

少数ラベルありデータからの語義曖昧性解消

藤田 早苗

藤野 昭典

NTT コミュニケーション科学基礎研究所

{fujita.sanae, fujino.akinori}@lab.ntt.co.jp

1 はじめに

多くの語義曖昧性解消タスクでは、対象語を辞書に定義された語義に対応付ける。語義曖昧性解消では、一般に十分なラベルありデータが存在する場合、教師あり学習により高い精度を得ることができるが [4, 10]、数万から数十万にもなるあらゆる語義に人手により多くのラベルありデータを用意することは容易ではない。また、日々出てくる新しい語義に対応するためにも、できる限り少ないラベルありデータから、高い精度を得ることが重要である。そのため、ラベルなしデータを活用する方法が多く提案されてきている [1, 3, 7, 9]。

特に [3] では、ラベルありデータを自動獲得し、更に半教師あり学習を適用することで、人手作成によるラベルありデータが少ない場合でも高い精度を得ている。また、未知語義の推定も可能 (再現率が最高で 50%) になることが示されている。しかし、評価データに出現する未知語義は、9 語義、18 例と少数だった。そこで、本稿では、与えられたラベルありデータから一部の語義を削除することで、疑似未知語を作り、[3] の提案手法の未知語義に対する性能をより多くのデータで評価する。特に、人手で正解付与されたラベルありデータと、自動獲得したラベルありデータを用いる場合の結果を中心に比較する。実験には、SemEval-2010 の日本語語義曖昧性解消タスク [8] のデータを利用する。

2 適用手法概要

本章では、本稿で適用する手法の概要を紹介する (詳細は [3] を参照)。本手法は、2 段階に分けられる。まず、Step-1 では、生コーパスからラベルありデータを自動的に獲得する。これにより、未知語義や低頻度語にもラベルありデータを獲得する。

次に Step-2 では、ラベルありデータとラベルなしデータを学習データとして、半教師あり学習 (MHLE) を適用する。ここでは、より簡単に獲得できるラベル

なしデータを活用することで、精度を向上させる。

2.1 ラベルありデータの自動獲得: Step-1

Step-1 では、辞書の例文を利用し、生コーパスからラベルありデータを獲得する。図 1 に辞書の例を示す。図 1 のような辞書が与えられた場合、まず、辞書の例文を文字列として完全に含む文を抽出し、形態素解析を行なう。更に、対象例文の見出し語と、基本形、および、品詞大分類が一致する形態素に、該当する例文の語義 ID を付与する。

例えば、図 1 の場合、例文「手に取って見る」を含む文として、(1) が得られ、語義 ID37713-0-0-1-1 のラベルありデータとして利用できる。特に人間用の紙の辞書の場合、省スペース化のため、例文は非常に短いことが多い。Step-1 では、例文より長くて情報量の多い文をラベルありデータとして自動獲得できることが利点である。[3] の人手によるサンプル評価によると、自動獲得したラベルありデータのラベル正解率は、94.3% である。

(1) ガラス碗を手に取って見ると [...]

2.2 半教師あり学習の適用: Step-2

Step-2 では、半教師あり学習法 (ハイブリッド法, *Maximum Hybrid Log-likelihood Expectation*: MHLE, [2]) を適用する。MHLE では、ラベルありデータで学習させた ME モデル (識別モデル) とラベルなしデータで学習させた NB モデル (生成モデル) を統合して分類器を得る。

ここで MHLE は、ラベルありデータと評価データの間に分布差があるような場合にも頑健であるという特徴がある。[1] では、自動獲得したラベルありデータの語義分布が評価データと差がある場合、精度が低下することが示されているが、[3] では、MHLE を適用することで、語義分布の差による影響を抑え、自動獲

図 1: 岩波国語辞典の例:「とる」から抜粋

得したラベルありデータを用いる場合でも精度向上できることを示している。

[学習データ] ラベルありデータには、人手で正解付与されたデータと、Step-1 により自動獲得されたデータを用いる。ラベルなしデータには、コーパスから、形態素解析した時に対象語の基本形を含む文を抽出する。例えば (2) の場合、太字部分の基本形は、対象語「とる」と一致するので、ラベルなしデータとして利用する。

