Vol.2014-NL-216 No.12
Vol.2014-SLP-101 No.12
2014/5/22

言語資源の追加：辞書かコーパスか

森 信介[1,a)]　ニュービッグ グラム[2,b)]

概要：本論文では、言語資源の追加戦略が日本語の形態素解析にどの程度の精度向上をもたらすかについて論じる。戦略は 2 つあり、1 つは辞書として未知語を追加する方法である。もう 1 つは、コーパスとして学習データに追加する方法である。実験の結果、コーパスとして学習データに追加する方法の方がよいことが分かった。レシピテキストに対しての適応実験の結果、未知語に関しては、部分的アノテーションコーパスとして、3 回程度の出現事例を文脈とともに追加すると辞書追加よりも明らかに改善することが分かった。この知見に基づき、発明開示書に対して実際にアノテーションを行い、単語分割精度の向上を確認した。

1. はじめに

機械学習に基づく言語処理が主流になり、言語資源の重要性が増している。また、単語分割や品詞推定などでは、一般分野のテキストに対する解析精度は非常に高い。これらの課題では、学習に用いる言語資源が少ない分野のテキストに対する解析精度が問題である。

本論文では、単語分割や品詞推定の課題に対して、辞書として言語資源を追加する場合とタグ付きコーパス (部分的アノテーションを含む) として言語資源を追加する場合について実験的に比較を行う

なお、本稿は [1] の概略である。詳細については当該論文を参照されたい。

2. 形態素解析

本論文で言語資源の追加による精度向上を試みる課題は、日本語の形態素解析である。この問題には大きく分けて 2 つの解き方が提案されている。1 つは、単語分割と品詞付与を同時に行う系列ラベリングとしての定式化である [4], [5], [6], [7], [8]。もう 1 つは、単語分割と品詞推定を個別に解く方法である [2], [3]。

本論文では、単語分割と品詞付与を CRF を用いて同時に解く手法 (MeCab)[6] と単語分割と品詞推定を点予測を用いて個別に解く手法 (KyTea) [3] について、言語資源の追加による分野適応の戦略について実験的に考察する。

[1] 京都大学学術情報メディアセンター
京都府京都市左京区吉田本町
[2] 奈良先端科学技術大学院大学情報科学研究科
奈良県生駒市高山町 8916-5
[a)] forest@i.kyoto-u.ac.jp
[b)] neubig@is.naist.jp

3. 評価

学習コーパスへの言語資源の追加方法の違いを観測するために、以下の実験を行った。

3.1 実験の設定

課題は形態素解析の分野適応である。実験に用いたコーパスは、BCCWJ のコアデータ [11] である。対象分野を Yahoo!知恵袋とし、一般分野の他の 5 分野とした (表 1 参照)。実験に用いた形態素解析は以下の 2 つである[*1]。

(1) MeCab: CRF に基づくモデル [6]

(2) KyTea: 点予測に基づく 2 段階モデル [3]

以下の戦略を試みた。

なし: 一般分野のコーパスと辞書のみを使用

辞書追加 (再学習なし): Yahoo!知恵袋の学習コーパスに出現する単語を辞書として追加 (MeCab のみ)

辞書追加 (再学習あり): Yahoo!知恵袋の学習コーパスに出現する単語を辞書として追加し再学習

コーパス追加: Yahoo!知恵袋の学習コーパスを学習データに追加し再学習

評価基準は、再現率と適合率の調和平均である F 値とした。

3.2 結果と議論

実験結果は表 2 の通りである。この表から、辞書として追加することにより、MeCab(再学習なし) では 1.35%、KyTea では 1.21%精度が向上することがわかる[*2]。さら

[*1] 緻密なパラメータ調整は行っていないので、さらなる精度向上の余地がある。
[*2] 表 2 からわかるように、CRF のパラメータを再学習すると精度

コーパス		
分野	単語数	
一般	784k	
一般+Yahoo!知恵袋	898k	
Yahoo!知恵袋 (テスト)	13.0k	
辞書		
分野	単語数	カバー率 (単語/品詞)
一般	29.7k	96.3%
一般+Yahoo!知恵袋	32.5k	97.9%

