument(a)	位置 a が項である
isPredicate(p)	位置 p が述語である
hasRole(p, a)	位置 p が位置 a を項に持つ
role(p, a, r)	(p, a) の潜在格は r
sense(i, s)	位置 i の意味は s
観測述語	**説明**
word(i,w)	位置 i の形態素が w
pos(i,p)	位置 i の品詞が p
possibleArgument(i)	位置 i は項になり得る
possiblePredicate(i)	位置 i は述語になり得る
dep(i, j, d)	(i, j) の合成関係ラベル
path(i, j, p)	(i, j) 間の最短パス

表 1: 潜在述語と観測述語

合成関係ラベルは M (修飾)，S (統合)，N (数) および I (複合語) の 4 種類がある．最短パスは構文木上での最短の移動経路で，例えば "UDD" なら「親への移動を 1 回，葉への移動を 2 回」となる．潜在格は，EDR で使われている agent, implement, object などの 27 種類である．

局所論理式・大域論理式については主に Meza-Ruiz らの研究 [3] および Che らの研究 [4] を参考に，日本語テキストおよび EDR で用いられているタグやラベルに沿うよう再構成したものを利用する．

5 実験と評価

EDR 日本語コーパスのなかから，ランダムに選んだ 30,000 レコードを訓練データに，3,000 レコードをテストセットとして利用する．学習および推論にはオープンソースの実装である thebeast*を使用した．

各潜在述語ごとの F 値を表 2 に示す．観測述語の性能に与える影響を観察するため，dep と path をそれぞれ省いた場合 (-dep, -path)，dep のラベルを省いた (-dep*) 場合，推定に対して人手でフィードバックを与えた場合 (+feedback) についても併わせて実験を行った．

	full	-path	-dep*	-dep	+feedback
isArgument	90.9	90.3	90.1	88.8	93.4
isPredicate	91.1	90.8	90.8	89.4	93.4
hasRole	88.1	87.1	87.2	86.0	91.4
sense	65.3	65.2	64.7	62.8	70.9
role	54.8	54.7	52.3	50.3	66.7

表 2: 各潜在述語ごとの F 値

*http://code.google.com/p/thebeast/

dep, path のいずれもが重要な素性として機能し，特に dep の貢献が大きいこと，人手のフィードバックも適切に学習に反映されることが確認できる．

hasRole と role での大きな性能差は，格同定については高い精度で推定できているものの，種類の多い深層格のラベルをまだ上手く推定しきれていないことを意味する．これについては格フレームの知識をより MLN に反映させることで，性能を向上させることができるだろうと考えている．sense についても，先行研究などと比べて概念がより細かい単位になっているため，概念階層をより活用するような論理式を追加することで，より良い手法へと発展させていくことができるだろう．

6 おわりに

本研究では日本語テキストを CDL 表現へ翻訳することを目標に，Markov logic を用いて日本語テキストの語義曖昧性解消および意味役割付与を同時に行う手法を提案した．その性能を実験的に評価し，提案手法が一定の性能を持っていること，特に格同定については高い性能を持つことを確認した．今後の研究では，概念階層・格フレームをより活用することで性能を向上させていくほか，照応解析など CDL 表現を抽出する上で重要な他のタスクも同時に解けるよう，手法を拡張していきたいと考えている．

参考文献

[1] Toshio Yokoi, Hiroshi Uchida, Koiti Hasida, Hiroshi Yasuhara, and Meiying Zhu. CDL (concept description language): A common language for semantic computing. In *Workshop on The Semantic Computing Initiative (SeC2005)*, 2005.

[2] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, Vol. 62, No. 1-2, pp. 107–136, 2006.

[3] Ivan Meza-Ruiz and Sebastian Riedel. Jointly identifying predicates, arguments and senses using markov logic. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 155–163, 2009.

[4] Wanxiang Che and Ting Liu. Jointly modeling WSD and SRL with markov logic. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 161–169, 2010.

[5] Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. Jointly extracting japanese predicate-argument relation with markov logic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 2011.

[6] 日本電子化辞書研究所．EDR 電子化辞書製品版 (第 1.5 版), 1996.