テキストから自動獲得した名詞の分類

村脇 有吾　　黒橋 禎夫
京都大学大学院情報学研究科
murawaki@nlp.kuee.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

1 はじめに

自然言語を計算機に解析させるには、人が持つ様々な常識的知識が必要となる。そうした知識は膨大であり、もっとも基本的な単位である形態素でさえ、人手で網羅するのは不可能に近い。そこで我々は、基本語彙は人手で整備し、残りを計算機に自動獲得させるという方針をとる。

人手で整備された形態素には様々な情報を付与されている。残りの語彙について、そうした情報を一気にに自動獲得するのは現実的でない。そこで、ひとまず基本的な品詞のみを付与した状態で、名詞「メル友」、「アキバ」や動詞「ググる」といった形態素をテキストから自動獲得する研究を行ってきた [7]。

この段階で未獲得だか重要な情報に、普通名詞と固有名詞の区別がある。そこで本稿では、自動獲得した名詞の分類というタスクを提案する。分類は大きく 2 軸、すなわち普通名詞と固有名詞の識別と、人、場所、組織といったカテゴリ分類からなる。このタスクにおいて特徴的なのは、普通名詞と固有名詞の識別が、関連研究 [2, 4, 1] が扱う英語と異なり、日本語では自明でないことである。

本稿では、このタスクを分類問題として定式化し、多クラスパーセプトロンによる分類を試みる。分類器の素性として形態素のテキスト中の振る舞いを与え、どのような手がかりが分類に有効かを考察する。

2 名詞の分類

2.1 分類ラベル

本稿が扱うタスクは、名詞に対して、そのテキスト中の振る舞いをもとに、あらかじめ定義されたラベルのいずれかを付与するという分類問題である。最初に、付与すべきラベルの集合を定義する。

分類対象の名詞は、普通名詞と固有名詞に大別できる。このうち、固有名詞の細分類は品詞として与えられている。形態素解析器 JUMAN の場合、固有名詞には、「人名」、「地名」、「組織名」、「固有名詞」のいずれかの品詞細分類が付与される。ここで「固有名詞」は、前三つのいずれにも当てはまらない固有名詞を表す。本タスクでは、この品詞細分類を固有名詞のラベルとして採用する[1]。

普通名詞には品詞による細分類はない。代わりに、JUMAN 6.0[2]は普通名詞に対して、「人」、「組織-団体」、「動物」、「抽象物」など 22 種類からなる意味カテゴリを付与している。本タスクでは、この意味カテゴリを普通名詞のラベルとして利用する。ただし、いくつかの意味カテゴリを統合し、固有名詞に対応する粗粒度のラベルとする。以上をまとめると、表 1 に示す 10 種類のラベルが定義される。

関連研究としては、超語義タグ付け (supersense tagging) と呼ばれるタスクが英語を対象に行われている [2, 4, 1]。超語義とは WordNet に基づく 26 種類の粗粒度のカテゴリである。例えば、chair は、person、artifact および act という 3 種類の超語義に属す。本タスクを超語義タグ付けと比較すると、普通名詞と固有名詞を識別するという点が特徴的である。英語の場合、固有名詞は大文字から始めるという正書法上の特徴から、両者の区別は自明である。表記の上だけでなく形態統語的にも、英語と日本語の名詞の振る舞いは大きく異なる。例えば、英語では冠詞などの限定詞により定性が義務的に示されるが、日本語では示されない [5]。また、日本語には文法範疇としての数も存在しない。こうしたことから、まず普通名詞と固有名詞の識別に有効な手がかりを探さなければならない。

2.2 分類の手がかり

名詞の分類に用いる手がかりは、その形態素のテキスト中の振る舞いである。ただし、「名詞」、「動詞」といった基本的な品詞分類 [7] と異なり、名詞の分類においては、制約として利用できる強力な形態統語的特徴はないと思われる。したがって、いくつかの弱い手がかりの組み合わせにより分類を試みる。

本稿では以下の 5 種類の素性を用いる。

[1] ipadic は「名詞-固有名詞-人名-姓」のようにより詳細な区別を与えるが、下位の区別を無視すれば同様のラベルが設定できる。
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

表 1: 分類ラベル一覧

分類ラベル	普通/固有	出典	例[1]
人名	固有名詞	品詞:名詞-人名	松井, 愛子, ジョージ, (佐祐理), (キョン)
地名		品詞:名詞-地名	京都, 銀座, ドイツ, (アキバ), (ウヴド)
組織名		品詞:名詞-組織名	日銀, トヨタ, NHK, (マツダ), (エリクソン)
固有名詞その他		品詞:名詞-固有名詞	ガンダム, スラブ, 平成, (ジプシー)
人	普通名詞	カテゴリ:人	被告, 先生, スタッフ, (メル友), (ニート)
場所		カテゴリ:場所-∗[2]	職場, 部屋, カフェ, (囲炉裏), (圃場)
組織		カテゴリ:組織-団体	政府, 仲間, チーム, (興信所), (弊所)
動物		カテゴリ:動物, 動物-部位	犬, ムカデ, 顔, (チワワ), (マンタ)
植物		カテゴリ:植物, 植物-部位	花, ハーブ, 葉っぱ, (ケナフ), (クレマチス)
普通名詞その他		その他のカテゴリの普通名詞	主張, 花火, ビジネス, (甚平), (着メロ)

