

実世界指向情報構造化支援のための情報抽出技術

渡邊 陽太郎[†] 佐々木 彬[†] 五十嵐 祐貴[‡] 岡崎 直観^{†*} 乾 健太郎[†]

[†] 東北大学大学院情報科学研究科 [‡] 東北大学工学部情報知能システム総合学科 * 科学技術振興機構さきがけ

{yotaro-w, aki-s, okazaki, inui}@ecei.tohoku.ac.jp yuki.i@so.tohoku.ac.jp

1 はじめに

自然災害に対峙するには、災害発生を前提とし、その後の応急対応、復旧・復興を迅速かつ的確に行い、災害による悪影響を極小化する減災力を強化し、災害に対するレジリエンスを高めることが必要であり、そのためには、防災情報の有効活用が必要不可欠である。国際的にも災害対応を含めた危機管理の情報マネージメントの標準化が進められており、我が国でも標準に準拠した災害対応体制とそれをサポートする情報システムの整備が強く求められている。東日本大震災を契機として、従来の紙やホワイトボードを主体とした防災情報の集約・共有方法から脱却し、国際的に標準化された災害対応のための情報システムの導入による迅速で的確な災害対応の実現が望まれている。また、災害対応の記録をアーカイブとして蓄積し、今後の教訓を引き出す仕組みを現場に導入する方法論とそれを実現する技術開発への重要性が高まっている。

防災情報を電子化し構造化するという問題は、言語処理の立場から考えると、実世界に存在する情報構造化問題の良い例題になっている。防災情報の種類は、人や建物の被害状況、避難所、物資、ライフラインなど多岐にわたり、これらの中には場所名、状況、数量などの抽出や電気使用可否など様々な種類の情報が含まれるため、情報抽出問題と分類問題が混在した複雑な情報構造化問題を解くことになる。また、防災情報構造化支援システムの設計に関して、システムの運用場所によって必要となる情報の構造化形式（データベーススキーマ）は異なる可能性があるため、構造の変更にも柔軟に対応できる汎用的なシステム設計が必要である。一方、場所名の辞書など特定の場所に依存した知識を柔軟に置き換えられる機能も必要である。システムを実際に運用するのは情報科学のエキスパートではない、例えば地方自治体の職員であるため、専門的知識を必要とせず、簡単な操作で運用ができることが必要であるし、システム出力を人手で修正することによってシステム性能が改善していく枠組を備えることも、システムを長期間運用させるためには重要な機能となる。

本稿では、実世界の情報構造化支援の例題として防災情報の構造化着目し、その支援のためのシステム設計・実装について述べる。

2 防災情報のデータベース化

災害時に報告される防災情報は、死者・行方不明者などの人的被害、建物の被害、公共交通機関の被害、災害対応状況、避難所の開設・避難状況、ライフライン、食料や物資に関する情報など多岐にわたる。これらの防災情報をデータベース化し管理・共有するためには、必要な情報を格納できる適切なデータベーススキーマ（テンプレート）の設計が必須である。これは、防災情報の種類ごとに異なるテンプレートが用いられ、それぞれのテンプレートに該当するレコードが、情報共有やとりまとめ等に利用される。自然言語処理の課題は、防災情報がテキストで与えられた時に、定義されたテンプレートの中から該当するものを選択し、テンプレートの各項目に対して適切な情報を埋めることである。防災ドメイン特有の性質として、複数の防災情報がまとめて報告される場合があるため、一つ

のテキストから複数のレコードを生成する必要がある例えば、“本公、避難者 148 人、電気ストップ。仙翁寺、炊き出しあり、水なし、ヘルプ。寺谷、湧き水使用、食糧なし、自家発燃料なし。“という防災情報には、本公、仙翁寺、寺谷という3箇所の避難所についての言及があるため、“避難所名:本吉公民館、避難者:148、電気使用:否”、“避難所名:仙翁寺、水道使用:否”、“避難所名:寺谷、水道使用:否、電気使用:否”の3種類のレコードを生成しなければならない。そのために、それぞれの避難所に関する情報が含まれる箇所を同定した上で適切な情報を抽出する必要がある。また、防災情報には、本公（＝本吉公民館）のように、地域特有の略称が含まれていることもあるため、場所に関する知識の活用が可能な設計も必要である。

