ユーザの興味があるカテゴリに応じた NER システム構築フレームワーク

芝原 隆善 [1,2]　大内 啓樹 [1]　山田 育矢 [2,3]　西田 典起 [2]
寺西 裕紀 [2]　古崎 晃司 [2,4]　渡辺 太郎 [1]　松本 裕治 [2]
[1] 奈良先端科学技術大学院大学 [2] 理化学研究所 [3]Studio Ousia [4] 大阪電気通信大学
{shibahara.takayoshi.sk4, hiroki.ouchi, taro}@is.naist.jp
{takayoshi.shibahara, hiroki.ouchi, ikuya.yamada, noriki.nishida,
hiroki.teranishi, kouji.kozaki, yuji.matsumoto}@riken.jp　ikuya@ousia.jp
kozaki@osakac.ac.jp

概要

異なるユーザーは異なる固有表現のカテゴリに関心を持つ．そのため本論文ではユーザーが関心を持つ固有表現を抽出可能な固有表現抽出システムを構築するフレームワークを提案する．このフレームワークでは，まずユーザーに関心のある固有表現カテゴリをシソーラスから選択してもらい，次に選択されたカテゴリを含むシソーラス全体のカテゴリ情報を活用した Distant Supervision を行い固有表現抽出モデルを学習する．本論文では UMLS シソーラスと MedMentions コーパスを利用したいくつかの実験を通じて提案フレームワークの有効性を確認した．

1　導入

固有表現抽出: Named Entity Recognition (NER) は自然言語処理の基本タスクの 1 つであり，質問応答 [1]，関係抽出 [2]，Entity Linking [3]，対話システム [4] など様々なタスクで応用されている．

本論文は，「異なるユーザは異なる固有表現 (Named Entity: NE) のカテゴリに興味を持つ」という着眼点に立っている．例えば，疫学者と薬学者が COVID-19 関連論文を調査する場合を考えてみる．疫学者は人間集団ごとの症状に関心を持ち，症状や人々の特徴 (性別，民族性，年齢など) に関する固有表現を求めるだろう．一方薬学者はコロナウイルスやその関連薬の作用機序に関心を持ち，解剖学的構造や化学物質に関連する固有表現に注目するだろう．

しかし，既存の固有表現認識手法では異なるカテゴリを欲するユーザーの要求に柔軟に対応できない．通常 NER システムは，固有表現カテゴリがあらかじめ定義されたデータセットを用いて構築される．それゆえ，ユーザにとって関心のある固有表現カテゴリが予め定義されたデータセットが存在しなければ，それらの固有表現は抽出されない．各ユーザが興味のある固有表現カテゴリを人手でアノテーションすることもできるが，これには大変なコストと時間がかかってしまう．

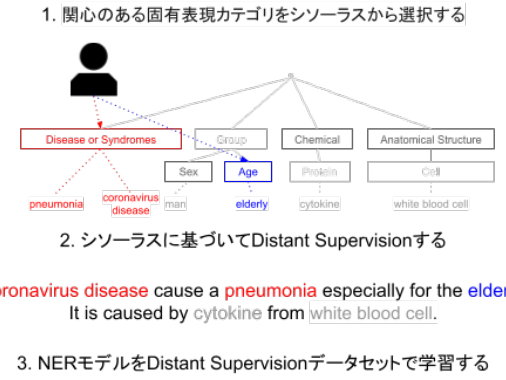

図 1　タスクの全体像

そこで本論文では，ユーザが興味をもつ固有表現を抽出可能な NER システム構築フレームワークを提案する (図 1)．まず，このフレームワークでは各ユーザーに興味のある固有表現カテゴリをシソーラスから選択してもらう．そして，その選択されたカテゴリに基づき，Distant Supervision によりテキストを擬似的にアノテーションし，その擬似アノテーションを用いてシステムを学習させる．その結果，システムはユーザの関心のある固有表現を抽出することができるようになる．

しかしながら Distant Supervision NER の問題とし

て，Unlabeled Entity Problem が指摘されている [5]．Unlabeled Entity Problem とは，シソーラスにない未知の固有表現を Distant Supervision でラベル付けすることが難しいという問題である．

我々はこの Unlabeled Entity Problem を擬似負例の信頼性の問題であると考える．従来の研究では，信頼性の低い擬似負例がもたらす影響を割り引くために，擬似負例の一部は本来正例であると仮定して NER タスクを実行していた．本研究では，シソーラスのカテゴリ情報を利用して，より信頼性の高い擬似負例を擬似データセットに追加し (3 節)，擬似負例の信頼性の低さの問題を緩和する．また，本フレームワークの有効性を検証するために，タスクと実験設定を提案・定式化する．本論文ではいくつかの実験を通じて提案法の有効性を確認した．