[1] 括弧付きの形態素は分類したい自動獲得の名詞。
[2] カテゴリ:場所-施設、場所-施設部位など。

call X について、「X という Y」、「X といった Y」などのパターン。「X という国」に対して「call:国」。

cf 係り受け関係にある用言と格要素。「X が言う」に対して「cf:言う:動:ガ格」。

demo 連体修飾する指示詞。「この X」に対して「demo:この」。

ncf1 ノ格による係り先の要素。「X の部屋」に対して「ncf1:部屋」。

ncf2 ノ格による係り元の要素。「すべての X」に対して「ncf2:すべて」。

suf 後続接尾辞。「X さん」に対して「suf:さん」。

分類対象の名詞 X は、単形態素、つまり複合的な名詞句の一部となっていない場合のみから抽出する。ただし、suf は「X+ 接尾辞」が名詞句を形成している場合とする。

call、ncf1、ncf2 については、名詞句中の数詞を汎化する。例えば、「二人の X」からは素性「ncf2:⟨ 数量 ⟩ 人」が抽出される。

cf の抽出手法は 2 通り考えられる。一つは用言格フレームの構築 [6] に用いられる曖昧性のない係り受け関係の抽出 (**dpnd**) である。もう一つは、用言格フレームに基づく格解析結果の利用である。両者を比べると、前者の方が精度が高いが、被覆率が小さい。特に「ガ格」の要素は、提題の助詞「ハ」によって隠されることが多いため、被覆率が小さいことが知られている。ncf1、ncf2 の抽出は、名詞格フレームの構築手法 [9] に従う。

2.3 分類器

本タスクは多値分類問題として定式化できる。分類器は、入力たる d 次元の素性列 (事例) $x \in \mathbb{R}^d$ に対して、出力 $y \in Y$ を返す。ここで、Y は分類ラベルの集合である。

本稿では分類器として多クラスパーセプトロン [3] を用いる。実装が単純であり、オンライン学習ゆえに大規模データへの適用が容易だからである。

多クラスパーセプトロンは、各ラベル y について重みベクトル $v_y \in \mathbb{R}^d$ を持つ。分類時には、入力 x と重み v_y の内積が最大となるような y を出力する。

$$\operatorname*{argmax}_y \langle v_y, x \rangle$$

学習はオンラインであり、訓練事例を一つずつ処理する。分類を誤った場合に重みベクトルを以下のように更新する。まず、正解ラベル y_i の重みを増やす。

$$v_{y_i} \leftarrow v_{y_i} + x$$

次に、正解ラベルよりも内積が大きい y の集合

$$E = \{y | \langle v_y, x \rangle > \langle v_{y_i}, x \rangle\}$$

を考えたとき、各 $y \in E$ について、重みを減らす。

$$v_y \leftarrow v_y - \frac{1}{|E|} x$$

2.4 事例

素性列からなる事例 x の与え方を決める。この際問題となるのが多義性である。多義性を持つ形態素は珍しくなく、ある文脈では「人名」、別の文脈では「地名」といった具合に使われる。したがって、大規模テキストから抽出した素性列全体を一つの事例とすると、複数の語義の混在してしまう可能性が高い。反対にテキスト中の個々の形態素を分類しようとしても、得られる素性はほとんどの場合高々1 個となってしまう。

そこで、同一文書中の形態素から抽出される素性列を 1 個の事例とする。同一文書中では大多数が同じ語義で使われることが、固有表現認識タスクについて知られており、この仮定は妥当と思われる。個々の事例が分類されると、その結果の集約により、形態素の静的な語義が得られる。

2.5 訓練データとテストデータ

分類器の学習には訓練データが必要となる。この訓練データをどのように用意するかが問題となる。一般に教師あり学習では、訓練データを人手で整備することが多い。しかし、2.2 節で述べた素性は種類が多く、

小規模なタグ付きコーパスからの正しい学習は期待できない。

そこで、未知語獲得 [7] と同様に、自動解析結果のうち、人手で整備された形態素に関する部分を訓練データとして利用する。解析結果から素性列を抽出し、閾値以上の数の素性が得られたものを正解事例とする。正解ラベルとしては、人手で付与された分類ラベルを用いる。ただし、複数の分類ラベルが付与された曖昧な形態素は訓練事例から取り除く。