(2) 処置 がとられた から である

[素性] 素性は [3] と同じものを用いた。すなわち、各対象語に対し、出現形、基本形、品詞、品詞大分類(名詞、動詞、形容詞など)を利用する。また、対象語が i 番目の語だとすると、前後 2 語 ($i-2, i-1, i+1, i+2$) の同じ情報も利用する。更に、前後 3 語以内の bigrams, trigrams, skipbigrams も利用する。また、各対象語と同一文内に出現する全内容語の基本形も素性として利用する。

3 実験

3.1 実験対象データ

評価対象として、SemEval-2010 の日本語語義曖昧性解消タスク [8] を利用した。本タスクの対象語は 50 語であり、ラベルありデータ、および、評価データは、各語につき 50 文ずつである。これらの出展は多岐にわたり、白書、新聞、本や雑誌からなる。更に、評価データには、Web 上の Q&A サイトである Yahoo! 知恵袋のデータも含まれる。これらのデータは、現代日本語書き言葉均衡コーパス (BCCWJ)¹ のうち、形態素解析の誤りを人手で修正したコアデータと呼ばれる部分から抽出されている。また、本データには、岩波国語辞典 [6] の語義を元に、語義 ID が付与されている(図 1 参照)。図 1 に示したように、各エントリは、表記、品詞、定義文や例文などの情報を含んでいる。

また、生コーパス(ラベルなしデータ)として、BCCWJ のモニター公開データを利用した。

¹<http://www.ninjal.ac.jp/kotonoha/>

3.2 実験方法

本稿では、まず、元のラベルありデータから、一部の語義を削除し、疑似未知語義を作る。次に、元のラベルありデータ(以下、TRN)、あるいは、自動獲得したラベルありデータ(以下、AUTO)をラベルありデータに少しずつ戻していき、精度変化を調査する。

ここで、ラベルなしデータ量は各語 300 文とした。これは、[3] のほとんどの実験で最も良い精度が出ているからである。

削除する語義(以下、疑似未知語義)は、各対象語の語義の中からランダムに選択した。疑似未知語義の、タスクデータセットでの出現頻度、および、自動獲得したラベルありデータの文数を表 1 に示す。

表 1: 疑似未知語義の出現頻度

	最小	最大	平均	TOTAL
評価データ	1	50 ²	19.1	937
ラベルありデータ (TRN)	1	49	20.2	991
自動獲得データ (AUTO)	1	18,675	502.9	24,644

3.3 結果と考察

3.3.1 ラベルありデータの比較

一度、疑似未知語義のラベルありデータをすべて削除した後、TRN、および、AUTOを追加して(戻して)いった場合の学習曲線を図 2, 3 に示す。また、図 2, 3 には、ベースラインとして、TRN を戻した場合に、MHLE ではなく、MEM³を適用した場合の精度も示した(以下、BL)。ここで、図 2 は、評価データ全体の精度を示し、図 3 は、疑似未知語の推定性能 (F 値) を示している。

また、図 2, 3 の横軸は、追加する文の数であり、30 まで表示している。但し、表 1 に示したように、そもそも多くの語義は、TRN や AUTO に 30 文も存在しない。そのため、横軸で示した文の数だけ追加すること

²評価データは各語 50 文だが、評価データの中に一つの語義しか出現しない語があった(対象語「外」)

³代表的な識別モデルの一つ。ラベルありデータを用いて教師あり学習を行う最大エントロピーモデル (Maximum Entropy Method)

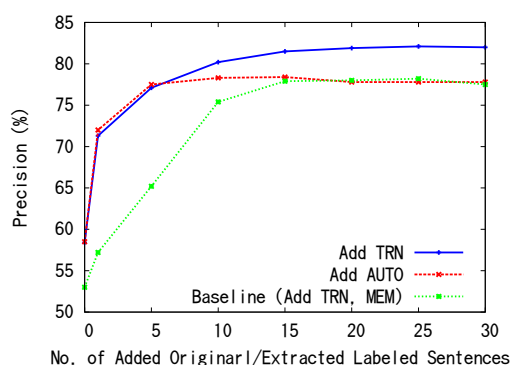


図 2: 全体の推定精度: ラベルありデータの比較 (TRN vs. AUTO)