我々が開発している防災情報データベース化システムの全体像を図1に示す。ここでは、音声で入力される防災情報をテキスト化し、様々な言語解析技術を利用して言語情報を付加する。次に、構造化テキストを入力とし、事前に定義されたテンプレート集合と情報抽出技術を利用して、テキストに対応するデータベースレコードを生成する。また、場所情報処理により、略称表現などを適切な場所表現に変換し、各表現が指し示す実世界の位置を同定する。これにより、災害状況を地図上で俯瞰することが容易となり、災害対応の意思決定を円滑におこなうことが可能となる。

以降、まずは防災情報管理のためのテンプレート設計について述べる。次にデータベース化のための言語解析技術、本ドメインで特に重要な場所情報の処理技術について触れた後、テキストからレコードを自動生成する手法について述べる。その後、本システムを社会実装すること目的とした様々な工夫について述べる。

2.1 テンプレート設計

災害対応のための意思決定および情報のとりまとめのために必要十分な情報を含むテンプレートがどのようなものかは明らかでない。我々は、様々な防災関連の文書や災害対応ログ、関連研究等の知見などをふまえ、表2.1に示されている15種類のテンプレートを策定した^{*1}。このテンプレートの妥当性は、米国の標準化された危機管理システムであるICS (Incident Command System) の項目^{*2}、東田らにより作成された檀原テンプレート [2]、京都府テンプレート、APPLIC (全国地域情報化推進委員会) 防災業務アプリケーションユニッ

*1 テンプレートの項目は、共通項目と個別項目に分けられる。共通項目は、ID、関連ID、受信日時、受信手段（電話、手渡、無線、口頭）、相手先、受付者名、信憑性（未確認、信憑性有、信憑性無、確認済）、重要性（高、中、低）、対応・未対応、受信内容、処理の11項目である。個別項目は、物的被害:ライフラインの場合、日時、地域名、施設名、ライフライン種別（電力、ガス、上水道、下水道、固定電話、携帯電話）、復旧状況区分（停止、作業中、復旧済）、対応方針の6項目が含まれる。括弧がある項目の値は、括弧内のいずれかから選択する。

*2 4種類の部門それぞれのUnitの中から外部から報告される防災情報に関連する項目を厳選し、対応関係を調べた。具体的には、資源管理部門のSupply Unit、Food Unit、Ground Support Unit、Facilities Unit、Medical Unit、情報計画部門のSituation Unitである。



図 1: 防災情報データベース化システム

テンプレート	サンプル文
ハザード情報	本吉分署です。津波第2波が襲来するので更に高台に避難する事を周知してください。
人的被害	津谷幼・馬籠幼・小泉幼、預かりけが人なし。
物的被害:公共施設等	本吉総合支所前、交差点信号機、停止。
物的被害:公共交通機関	J R小泉駅前が冠水し、通行止めとなっております。
物的被害:ライフライン	三日町・八日町、通電。
物的被害:火災	西川内、炭釜、火災、鎮圧状態。
道路・交通規制	国道 45 号、小泉・大谷、冠水。
医療	三峰病院や大友病院でも患者を受け入れている。
避難状況	ホテル望洋に 80 名ほど避難している。可能であれば電気をお願いしたいとの話があった。
避難所	林の沢会館・坊の倉会館、避難者なし、受け入れ準備 OK。
人的資源	山形自衛隊、70 人到着。響高校、テント張る。
物的資源	現在稼働しているスタンドは、JA、安全、気仙沼商會、丸和。
広報:交通機関	南町交番交差点前から河北ビル交差点前の河原田通り通行可能。
広報:ライフライン	本日の地域は、旭が丘学園、食料確保の観点からマイヤ、デイルーポートを通電したい。
広報:暮らしの情報	市民にごみ焼きをしないよう周知して欲しい。

表 1: テンプレート一覧

ト標準仕様^{*3}等と比較し、その妥当性を確認している。また、各テンプレートについて、気仙沼市本吉支所災害対応記録、気仙沼市災害対策本部概要記録、危機管理監メモ、気仙沼市地域防災計画等の文書に基づき、データ作成（サンプル文の作成およびデータベースレコードの各項目に対するアノテーション）も進めている。

2.2 言語解析

防災情報テキストは、様々な言語解析を適用することでデータベース化のために必要な情報を得る。具体的には、形態素解析 (MeCab[5])、係り受け解析 (CaboCha[4])、述語項構造解析 (ChaPAS[14])、固有表現抽出、数量・時間表現抽出 (normalizeNumexp^{*4})、イベント抽出、関係抽出が適用される。固有表現抽出は、人名、地名、施設名、名称から場所を特定可能な組織名 (企業名等) の認識を目的に利用する。固有表現抽出器は、関根の拡張固有表現階層 [13] に基づき固有表現情報が付与された日本語コーパス [15] から、人名、地名、施設名、企業名に該当するタグだけを残し、条件付確率場 [6] を

