絵本のテキストを対象とした形態素解析

藤田 早苗　　平 博順　　小林 哲生
NTT コミュニケーション科学基礎研究所
{fujita.sanae, taira.hironori, kobayashi.tessei}@lab.ntt.co.jp

1　はじめに

これまで、主に新聞などのテキストを対象とした解析では、形態素解析器を始めとして高い解析精度が達成されている。しかし近年、解析対象は Web データなど多様化が進んでおり、これらのテキストに対しては既存の解析モデルで、必ずしも高い解析精度を得られるわけではない。

本稿では、そうしたテキストの一つである絵本を対象とした形態素解析への取り組みについて紹介する。絵本は幼児の言語発達を支える重要なインプットの一つであり [3]、高い精度で解析できれば、発達心理学における研究や教育支援などへの貢献が期待できる。

絵本の多くは子供向けに書かれており、わかりやすい文章になっていると考えられる。それにも関わらず、一般的な形態素解析モデルでは、必ずしもうまく解析できない。そこで本稿では、新聞などのテキストと絵本のテキストを比較、違いを調査する。また、その違いに基づき、既存の大人向けテキストからなる学習データを変換・利用し、絵本のテキストの形態素解析精度の向上、および、精度低下の原因調査を試みる。

2　絵本データベース

本稿では構築中の絵本データベースを解析対象とする [7]。絵本データベースは、発達心理学における研究や、子供の興味や発達に応じた絵本リコメンデーションを目的として構築されている。

含まれる絵本は、2010 年度の紀伊国屋書店グループの売上冊数が上位のファーストブック (以下、FIRST) と絵本 (以下、EHON) 計 1,000 冊以上[1]、および、福音館書店の月刊誌 (以下、KODOMO)190 冊である。これらの選定理由は、前者は多くの子供に読まれていると考えられること、後者は対象年齢が比較的はっきりしていることである[2]。本稿では、これらをまとめて絵本と呼ぶこととする。

本稿では、絵本データベースに含まれる情報のうち、本文のテキストを解析対象とする。本文のテキストは、人手で入力されている[3]が、元のページのレイアウトに忠実に入力されている。そのため、文や文節の途中での改行なども、忠実に再現されている (例 (1))。なお、2012 年 12 月 25 日時点で本文のテキスト入力済みの絵本 958 冊のサイズは表 1 の通りである。

(1)　「おや、かばくん。ぞうくんを　のせて
なにしてるの」
(なかのひろたか「ぞうくんのあめふりさんぽ」p.15 (2004, 福音館書店))

	計	平均/冊	最大/冊	最小/冊
ページ数	29,215	30.5	332	9
行数	102,707	107.2	1,844	10
文字数	1,486,102	1,551.3	23,308	33

表 1: 絵本データベースのサイズ

3　絵本と他のコーパスの比較

絵本のテキストの特徴を調べるため、絵本と一般的なコーパスにおける文字種の割合を比較する。表 2 に、データ入力済みの絵本 (表 1) における文字種と、現代日本語書き言葉均衡コーパス[4](以下、BCCWJ) 、京都大学テキストコーパス[5](以下、京大コーパス) 、および、基本語データベース Lexeed[2] の定義文、例文に出現する文字種の数と割合を示す。

表 2 から、他のコーパスに比べ、絵本の場合、ひらがなと空白が占める割合が圧倒的に高いことがわかる。また逆に、漢字が占める割合は非常に低い。表 2

[1]絵本とファーストブックの分類は紀伊国屋書店による。

[2]対象年齢は 0・1・2 歳向け、年少向け (3 歳児)、年中向け (4 歳児)、年長向け (5 歳児) とわかれている。

[3]当初、既存 OCR による自動的な文字認識を試したが、絵と文字部分の判別が難しく、高精度な自動認識は困難だった。

[4]http://www.ninjal.ac.jp/kotonoha/

[5]http://nlp.ist.i.kyoto-u.ac.jp/index.php

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

文字種	絵本 No.	絵本 %	BCCWJ No.	BCCWJ %	京大コーパス No.	京大コーパス %	Lexeed No.	Lexeed %
ひらがな	1,118,635	**75.3**	91,872,255	49.2	714,357	42.6	929,088	48.5
カタカナ	90,168	6.1	14,758,557	7.9	118,342	7.1	92,562	4.8
漢字	23,446	**1.6**	56,457,392	30.3	716,972	42.8	731,820	38.2
アルファベット	2,257	0.2	2,784,843	1.5	4,803	0.3	554	0.0
数字	4,271	0.3	3,112,390	1.7	6,441	0.4	7,349	0.4
空白	114,832	**7.7**	2,432,456	1.3	0	0.0	504	0.0
記号	52,730	3.5	6,218,324	3.3	24,969	1.5	2,360	0.1
句読点	79,763	5.4	8,959,297	4.8	90,729	5.4	151,276	7.9
TOTAL	1,486,102	100.0	186,595,514	100	1,676,613	100	1,915,513	100
(参考) 平均文字数/文	14.5		33.2		43.6		16.0	