表 1 言語資源の諸元

言語資源追加戦略	MeCab	KyTea
なし	95.20%	95.54%
辞書追加 (再学習なし)	96.59%	-
辞書追加 (再学習あり)	96.55%	96.75%
コーパス追加	96.85%	97.15%

表 2 単語分割精度 (F 値)

	文数	用語数	単語数	文字数
学習	1,760	13,197	33,088	50,002
テスト	724	–	13,147	19,975

表 3 レシピコーパスの諸元

に、コーパスとして追加することにより、さらになる精度向上が実現できることがわかる (MeCab の場合は 0.30%、KyTea の場合は 0.40%)。つまり、コーパスの追加による精度向上幅の 75~80%は辞書として追加することにより実現できているが、残りの 20~25%は文脈情報を含むコーパスとして追加することによってのみ実現されるといえる。

4. 実際の分野適応

前節までの分野適応では、フルアノテーションを想定している。本節では、より現実的な部分的アノテーション [10] による言語資源の追加について述べる。MeCab には部分的アノテーションからの学習機能がないので、KyTea を用いた結果のみを報告する。また、日本語の形態素解析の曖昧性の主たる要因である単語分割を課題とする。

4.1 レシピテキスト

最初の適応分野はレシピである。実験にはレシピフローグラフコーパス (r-FG コーパス) [12] を用いた。r-FG コーパスでは、各文に含まれる重要語の両端と種類の情報が付与されている。また、重要語の内部のみ単語分割されている。(図 1 参照)。

4.1.1 実験の設定

表 3 は r-FG コーパスの諸元である。分野適応をしない「**なし**」に加えて、以下の 2 つの分野適応手法を試みた。

辞書 学習データを辞書として用いる

部分的単語分割コーパス 頻度の上限を n として学習データを部分的単語分割コーパスとして用いる

が低下した。

分野適応	頻度		単語数	F 値	
戦略	上限 (n)	平均		BCCWJ	レシピ
なし	–	–	0	98.87%	94.35%
辞書	–	–	1,999	98.90%	94.54%
部分的単語分割コーパス	1	1.00	1,999	98.89%	95.56%
	2	1.60	3,191	98.89%	95.81%
	3	2.02	4,046	98.89%	95.94%
	4	2.36	4,727	98.89%	96.01%
	8	3.26	6,523	98.89%	96.07%
	16	4.26	8,512	98.89%	96.14%
	32	5.10	10,203	98.89%	96.21%
	64	5.77	11,542	98.89%	96.28%
	∞	6.60	13,197	98.89%	96.29%

表 4 各分野適応戦略による単語分割精度

	文数	単語数	文字数
テスト	500	20,658	32,139

表 5 発明開示書コーパスの諸元

4.1.2 結果と議論

単語分割の精度を表 4 に掲げる。この結果から、辞書として重要語を学習データに追加することにより精度向上が図れることがわかる。これは、表 2 の結果と一貫性がある。部分的単語分割として各 1 回の出現となる部分的単語分割された文の追加は、辞書追加と同等である。部分的単語分割コーパスに含まれる頻度 (n) を増加するに従って精度は向上するが、その程度は逓減していく。したがって、アノテーションコストを考慮すれば、新語を頻度の降順にアノテーションしていくことが良い戦略と考えられる。

4.2 発明開示書

最後に、発明開示書に対する分野適応の結果について報告する。

4.2.1 実験の設定

上述の知見に基づき、以下の分野適応戦略を実施した。

(1) 分布分析 [13] に基づき、生コーパスから未知語候補を抽出する。

(2) 未知語候補を期待頻度の降順に整列し、各未知語候補に対して 3 回の出現箇所をアノテーションする。

この戦略では、作業の初期は頻度が 3 以上の未知語候補へのアノテーションとなるので、表 4 の頻度 4 と 8 の間 (平均が 3) に対応する。

各 1 時間の作業の後に、それまでにアノテーションされたコーパスを追加して単語分割器を構築し、表 5 のテストデータに対する単語分割精度を測った。

4.2.2 結果と議論

図 2 は、実験の結果得られた学習曲線である。部分的単語分割コーパスの追加に応じて精度がほぼ単調に増加している。12 時間の作業の結果、20%の誤り削減に成功し、絶対的な精度は一般分野と同等となった [3]。この結果得られ