用いて学習、構築したモデルを利用している^{*5}。イベント抽出は災害関連表現を抽出する目的で適用される。ここで、災害関連表現とは、“被害、配送、受入、復旧、確認、問合せ、要求、状態、病気、移動、死亡”等、防災情報によく出現するイベントを表す表現のことを指す。例えば、被害に該当するものとしては“亀裂”や“陥没”などがある。これらの表現の抽出は人手で整備した辞書を用いる。関係抽出はモノと数量の関係、ここでは主に<避難者一人数>、<物資一数量>の関係の抽出を目的に適用する。これには、木構造を扱えるパターンマッチャーを利用し、係り受け木上でモノと数量の関係を捉えるパターンを作成し抽出している。

2.3 場所情報処理

災害状況の把握および災害対応の意思決定を行う上で、各地域の被害状況を俯瞰的に把握することは必要不可欠であり、そのためにはどの場所にどの程度の被害が発生しているかを可視化することが必要となる。ここでは、防災情報に含まれる場所表現が、実際のどの領域または地点を指しているかを同定する問題を扱う。ある場所表現の領域や地点を同定するためには、場所に関する知識が必要不可欠である。本研究では、Yahoo!ロコ、i タウンページから地理情報を収集し、利用している。規模は、47 都道府県でそれぞれおよそ 1236 万件、835 万件であり、各エントリは場所名称、住所、ジオコード情報で構成されている。収集したデータから、以下の二種類のデータベースを構築した。

1. 地名データベース 地名から、都道府県名を含む正式な住所を検索するためのデータベース。例えば、“本吉”という表現から“宮城県気仙沼市本吉”を得る。
2. ランドマークデータベース 施設などのランドマークの曖昧性を解消し、住所を同定するためのデータベース。例えば、“本公”から“本吉公民館”を同定する。

固有表現抽出で地名と判定された表現については 1 を、施設名や組織名と判定された表現については、2 のデータベースを利用して、実際の位置を特定する^{*6}。

^{*3} <http://www.applic.or.jp/URN/APPLIC-0009-2010/APPLIC-0009-2010-10/APPLIC-0009-2010-10-00.pdf>

^{*4} <http://www.cl.ecei.tohoku.ac.jp/~katsuma/software/normalizeNumexp/>

^{*5} 実装は CRFsuite [11] を利用している。

^{*6} 入力の場合所表現を x 、データベース D の各要素 y' とした時、最尤の表現は、式 $\hat{y} = \arg \max_{y' \in D} \cos(f(x), f(y')) + h(x, y')$ で求める。ここで、 $\cos(f(x), f(y'))$ はコサイン類似度、 $f(x)$ と $f(y')$ は x と y'

2.4 データベースレコード生成

言語解析された防災情報テキストから、データベースに追加するレコードを生成する。前述のように一つのテキストから生成されるレコードは複数の可能性もあるため、テキスト x が与えられた時のここでの出力はレコード集合 r である。各レコード $r_i \in r$ は、テンプレート s とテンプレートの持つ各属性値 $a^s = \{a_1^s, a_2^s, \dots\}$ のペアから構成される ($r = \langle s, a^s \rangle$)。したがって、レコードを生成するためにはレコード生成数、テンプレート分類、各属性値の同定の全てをおこなう必要がある。この問題は一種の情報抽出問題とみなすことができるが、“電気使用可否”のように、有限の集合（この場合可か否）から適切な値を選択するという分類問題も含んでいるという点で従来の情報抽出問題とは異なる。

提案手法は、(1) テンプレート分類、(2) トリガ表現認識、(3) テキスト分割、(4) テンプレート埋めの4つの処理によりレコードを生成する。以下、各処理について詳述する。

2.4.1 テンプレート分類

テキスト x が与えられた時、それに対応するテンプレート \hat{s} を、ロジスティック回帰モデルを用いて得られる条件付確率 $p(s|x)$ が最大となるテンプレート \hat{s} を選択する。実装は、liblinear-java^{*7}を利用している。分類の手がかりには、bag-of-words 素性、単語の確率的クラスタリング情報 [3] を利用している。