表 2: 文字種の割合: 絵本と他のコーパスの比較

には、参考として、一文に含まれる平均文字数も記載した。但し、絵本の場合は、一行に含まれる平均文字数を記載しており、必ずしも文単位ではない。

4 形態素解析

本稿では、京都テキスト解析ツールキット KyTea[6][5, 6] の学習機能を利用し、絵本用の形態素解析モデルを構築する。ここで、KyTea を選択したのは、再学習が容易なこと、複数の言語資源を利用しやすいことが主な理由である。なお、既存言語資源との整合性を考慮し、品詞体系は IPA 品詞体系に準拠した。

学習データ　学習には、コーパス檜 [1] を用いた。檜には、Lexeed の定義文、例文、京大コーパスの全文[7]が含まれている。3 章で述べたように、絵本の場合、一般的なコーパスに比べ、ひらがなや空白が非常に多く、逆に漢字が少ない。そこで、檜をこれらの特徴に併せて変換したデータも学習データとして用意する。例えば、文 (2) は、Lexeed での見出し語「きしめん」に付与された例文である。ここで、”,” は形態素区切りを示している。これに、まず句読点の直後を除く文節毎に空白を挿入する (文 (3))。また、すべての漢字をひらがなのよみに変換する (文 (4))。更に、ひらがなに変換し、かつ、句読点の直後を除く文節毎に空白を挿入し (文 (5))、学習データを作成した。

(2)　寄せ鍋, に, きしめん, を, 入れる,。

(3)　寄せ鍋, に,　, きしめん, を,　, 入れる,。

(4)　よせなべ, に, きしめん, を, いれる,。

(5)　よせなべ, に,　, きしめん, を,　, いれる,。

学習データを変換した場合の効果を検証するため、これらの学習データの組み合わせを変えて利用した場合の精度評価を行なう。

辞書　辞書として、NAIST Japanese Dictionary[8] (以下、NAISTJ) と、Lexeed、および、日本語語彙大系 [4] の固有名詞、および、動植物名[9]を利用した。但し、Lexeed と日本語語彙大系は、本来 IPA 品詞体系ではないが、自動的に IPA 品詞に変換した。

また、NAISTJ と Lexeed の場合、漢字やカタカナのエントリは、ひらがなに変換したエントリも作成・追加した。日本語語彙大系も、同様に、カタカナ、ひらがなに変換したエントリも作成・追加した。

例えば、日本語語彙大系から得られる「伊予柑」の場合、もともとの見出し語から得られるエントリは (6) となるが、ひらがなのエントリ (7) とカタカナのエントリ (8) も追加した。最終的に利用した辞書サイズは、表 3 の通りである。

(6)　伊予柑/名詞-一般/イヨカン/伊予柑

(7)　いよかん/名詞-一般/イヨカン/伊予柑

(8)　イヨカン/名詞-一般/イヨカン/伊予柑

知識源	NAISTJ	Lexeed	日本語語彙大系 固有名詞	日本語語彙大系 動植物名
エントリ数	810,907	73,568	624,502	14,282

表 3: 辞書サイズ: ひらがなやカタカナに展開済み

[6]http://www.phontron.com/kytea/index-ja.html, ver. 0.3.2 を利用

[7]但し、IPA 品詞体系で解析しなおしてある。

[8]http://sourceforge.jp/projects/naist-jdic/

[9]具体的には、日本語語彙大系の日本語辞書のうち、＜**543:生物**＞配下の意味クラスが付与されている語を追加した。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

5 評価用データ

FIRST, EHON, KODOMO の各クラスから、評価用の絵本をランダムサンプリングした。さらに各絵本から、文のあるページを 1 ページずつランダムサンプリングし、人手で形態素情報をアノテーションした。評価用データのサイズを表 4 に示す。

	FIRST	EHON	KODOMO	TOTAL
ページ (本) 数	100	100	50	250
行数	287	437	176	900
文字数	2,699	6,630	2,770	12,099
形態素数	1,306	3,722	1,553	6,581

表 4: 評価データサイズ

また、絵本に特徴的な、ひらがなや空白の影響を調査するため、評価データをいくつかのルールに沿って変換したものも作成した。例えば、文 (9) は、絵本にもともと出てくる文である。こうした評価データの空白を削除したもの (文 (10))、ひらがなをできるだけ漢字に変換したもの (文 (11))、漢字に変換し、かつ、空白を削除したもの (文 (12)) を作成した。