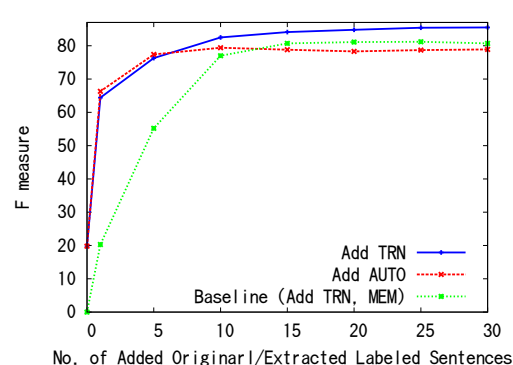


図 3: 疑似未知語の推定性能 (F 値) (TRN vs. AUTO)

ができない場合、追加できるすべての文をラベルありデータに追加した結果を用いている。つまり、追加文数が 30 とあっても、実際には 1 文しか追加できていないような語義もある。

[人手作成と自動獲得データの比較] 図 2 から、5 文追加程度までは、AUTO を追加する場合も TRN を戻す場合もほぼ同程度だが、若干 AUTO の方が精度が高くなっている。しかし、10 文追加からは、TRN の方が精度が高くなる。AUTO の場合、5 文追加以降は学習曲線はほぼフラットになっている。これは、Step-1 で獲得するデータに「例文を含む」という制約を設けているため、バリエーションに限りがあることや、間違いが含まれることによる影響だと考えられる。また、TRN 追加の場合も、15 文程度からほぼフラットになっている。ラベルありデータの文数が平均 20 文程度なので、そもそも追加できる文がなくなってきていることが大きいと思われる。

[ベースライン (MEM) との比較] MEM によるベースラインには、TRN と同じラベルありデータを用いているにも関わらず、MHLE による精度の方が常に高い精度となった。最も差分が大きいのは、1 文追加時点で +14.1% 差だった。これは、MHLE がラベルありデータが非常に少数でも、高い分類性能を持つことを示している。言い替えると、MHLE を適用する場合、未知語義に対してラベルありデータを 1 文 (あるいは、5 文) 追加すれば、MEM を適用する場合に 8 文 (あるいは、15 文) 追加しなければ得られない分類性能を得ることができる。しかも、5 文追加時点では、AUTO でも同等の性能が得られている。また、AUTO と BL を比較すると、ラベルありデータが少ない時点の優位性はもちろん、それ以降も同等の精度を得ることができた。

[疑似未知語義の推定性能] 疑似未知語義に対する推定

性能 (図 3) を比較すると、全体精度と同様に、立上りは TRN、AUTO の方が圧倒的に良い。対象語義のラベルありデータを追加するに従い、BL は AUTO を上回るようになるが、TRN を上回ることではない。なお、追加文数が 0 の場合、疑似未知語義のラベルありデータは存在しないため、教師あり学習である MEM の場合 F 値は 0 % である。しかし、MHLE は、F 値 19.8 % (再現率 11.1 %) で疑似未知語義を推定できている。これは、半教師あり学習である MHLE では、ラベルなしデータから周辺分布を学習する時に、既知クラスと非常に異なる場合、未知クラスに分類するからである。更に、1 文でもラベルありデータが与えられると、AUTO の場合で、F 値 66.3 % (再現率 51.8 %) と格段に向上する。

また、AUTO の結果は、TRN や BL より最終的には低くなるが、未知語に対する性能であり、人手によるラベルありデータを用いていないことを考えると非常に高い精度だといえる。

3.3.2 自動獲得データの事前追加の有効性

図 2, 3 では、疑似未知語義の学習データを完全に削除し、そこに、TRN、AUTO を追加していった。しかし、そもそも、Step-1 の手法の利点は、全自動でラベルありデータを獲得できることにある。そこで、自動獲得データが悪影響さえ与えないのであれば、あらかじめ自動獲得データを与えておくことで、一定以上の精度を担保し、時間とコストに余裕があれば、人手で訓練データを追加して精度を向上させることが理想的である。

そこで、図 4, 5 に、あらかじめ自動獲得データ (AUTO) を一定数与えておき、そこに TRN (つまり、人手で正解付与されたラベルありデータ) を追加した場