テストデータも、訓練データと同様の手順で自動解析結果から作成する。ここで、分類すべき名詞が、テキストから自動獲得 [7] により、形態素解析器にとって既知となっていることに注意する。分かち書きされない日本語では、形態素解析における未知語は解析を誤りやすい。形態素解析を誤ると、2.2 節で述べたような係り受けに基づく素性が正しく得られない。これに対し、形態素解析において既知であれば、形態素解析で誤る場合が減り、係り受け解析結果を利用できるようになる。

3 実験

3.1 データ

対象テキストは、ウェブコーパスの一部約 2.5 千万ページとする。テキストを形態素解析器 JUMAN、および構文解析器 KNP により解析する。ここで、JUMAN は、語彙の自動獲得 [7] により拡張された形態素辞書を用いる。自動獲得された形態素は 12,684 個、うち分類対象となる名詞は 12,390 個である。

解析結果から、まず 2.2 節で述べた素性を抽出する。頻度 300 以上の素性 126,643 個を分類に用いる。次に、人手で登録された形態素から訓練データを、自動獲得された名詞からテストデータを生成する。事例に必要な最低素性数を 10 とし、事例作成単位たる文書をウェブページとする。ただし、十分な数の素性が得られない場合、同一ドメインの別ページをあわせて擬似的に一つの文書とみなす。

3.2 自動獲得名詞の分類

自動獲得された名詞の分類を評価する。複数事例の集約結果は判定が難しいため、個々の事例の分類結果を判定する。素性としては、2.2 節で述べたすべての種類の素性を用いる。また、cf の素性抽出には格解析結果を用いる。訓練データとして 7,039,826 個の事例が得られた。テストデータは 302,303 個の事例が得られ、うち 500 個の事例に人手で正解を付与した。

表 2: 分類結果の混同行列

	人名	地名	組織名	固有他	人	場所	組織	動物	植物	普通他
人名	36				4			1		
地名	2	3	5			1				1
組織名	2									1
固有他			1	2		1		1		
人	53		1	2	32		1	4		12
場所	3	2	1	1		5		1		5
組織			3		1					2
動物	2				5			26		5
植物	2							1	4	
普通他	10	2	7	1	6	5		12	2	220

表 3: 集約された分類結果の例

形態素	事例数	ラベル所属割合
加持	98	人名:0.776, 人:0.133, 普通他:0.071
ラスベガス	121	地名:0.752, 場所:0.116, 普通他:0.050
グーグル	128	組織名:0.375, 普通他:0.300, 人:0.117
かみさん	153	人:0.869, 普通他:0.065, 人名:0.039
チワワ	159	動物:0.761, 普通他:0.082, 人:0.057
メルマガ	1,743	普通他:0.950, 場所:0.024, 組織:0.008

精度は 65.6%であった。表 2 に混同行列を示す。もっとも多い誤りは、「人名」を「人」と誤判定するもので、53 例あった。自動獲得形態素の集約された分類結果の例を表 3 に示す。集約するとかねがね妥当だが、「人名」を「人」と誤判定する部分が依然目立つ。

3.3 素性の調査

素性の効果を確認するために、各種類の素性を抜いたり、cf の抽出手法を曖昧性のない係り受け関係 (dpnd) に代えて実験する。ここで、素性の種類を変えると作成される事例も変わるため、統一的な評価データが作りにくいという問題がある。そこで、人手で整備された形態素の分類によって評価する。すなわち、訓練データを分割し、新たな訓練データとテストデータとする。3.2 節の訓練データからは、人手で整備された形態素が 20,931 個得られたが、このうち頻度上位 2,000 形態素を除き、残りから 2,000 形態素を無作為に抽出した。これらの形態素の事例を分類ラベルを隠してテストデータとし、残りを訓練データとした。

結果を表 4 に示す。自動獲得名詞と人手で整備された形態素の分類で精度が大幅に異なる原因は、誤りやすい「人名」が事例に占める割合に求められる。前者に占める「人名」は 22.0% (すべての素性を用いた場合) に対し、後者では 8.2%にすぎない。

訓練データの事例数の減少率から、cf が素性の多くを占めていることがわかる。cf の素性抽出を dpnd とすると、事例数が 4 分の 1 近くに減少するが、逆に精度がやや上がっている。また、ncf1 よりも ncf2 の方が分類に効果的と推測できる。自分が修飾する要素よりも、自分を修飾する要素の方が、自分の性格を表すことを意味し、直感に合致する。