2.4.2 トリガ認識

各テンプレートには、レコードを生成するトリガとなる属性が少なくとも一つ存在する。例えば、避難所テンプレートでは、避難所ごとに情報が格納されるため、2節の例のように“避難所名”属性とレコードには対応関係がある。ここでは、ある属性が出現すれば必ず対応するレコードが生成されるような属性をトリガ属性、トリガ属性に該当する表現をトリガと呼ぶ。本手法は、あるテンプレート s の属性 a_k^s がトリガ属性であるかどうかを、サンプルデータに基づき判断する。具体的には、サンプルデータ中の各レコードで各属性値が埋められている頻度と複数レコードが生成される事例では異なる値が埋められている頻度から属性 a_k^s が埋められる確率 $p(\text{fill}|a_k^s, x)$ を計算し、この確率値が閾値以上となる属性をトリガ属性と判定する。トリガの認識は、後述するテンプレート埋めモデルを利用してテキストに対して属性のラベル付けをおこない、トリガ集合 $t = \{t_1, t_2, \dots\}$ を得る。

2.4.3 テキスト分割・生成

ここでは、各トリガ $t_i \in t$ に関する情報を含む部分テキスト $x'_i \in \mathcal{X}'$ を、テキスト x を分割することによって得る。テキストの分割は、下記の3種類の規則に基づきおこなう。

1. 複数文から構成され、それぞれに異なるトリガ表現が存在する場合、それぞれの文を異なるグループとして扱う。
2. 文中にトリガ表現の並列句がある場合^{*8}は、その並列構造を分解し、要素ごとに文を生成する。
3. 一文中でトリガと属性値がペアで並列構造になっている場合、ペアごとに分解して文を生成する。

の unigram, bigram により構成されたベクトルである。 $h(x, y')$ はヒューリスティックルールに基づくスコアを出力する関数であり、 x と y' の末尾文字列が一致していれば 0.1 を返す。コサイン類似度の高い候補を効率的に列挙するため、SimString[12] を利用している。

^{*7} <http://liblinear.bwaldvogel.de/>

^{*8} 並列関係は、CaboCha の係り受け関係を分類することで同定している。ここでは、ChaPAS の解析結果を利用している。

これらの規則を、適用できるまで分割を繰り返し、分割された部分テキスト集合 \mathcal{X}' を得る。ここで、 $|\mathcal{X}'| = |t|$ である。

2.4.4 テンプレート要素埋め

テンプレート s 、トリガ t_i と対応する部分テキスト x'_i が与えられた時に、属性値 a_k^s に対応する値を x'_i の情報抽出・分類によって得て、レコードを生成する。各属性の属性値は、そのタイプに応じて二種類の方法のいずれかによって埋められる。

情報抽出問題: テキスト中の部分文字列が属性値に対応する問題。この問題に該当する属性値は、各テンプレートごとに条件付確率場 $p(y|s, t_i, x'_i)$ をサンプルデータから学習し、各形態素に対して属性値のラベル付けをおこなうことで抽出する。

分類問題: 規定された値のセットの中から適切な値を選択する問題。この問題は、それぞれの属性ごとにロジスティック回帰モデル $p(a_k^s|s, t_i, x'_i)$ を作成し、適切なラベルを分類する。

上記の解析によって得られたレコード集合 $r = \{r_1 = \langle s, a_1^s \rangle, r_2 = \langle s, a_2^s \rangle, \dots\}$ をデータベースへ登録する。

2.5 社会実装へ向けて

本システムの社会実装に向けて、1節で述べた3つの要件に対して、以下の点を工夫している。

汎用性・ポータビリティ 異なるテンプレートの利用にも対応できるように、データベース化部分を機械学習化し、データさえ用意すれば人手で規則を書くこと無くシステムを構築することができる。また、場所情報辞書を置き換えれば、異なる自治体での利用を可能にしている。

低コストでの運用の実現 簡単なマウス操作で効率的な情報の修正・登録が可能なユーザーインターフェースを整備している (図 1)。また、解析誤りの修正結果を使用してモデルの再学習をおこなうことでシステムを改善できる処理フローを取り入れている。

将来的には、防災情報の入力時のボトルネックを解消するための音声認識システムの導入、テンプレートの各要素ごとに最適な手がかりを得るための素性選択アルゴリズムの導入などを予定している。

3 評価実験

2.1節で述べたデータセットを利用して、レコード生成の評価実験をおこなう。使用するデータは、表 2.1 で示したテンプレートのうち、サンプルデータがある 8 テンプレートである。データ数は全 329 事例である。評価には leave-one-out 方式を採用し、テンプレート分類とテンプレート埋めそれぞれの性能を評価する。また、テンプレート埋めの評価では、正解のテンプレートを与える。