(9) めには、いちごの　あかい　みを　いれました。
(舟崎靖子 「もりのおかしやさん」 p.11 (1979, 偕成社))

(10) めには、いちごのあかいみをいれました。

(11) 目には、苺の　赤い　実を　入れました。

(12) 目には、苺の赤い実を入れました。

6 評価結果と分析

表 5 に、各条件でモデルを構築した場合の評価データに対する精度を示す。

表 5 の左端「元データ (9)」の列が、絵本のもともとの文章に対する精度を示している。また、表 5 の網かけ部分は、空白追加やひらがな変換などを行なわない、元の学習データを用いた結果であり、一般的な形態素解析モデルと同じような学習条件に相当するだろう。このままだと、精度は 66.8 %と非常に低いが、空白を追加したり、ひらがなに変換した学習データを利用することで、86.3 %まで精度を向上できた。つまり、新聞データなどの大人向けの文章を学習データに利用する場合でも、絵本での出現傾向にあわせて変換することで、相当の精度向上が出来ることがわかった。

ここで、空白を追加した学習データだけを利用する場合 [B] より、空白を追加しない学習データも利用する [C] の方が精度が高かった。これは、すべての絵本で全文節ごとに空白が入るわけではないので、両方を学習に利用した方が良かったのだろうと考えられる。そのため、今後は、この最も良かったモデルをベースに、更に改良を加えることを検討したい。また、絵本によって、空白や漢字の含有率は非常に異なるため、これらの含有率によってモデルを変更することも考えられる。

なお、よくある間違いは、擬人化（例えば、「ぞうさん」「くろくん」など）や、ひらがなの固有名詞 (例えば、「ぐり」「ぐら」など)、口語体 (13) や方言 (14) が出てくる部分が多かった。いずれも檜コーパスにはほとんど出現しないため、対応する学習データの増強や方言用モデル構築が必要だろう。

(13) うーん、しろかぶくんは、まだか のう？
(やなせたかし「しろかぶくんとアンパンマン」 p.12 (2011, フレーベル館))

(14) 番長は、いっちょん いうこと ば 聞き より ません。
(よしながこうたく 「給食番長」 p.9 (2007, 長崎出版))

6.1 空白・ひらがなの影響

絵本のテキストは従来の形態素解析モデルでは、うまく解析することが難しい。その原因として、絵本に特徴的な空白とひらがなのどちらがより影響するのか調査した。

まず、空白を削除することでその影響を調査した。空白を削除した場合、訓練データに空白追加のない [A] では精度が向上する。しかし、空白の学習は比較的容易であり、[B] や [C] のように空白を追加した学習データを用いると、精度はすぐに向上する。参考にあげた MeCab でも、空白は区切りの判別のための手がかりとして有効に利用されているようであり、空白を削除するとむしろ精度は低下する。

特に、(15) のように、擬音語や擬態語が連なる場合、空白を削除すると、解析が非常に困難になっており、空白の有無が形態素の判別に有効な手がかりであることがわかる。

(15) 「こちょ　こちょ　こちょ　こちょ
(豊田一彦「こちょこちょももんちゃん」 p.24 (2010, 童心社))
COR:「,こちょ，　,こちょ，　,こちょ，　,こちょ
RES:「,こ, ちょこ, ちょこちょこ, ちょ
(但し、COR: は正解、RES: は空白を削除した場合の結果)

次に、ひらがなが多いことによる影響を調査する。評価データ中のひらがなを漢字に変換した場合、[A] 以外では、比較的高い精度を得ている。特に、最も高い精度は 88.0 %あり、漢字は大きな手がかりとなっていることがわかる。つまり、一般的なテキストとの大きな違いのうち、ひらがなによる曖昧性の増大が解析精度の低下の主な要因だといえる。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

学習データ	ひらがなのままの評価データ 元データ (9) No.	%	空白削除 (10) No.	%	漢字に変換した評価データ 漢字変換 (11) No.	%	空白削除 (12) No.	%
[A] 空白追加なしの学習データを使った場合								
漢字のまま (2)	4,397	66.8	4,296	76.9	4,813	73.1	4,758	85.2
ひらがな (4)	4,675	71.0	4,555	81.6	4,143	63.0	4,170	74.7
両方利用 (2)+(4)	4,728	71.8	**4,618**	**82.7**	4,786	72.7	4,708	84.3
[B] 文節で空白を入れた学習データだけを使った場合								
漢字のまま (3)	5,313	80.7	3,941	70.6	5,730	87.1	4,475	80.2
ひらがな (5)	5,407	82.2	3,768	67.5	5,016	76.2	3,332	59.7
両方利用 (3)+(5)	5,492	83.5	3,917	70.2	5,608	85.2	4,144	74.2
[C] 空白追加なしの学習データと、文節で空白を入れた学習データを使った場合								
漢字のまま (2) + (3)	5,406	82.1	4,295	76.9	**5,794**	**88.0**	**4,762**	**85.3**
ひらがな (4) + (5)	5,629	85.5	4,516	80.9	5,282	80.3	4,155	74.4
両方利用 (2)〜(5)	**5,682**	**86.3**	4,580	82.0	5,742	87.3	4,694	84.1
(参考) MeCab[10]	5,263	80.0	4,077	73.0	5,753	87.4	4,732	84.8