表 4: 素性の調査

素性	事例数	減少率	精度	
-call	6,969,353	0.05%	74.4%	(49,798 / 66,919)
-cf	30,7857	95.6%	73.5%	(1,768 / 2,407)
-demo	6,603,441	5.30%	75.0%	(47,857 / 63,837)
-ncf1	6,251,402	10.3%	74.4%	(42,360 / 56,929)
-ncf2	6,119,787	12.2%	72.9%	(42,191 / 57,856)
-suf	6,587,966	5.52%	72.6%	(44,489 / 61,307)
dpnd	1,672,179	76.0%	76.7%	(9,438 / 12,301)
すべて	6,972,805		74.7%	(50,035 / 67,021)

3.4 議論

分類精度を低下させる原因として、訓練データ中のノイズが挙げられる。提案手法では訓練データから明示的な曖昧性を排除しているが、明示されない曖昧性が残っている。例えば、人手で整備された形態素「愛」には「普通名詞その他」のみがラベルとして付与されるが、実際には「人名」として用いられる事例が少なくない。従来研究 [2, 4, 1] は英語を対象としており、そもそも普通名詞と固有名詞の語義が混在しない。加えて、一般の辞書は見出し語に対して 1 個以上の語義を記述するのに対し、本タスクでは、普通名詞と固有名詞が別々に登録された形態素辞書から、形態素を見出し語ごとに集約することによって多義性を発見している。したがって、語義の登録漏れが起きやすい。

どの手がかりが分類に有効かは、各素性の重みをラベルごとに比較すれば分析できる。例えば、ラベル間で「人」の重みが最大の素性のうち、「人名」の重みとの差が大きいものには、「ncf2:多く」、「ncf2:〈数量〉人」、「ncf2:全て」、「suf:向け」などがある。

「ncf2:〈数量〉人」のように類別詞を含む数量詞は、言語哲学や言語学において、日本語における可算名詞と不可算名詞の区別に関係して注目されてきた [8, 5]。本タスクにおいては、類別詞は、普通名詞の固有名詞からの区別、およびカテゴリ分類への有効性が確認できる。例えば、数量詞「人」からは普通名詞の人、「軒」からは普通名詞の場所と推測できる。ただし、若干のノイズが含まれる。助詞「の」の用法は広く、「二人の問題」のように数量詞として用いられない場合がある。「X が二人 (で)」のように遊離した数量詞を利用すれば、この問題は回避できるかもしれない。

「ncf2:多く」や「cf:増える:動:ガ格」は、カテゴリの分類には役立たないが、普通名詞の固有名詞からの識別に有効である。これらの素性は、固有名詞に対して用いると、多くの場合に意味的整合性が取れない。ただし、「固有名詞その他」に含まれる「スラブ」のような人の集団については、「X が増える」といった表現は自然である。

「この」、「そんな」などの指示詞は、普通名詞と固有名詞の区別、カテゴリ分類のいずれにも役に立たない。日本語では、「あの山田」のような指示詞と固有名詞の組み合わせは、非限定的用法で自然に用いられる。ただし、「どの」、「どんな」などの疑問の系列は、普通名詞の固有名詞からの識別に有効である。

「人名」を「人」から識別する素性を調べると、過学習によると見られるノイズが多い。「cf:ゴロ:動:ガ格」や「cf:失点:動:ガ格」などは、訓練データ中では固有名詞に対して使われる場合が多いものの、普通名詞に対して用いても意味的に問題がない。ノイズでないと見られる素性としては、「call:男」、「suf:両氏」などがあったが、全体として固有名詞を普通名詞から区別するために有効な手がかりが乏しい。

4 おわりに

本稿では、自動獲得された名詞をテキスト中の振る舞いをもと分類するというタスクと、その解法を提案した。分類精度はまだまだ低く、改善の余地が残っている。特に、固有名詞を普通名詞から識別するための手がかりは何かという問題は興味深い。

参考文献

[1] Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP 2006*, pp. 594–602, 2006.

[2] Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in WordNet. In *Proc. of EMNLP 2003*, pp. 168–175, 2003.

[3] Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, Vol. 3, pp. 951–991, 2003.

[4] James R. Curran. Supersense tagging of unknown nouns using semantic similarity. In *Proc. of ACL 2005*, pp. 26–33, 2005.

[5] David Gil. Definiteness, NP configurationality and the count-mass distinction. In Eric J. Reuland and Alice G. B. ter Meulen, editors, *The Representation of (In)definiteness*, pp. 254–269. MIT Press, 1987.

[6] Daisuke Kawahara and Sadao Kurohashi. Japanese case frame construction by coupling the verb and its closest case component. In *Proc. of HLT 2001*, pp. 204–210, 2001.

[7] Yugo Murawaki and Sadao Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proc. of EMNLP 2008*, pp. 429–437, 2008.

[8] Willard Van Quine. *Ontological Relativity and Other Essays*. Columbia University Press, 1969.

[9] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *Proc. of COLING 2004*, pp. 1201–1207, 2004.