各タスクの評価実験の結果を表 3 に示す。テンプレート分類性能は必ずしも十分であるとは言えないが、実際のシステムの運用では N-best の上位の結果を優先的にオペレータに提示し、正しいテンプレートを選択してもらうことを想定している。上位 4-best までの正解率は 97.8% であり、運用上の問題は少ないと考えられる。

テンプレート埋めの情報抽出問題に対する性能は、全体の F 値が 0.36 と十分な性能が得られているとは言えない。物的被害: 火災や避難所は比較的高い数字が得られているが、この理由は、これらのテンプレートでは抽出問題に該当する属性が、火災の場所名め避難所名など、比較的容易に抽出できるものしか無いためである。一方で、その他のテンプレートでは、長い文字列を抽出する必要のある難しい問題が多いため、性能が

表 2: 評価結果.

テンプレート名	テンプレート分類			テンプレート埋め 情報抽出問題			テンプレート埋め 分類問題		
	精度	再現率	F1	精度	再現率	F1	精度	再現率	F1
人的被害	0.70 (39/56)	0.71 (39/55)	0.70	0.37 (28/76)	0.17 (28/165)	0.23	0.80 (40/50)	0.53 (40/75)	0.64
物的被害:公共施設等	0.60 (30/50)	0.67 (30/45)	0.63	0.25 (22/87)	0.15 (22/145)	0.19	0.34 (25/74)	0.23 (25/108)	0.27
物的被害:公共交通機関	0.73 (8/11)	0.53 (8/15)	0.62	0.33 (2/6)	0.06 (2/36)	0.10	0.68 (13/19)	0.50 (13/26)	0.58
物的被害:ライフライン	0.93 (14/15)	0.67 (14/21)	0.78	0.28 (5/18)	0.14 (5/35)	0.19	0.73 (30/41)	0.59 (30/51)	0.65
物的被害:火災	0.80 (8/10)	0.62 (8/13)	0.70	0.72 (13/18)	0.52 (13/25)	0.60	0.62 (8/13)	0.62 (8/13)	0.62
道路・交通規制	0.80 (36/45)	0.84 (36/43)	0.82	0.42 (33/79)	0.19 (33/172)	0.26	0.50 (20/40)	0.28 (20/71)	0.36
避難所	0.89 (119/134)	0.94 (119/127)	0.91	0.69 (140/202)	0.49 (140/284)	0.58	0.83 (109/132)	0.74 (109/147)	0.78
人的資源	0.75 (6/8)	0.60 (6/10)	0.67	0.47 (8/17)	0.23 (8/35)	0.31	0.00 (0/1)	0.00 (0/7)	0.00
合計	正解率: 0.79			0.50 (251/503)	0.28 (251/897)	0.36	0.66 (245/370)	0.49 (245/498)	0.56

低くなっている。例えば、「建物、ある、中は使える状態でない」（物的被害:公共施設等-被害内容）や「東北電力、電源車、電気復旧。119 番通報、受理可能」（物的被害:ライフライン-対応方針）といった文字列の抽出が必要となる。システムの出力結果を調査したところ、正解:「自衛隊、通行できない」、システム出力:「通行できない」のように部分一致している結果が散見された。そこで部分一致も正解とした場合の性能を評価したところ、精度:0.75, 再現率:0.42, F 値 0.54 であった。

分類問題については、情報抽出問題と比較すると高い性能が得られているものの、まだ十分であるとは言い難く、特に物的被害:公共施設等の性能が低い。この理由は、分類問題の中に比較的種類が難しい項目が含まれている点が挙げられる。その一つである建物被害状況区分は、{全壊・全焼, 半壊・半焼, 一部破損, 床上浸水, 床下浸水, その他}の中から一つ選択する分類問題であるが、被害状況の表現の仕方は様々なバリエーションがある一方で、学習データに含まれるバリエーションが限られていたため、十分な性能が得られなかったものと考えられる。