要約すると，我々の貢献は主に以下の 3 つである．

- 本論文は，ユーザーの興味に応じた固有表現を抽出するオーダーメイドの NER システムを構築するためのフレームワークを開発した最初の研究である．
- Unlabeled Entity Problem を解決するために，シソーラスのカテゴリ情報を有効に活用する新しい手法を提案した．
- 「ユーザが興味を持つ固有表現カテゴリに応じた NER システム構築」の状況をシミュレートできるタスクと実験設定を策定・実施した．

2 タスク設定

提案法は Distant Supervision NER (DS NER) に基づいている．固有表現抽出: Named Entity Recognition (NER) は n 長の単語列: $X = \{w_j\}_{j=1}^{n} \in V^n$ に対し m 個のスパンとそれらのラベル: $Y = \{(s_i, e_i, l_i) \in [0..n-1] \times [1..n] \times L\}_{i=0}^{m-1}$ を事前に定義されたラベル集合: L から予測するタスクとして形式的には定義される．通常の教師あり設定ではアノテーションされたデータセットを元に学習するが，DS NER では，文書と用語集から構築される擬似データセットを利用することでアノテーションを省く．擬似データセット:$\{X_k \in V^*, \tilde{Y}_k = \{(\tilde{s}_{ki}, \tilde{e}_{ki}, \tilde{l}_{ki})\}_{i=0}^{m_k-1}\}_{k=1}^{|\mathfrak{D}|}$ [1] は文書: $\mathfrak{D} = \{X_k\}_k \subset V^*$ と用語集 $g : T \to L$ から構築される．ただしここで $T \subset V^*$ は用語集の中で記述される用語の集合である．用語集 g は文書 $\mathfrak{D}$ に対する文字列マッチによって擬似データの構築に利用される．しかし，この DS NER のタスク設定には，Unlabeled Entity Problem [5] と呼ばれる問題がある．これは，用語集の被覆率が低いために，擬似データセット中の固有表現が見逃され，学習モデルがこれらのラベル付けされていないスパンを無視してしまうという問題である．例えば，使用した用語集に新しい化学物質が記述されていない時に，DS NER モデルがこの化学物質を見逃すという可能性がある．この問題に対処するため，先行研究では信頼性の低い擬似負例が与える影響を割り引き，擬似負例の一部が本来正例であると仮定して NER タスクを実行している．具体的には，lenient CRF [6]，Self-Training [7]，PU 学習 [8]，擬似負例スパンのアンダーサンプリング [5] によって疑似負例に正のラベルを予測する手法が知られている．しかし，これらの擬似負例の影響を割り引く手法は，本来負例であるスパンまで固有表現として予測してしまうという危険性がある．

本研究では，ユーザーが選択したカテゴリ L に応じて用語集を作成することで，ユーザーごとにオーダーメイドな NER モデルを実現する．ここでシソーラスは Directed Rooted Tree DRT = (C, E, r) であるとする．つまり，シソーラスは，概念の集合 C，C 上の辺としての is-a 関係の集合 E，ルートノード r からなるものとする．本研究では L だけではなく DRT を利用して負例のカテゴリ L_{neg} を作成し，L_{neg} と L 両方の語句を含む用語集 g' を利用した Distant Supervision を行う．このことにより提案手法がシソーラス全体の情報やクラス間の排他的な関係を捉えられるようになることを目指す[2]．ここで L_{neg} は L と組み合わせることでシソーラス全体を被覆するように定義する．つまり L に含まれるいずれかの概念の子孫の全ては L によって被覆されているので，負例カテゴリ L_{neg} を L によって被覆されていない DRT に含まれる概念の最小の集合とする[3]．例えばユーザーが {"Diseases or Symptoms", "Age"} を正例の概念として選んだ場合，提案手法は L_{neg} として {"Sex, "Chemical", "Anatomical Structure"}

1) m_k は k 番目の文に含まれているスパンの個数である．

2) より形式的には，提案手法のタスク設定は用語集 $g' : T \cup T_{neg} \to L \cup L_{neg}$ が負例クラス L_{neg} を含んでいるという点で通常の Distant Supervision と異なっている．ただし T_{neg} は L_{neg} に含まれる語句の集合である．

3) 形式的には L_{neg} は L' を利用して次のように定義できる．$L' = \{a | l \in L \wedge (l, a) \in E^*\}$ $L_{neg} = \{c | l' \in L' \wedge (l', p) \in E \wedge (c, p) \in E \wedge l' \neq c \wedge c \notin L'\}$ ただし E^* は E の反射推移閉包であるとする．

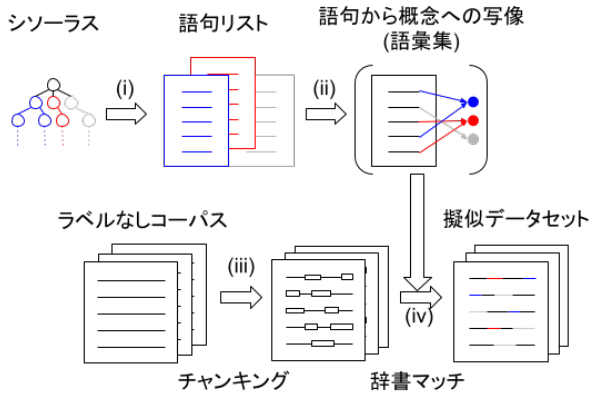

図 2　Pseudo Dataset Construction

を選ぶ (図 1).