但し、(2) から (5) は、対応する学習データの例の番号、(9) から (12) は、対応する評価データの例の番号を示している。また、[A]-[C] は参照用に付与した記号である。

表 5: 評価結果: 区切り+品詞が一致した数と割合

なお、元データのままだと解析に失敗するが、漢字に変換すると正解する例には、(16) などがあった。

(16) みずを　のみにきた　うしさんに
(たちもとみちこ「おほしさま」p.10 (2006, 教育画劇))
COR: みず, を,　, のみ, に, き, た,　, うし, さん, に,
RES: みず, を,　, のみ, に,きた,　,うしさん, に
RES2: 水, を,　, 飲み, に,来, た,　,牛, さん, に
(但し、COR: は正解、RES: は結果、RES2: は漢字に変換した場合の結果)

7　まとめと今後の課題

本稿では、絵本特有の文章を形態素解析するための取り組みについて紹介した。

絵本の文章の形態素解析は、新聞データなどの大人向けの文章を学習データに用いた一般的な形態素解析モデルでは難しい。そこで、本稿では、まず絵本と一般的な大人向けの文章を比較し、絵本では空白やひらがなの含有率が非常に高いことを示した。そして、新聞などの大人向けの文章からなる学習データに、空白を挿入したり、ひらがなに変換して用いることで、精度を 86.3 %まで向上できることを示した。しかし、新聞などに対する形態素解析精度は、一般に 96 %を越えると言われており、比較すると精度はまだまだ低い。これは、ひらがなの固有名詞や擬人化、口語文など、既存の学習データの改変だけでは難しい部分の影響だと思われる。これらの問題は、絵本以外の文章でも共通する問題であり、これらへの対応を進めれば、形態素解析モデルのロバスト化にもつながると期待している。

今後は更に、絵本自体のデータを学習データに追加し、段階的に解析精度を向上させながら、解析を進めたい。特に、絵本には繰り返しが多いという特徴を利用し、最初の数ページを先にアノテーションし、学習データに追加することで効率的な精度向上を目指す予定である。

また、形態素解析の次の段階として、絵本に出現する語の意味クラスの推定・分析、「オノマトペ」や「社会語」といった赤ちゃんへのインプット分析に利用されるラベルの付与などを行なっていく予定である。

謝辞　KyTea の利用に際して大変ご協力をいただいた京都大学森信介先生、奈良先端大 Graham Neubig 先生に感謝する。

参考文献

[1] Francis Bond, Sanae Fujita, and Takaaki Tanaka. The Hinoki Syntactic and Semantic Treebank of Japanese. *Language Resources and Evaluation (Special issue on Asian language technology)*, 2007.

[2] 笠原要, 佐藤浩史, Francis Bond, 田中貴秋, 藤田早苗, 金杉友子, 天野昭成. 「基本語意味データベース:lexeed」の構築. 情報処理学会 自然言語処理研究会 (2004-NLC-159), pp. 75–82, 2004.

[3] Helen Raikes, Barbara Alexander Pan, Gayle Luze, Catherine S. Tamis-LeMonda, Jeanne Brooks-Gunn, Jill Constantine, Louisa Banks Tarullo, H. Abigail Raikes, and Eileen T. Rodriguez. Mother-child bookreading in low-income families: Correlates and outcomes during the first three years of life. *Child Development*, Vol. 77, No. 4, pp. 924–953, 2006.

[4] 池原悟, 宮崎雅弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.

[5] 森信介. 自然言語処理における分野適応. 人工知能学会誌, Vol. 27, No. 4, 2012.

[6] 森信介, 中田陽介, Graham Neubig, 河原達也. 点予測による形態素解析. 自然言語処理, Vol. 18, No. 4, pp. 367–381, 2011.

[7] 平博順, 藤田早苗, 小林哲生. 絵本テキストにおける高頻度語彙の分析. 情報処理学会関西支部 支部大会, 2012.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.