情報抽出問題、分類問題の双方において、総じて学習データ量の不足が起因すると思われる誤りが多く見られたため、今後データの規模を拡大していきたいと考えている。

4 関連研究

本研究で扱っているテキストを入力とした情報のデータベース化は、情報抽出の問題の一種とみなすことができる。情報抽出の問題は、Automatic Content Extraction (ACE) *9をはじめ、医療ドメイン文書からの情報抽出 [9] など広く取り組まれている課題である。この問題は構造予測問題の一種とみなすことができ、条件付確率場などの系列ラベリングに基づく手法 [1]、関係から得られるグラフに基づく手法 [8]、構造化パーセプトロンを利用して、トリガ表現の分類とテンプレート要素埋めして同時学習する手法 [7]、メンションの同定、単一化、テンプレート要素埋めを同時におこなう手法 [10] などが提案されている。本研究で扱っている情報抽出課題は、抽出問題だけでなく分類問題も含んでいる点、また複数レコード生成が必要である点など、従来の情報抽出の問題設定よりも複雑であると言える。

5 まとめ

本稿では、実世界の情報構造化支援の例題として、災害時における防災情報の効率的な共有のための防災情報データベース化システムについて述べた。本システムは、様々な自治体において低コストで導入できるように、特定のデータベーススキーマに依存しない汎用的なレコード生成、場所情報を置き換えだけで他の自治体でも利用可能とする場所情報の管理、簡単な操作で情報をデータベース化することを可能とするユー

ザーインターフェースなどを特徴とする、汎用的なシステム設計となっている。今後は、音声認識との連携、場所情報の修正インターフェース、レコード生成部分に対して素性選択アルゴリズム等を実装することでシステムを洗練させる。また、本システムを気仙沼市の機能訓練での利用を予定しており、そこでシステムの課題を洗い出し、さらなる改善を進めていく予定である。

謝辞

本研究は、RISTEX 社会技術研究開発センターの研究開発領域「コミュニティがつなぐ安全・安心な都市・地域の創造」一環として、また JST 戦略的創造研究推進事業「さきがけ」から部分的な支援を受けて行われた。本研究で利用した防災情報データベース化のためのテンプレートおよびサンプルデータは、京都大学防災研究所鈴木進吾助教、NTT セキュアプラットフォーム研究所小阪尚子様、小山晃様、前田裕二様、富士常葉大学河本尋子講師、株式会社サイエンスクラフト大橋真武様にご作成頂いた。ここに感謝の意を表する。

参考文献

- [1] Feng, D., Burns, G. and Hovy, E.: Extracting data records from unstructured biomedical full text, *EMNLP-CoNLL 2007* (2007).
- [2] Higashida, M., Sugiyama, M., Takeda, H., Yamamoto, T., Maeda, Y. and Hayashi, H.: Analysis of Information Processing Patterns Appeared at Emergency Operation Center Training, *Outline of Social Safety and Science*, Vol. 30 (2012).
- [3] Kazama, J., Saeger, S. D., Kuroda, K., Murata, M. and Torisawa, K.: A Bayesian Method for Robust Estimation of Distributional Similarities, *Proc. of ACL 2010* (2010).
- [4] Kudo, T. and Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking, *Proc. of CoNLL 2002* (2002).
- [5] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. of EMNLP 2004* (2004).
- [6] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. of ICML-2001* (2001).
- [7] Li, Q., Ji, H. and Huang, L.: Joint Event Extraction via Structured Prediction with Global Features, *Proc. of ACL 2013* (2013).
- [8] McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y. and White, P.: Simple algorithms for complex relation extraction with applications to biomedical IE, *Proc. of ACL 2005* (2005).
- [9] McDonald, R. T., Winters, R. S., Mandel, M., Jin, Y., White, P. S. and Pereira, F.: An Entity Tagger for Recognizing Acquired Genomic Variations in Cancer Literature, *Bioinformatics*, Vol. 20, No. 17 (2004).
- [10] Minkov, E. and Zettlemoyer, L.: Discriminative learning for joint template filling, *Proc. of ACL 2012* (2012).
- [11] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007).
- [12] Okazaki, N. and Tsujii, J.: Simple and Efficient Algorithm for Approximate Dictionary Matching, *Proc. of Coling 2010*, Beijing, China, pp. 851-859 (2010).
- [13] Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy, *Proc. of LREC 2002* (2002).
- [14] Watanabe, Y., Asahara, M. and Matsumoto, Y.: A Structured Model for Joint Learning of Argument Roles and Predicate Senses, *Proc. of ACL 2010* (2010).
- [15] 橋本泰一, 中村俊一: 拡張固有表現タグ付きコーパスの構築 - 白書, 書籍, Yahoo! 知恵袋コアデータ -, 言語処理学会 第 16 回年次大会予稿集, pp. 916-919 (2008).

*9 <http://projects.ldc.upenn.edu/ace/>