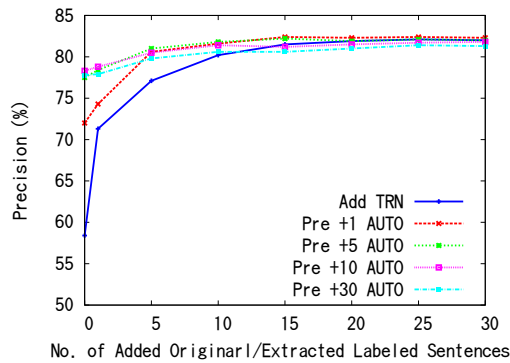


図 4: 全体の推定精度: あらかじめ自動獲得データ (AUTO) を追加しておく場合

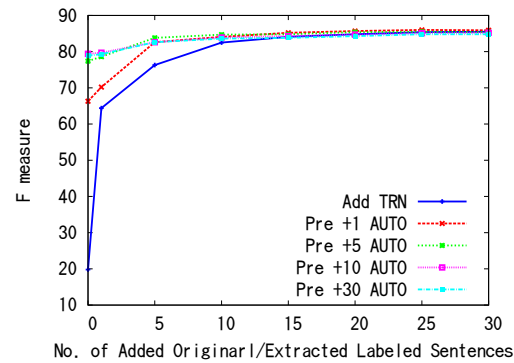


図 5: 疑似未知語の推定性能 (F 値): あらかじめ自動獲得データ (AUTO) を追加しておく場合

合の学習曲線を示す。図 4, 5 では、あらかじめ AUTO を 1, 5, 10, 30 文追加しておいた場合を示している。また、比較のため、図 2, 3 の TRN の結果も再掲している。

図 2, 3 から、AUTO をあらかじめ追加しておいても、精度は下がらないことがわかる。また、あらかじめ 5 文も追加しておけば、全体精度が 77.5%、未知語義の推定性能も F 値で 82.6%、あらかじめ 10 文追加しておけば、全体精度が 78.3%、未知語義の推定性能は F 値で 84.1%と、かなり高い初期値を得られる。更に、このようにあらかじめ AUTO を追加しておけば、TRN を 10 文程度追加するだけで、TRN のみの場合のほぼ最高精度となるため、人手で作成しなければならないラベルありデータ量の削減が可能である。

すなわち、Step-1 の手法により、あらかじめ、5 文から 10 文程度のラベルありデータを獲得しておき、コストと時間に応じて人手でラベルありデータを追加していくことが良いと考えられる。

4 おわりに

本稿では、ラベルありデータの自動獲得と、半教師あり学習 (MHLE) の組合せによる語義曖昧性解消方法 [3] について、ラベルありデータの質、量と、性能の関係を評価した。

本稿では疑似未知語を選び、疑似未知語義のラベルありデータをすべて削除した後、人手作成によるラベルありデータと、自動獲得データを追加した場合の性能比較を行なった。その結果、ラベルありデータとして、自動獲得データを用いた場合でも、人手作成によるラベルありデータほどではないが、高い性能を得た。更に、あらかじめ自動獲得データをラベルありデータ

に追加しておいても悪影響はなく、人手で作成するラベルありデータ量の削減が可能であることをしめた。

本稿の実験では、未知語義のラベルありデータが高々数文でも高い精度を得られている。今後は、実データを用いて、本手法が未知語に対して容易に対応可能であることを実証していきたい。

参考文献

- [1] Eneko Agirre and David Martinez. Unsupervised WSD based on Automatically Retrieved Examples: The Importance of Bias . In *Proceedings of EMNLP-2004*, pp. 25–32, 2004.
- [2] Akinori Fujino, Naonori Ueda, and Masaaki Nagata. A Robust Semi-supervised Classification Method for Transfer Learning. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pp. 379–388, 2010.
- [3] Sanae Fujita and Akinori Fujino. Word Sense Disambiguation by Combining Labeled Data Expansion and Semi-Supervised Learning Method. In *Proceedings of IJCNLP-2011*, pp. 676–685, 2011.
- [4] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. 技術資料 SENSEVAL-2J 辞書タスクでの CRL の取り組み- 日本語単語多義性解消における種々の機械学習手法と素性の比較. 自然言語処理, Vol. 10, No. 3, pp. 115–134, 2003.
- [5] Kamal Nigam, John Lafferty, and Andrew McCallum. Using Maximum Entropy for Text Classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67, 1999.
- [6] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典. 岩波書店, 1994.
- [7] Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proceedings of ACL-2005*, pp. 395–402, 2005.
- [8] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. SemEval-2010 Task: Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 69–74, 2010.
- [9] Thanh Phong Pham, Hwee Tou Ng, and Wee Sun Lee. Word Sense Disambiguation with Semi-Supervised Learning. In *AAAI-2005*, pp. 1093–1098, 2005.
- [10] Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of EMNLP-CoNLL-2007*, pp. 477–485, 2007.