各 /ホット ドッグ/F に /チリ/F 、 /チーズ/F 、 /オニオン/F を /ふりかけ/Ac る
/ホット ドッグ/F を /アルミ ホイル/F で /覆/Ac う

図 1 r-FG コーパスの文の例

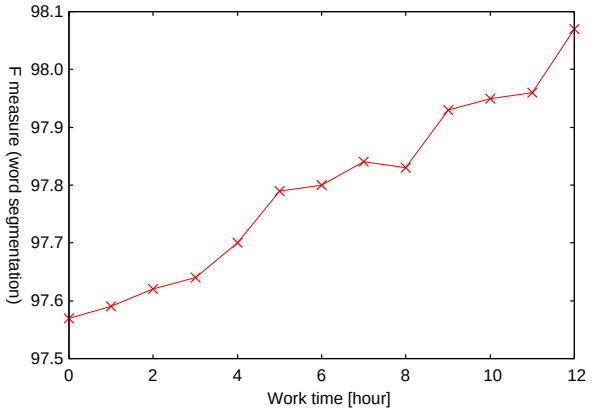

図 2 学習曲線

た単語分割器のモデルは、発明開示書に対する様々な言語処理に貢献することができる。さらに、図 2 から精度は向上し続けていることがわかるので、単に作業を継続するだけでさらなる精度向上が図れることがわかる。

形態素解析 (単語分割と品詞推定) は公開されているツールとモデルで行うことが一般的であることを考えると、本論文の知見を活かすためには、独自に準備したコーパスを配布モデルに追加して再学習ができる設計が重要である。

5. おわりに

本論文では、日本語の形態素解析の精度向上を目的とした言語資源追加の戦略について実験的に比較した。実験の結果、単に辞書として追加するのではなく、数回の出現に対して文脈も含めた部分的アノテーションコーパスとして学習コーパスに追加することが効果的であることが分かった。

6. 謝辞

本研究は JSPS 科研費 23500177,24240030,26280084 と日本電信電話株式会社との共同研究契約 (2013/05/23) の助成を受けたものである。ここに謝意を表する。

参考文献

[1] Mori, S. and Neubig, G.: Language Resource Addition: Dictionary or Corpus?, *Proceedings of the Nineth International Conference on Language Resources and Evaluation* (2014).

[2] Ng, H. T. and Low, J. K.: Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2004).

[3] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011).

[4] Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 201–207 (1994).

[5] Mori, S. and Kurata, G.: Class-Based Variable Memory Length Markov Model, *Proceedings of the InterSpeech2005*, pp. 13–16 (2005).

[6] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 230–237 (2004).

[7] Nakagawa, T.: Chinese and Japanese word segmentation using word-level and character-level information, *Proceedings of the 20th International Conference on Computational Linguistics* (2004).

[8] Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K. and Isahara, H.: An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging, *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics* (2009).

[9] Kaji, N. and Kitsuregawa, M.: Efficient Word Lattice Generation for Joint Word Segmentation and POS Tagging in Japanese, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 153–161 (2013).

[10] Neubig, G. and Mori, S.: Word-based Partial Annotation for Efficient Corpus Construction, *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (2010).

[11] Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102 (2008).

[12] Mori, S., Maeta, H., Yamakata, Y. and Sasada, T.: Flow Graph Corpus from Recipe Texts, *Proceedings of the Nineth International Conference on Language Resources and Evaluation* (2014).

[13] Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, *Proceedings of the 16th International Conference on Computational Linguistics* (1996).