文字列マッチによってこの負例カテゴリ L_{neg} を利用して得られた擬似負例は，確実に L に含まれないカテゴリであることが保証されるので，単純に辞書マッチしなかったという従来の擬似負例よりもより負例として正しいことが期待される．また負例クラスを利用することで Distant Supervision NER モデルが本来の負例に対して過剰な固有表現の予測を行うという，従来の Unlabeled Entity Problem 対策手法の欠点を補うことができると考えられる．例えば先の例であれば，Distant Supervision NER モデルが本来 Chemical であるスパンに対して，Disease or Syndrome とラベル付けすることを防ぐことができると期待される．

3　手法

3.1　擬似データセット構築

擬似データセットを構築する手順を，図 2 に記述する．(i) シソーラスから L，L_{neg} の各カテゴリの用語リストを構築する．例えば，UMLS は固有表現のカテゴリを階層構造で定義しており，階層を推移的に利用することで各カテゴリの用語リストを容易に得ることができる．また，用語のリストに対し，inflector[4] という python ライブラリを用いて，用語リスト上の各用語の複数形・単数形を追加する．(ii) L，L_{neg} の用語リストを用いて，用語からカテゴリへの対応付け (用語集)g，g' を計算する．ただし，複数のクラスに含まれる曖昧な用語がある[5]．このような曖昧な用語を用語集から削除することで，エラー伝搬を低減する．(iii) ラベルの付与されていないコーパスに対して Noun Phrase (NP) Chunker を適用し，擬似アノテーションの対象となる候補スパンを獲得する．(iv) (ii) の用語集 g，g' と (iii) の候補 NP チャンクとの間に文字列マッチアルゴリズムを適用し，対象の NP チャンクの末尾部分文字列に基づいて分類する．文字の大小で異なる用語が存在する場合は大文字と小文字を区別して NP チャンクを分類し，そうでない場合は区別せずに分類する．

3.2　スパン分類モデル

本研究では BERT による符号化を利用した簡素なスパン分類モデルを利用しており，n トークンからなる文 $X = \{w_j \in V\}_{j=1}^{n}$ の候補スパン $(s, e) \in [0..n-1] \times [1..n]$ をラベル集合:$L \cup \{\text{“}O\text{”}\}$[6] に分類する．

まず，BERT の最終隠れ層: $v = \{v_j\}_{j=1}^{n}$ を利用し入力文を符号化する[7]．この符号化ベクトルに対し Dropout を適用した: $\tilde{v} = \text{Dropout}(v)$．Dropout の後で，スパンの開始位置と終了位置のベクトルを連結したスパン表現を分類した: $\hat{y} = \text{Softmax}(W^T(\tilde{v}_s \oplus \tilde{v}_e) + b)$．この予測確率に対し交差エントロピーを利用して学習を行った．ただし，Li ら [5] と同様に擬似ラベルが O である，つまりいかなる用語集の用語にもマッチしなかったスパンに対して，そのロスを事前に定義された確率で無視することで擬似負例のノイズを軽減する．

3.3　前処理・後処理

前処理として，NER データセットをスパン分類データセットに変換する．長さがあらかじめ定義された最大長より小さいスパンを列挙し，ラベルを付ける．NER データセットでラベル付けされているスパンにはそのラベルを付与し，そうでないスパンには O ラベルを付与する．

後処理として，予測時にスパンの分類スコアをスパン重複のない NER 出力に変換する．具体的には，学習したモデルによって予測される最尤確率の高い順に，貪欲に列挙されたスパンを埋めていく．ただし，最大尤度のラベルが O であったり L_{neg} に含まれるスパンは予測に利用しない．

4)　https://github.com/ixmatus/inflector
5)　例えば，AAA は Abdominal Aortic Aneurysms という病名の頭文字であると同時に，APP gene という遺伝子名の別名でもある．
6)　提案手法の場合，ラベル集合は $L \cup L_{neg} \cup \{\text{“}O\text{”}\}$ である．
7)　トークンの最後に位置するサブワードのベクトルをトークンの表現として利用した．

Method	Strict P.	Strict R.	Strict F.	Lenient P.	Lenient R.	Lenient F.
Chunker Match	29.81	16.78	21.47	62.00	36.22	45.73
Span Classif. w/ N.U.	24.19	22.94	23.54	51.70	51.25	51.47
+Thesaurus Negatives	24.04	23.35	23.69	52.60	53.64	53.12

表 1 MedMentions で指定された 21 クラスを選択したときの NER スコア: Strict はスパン完全一致，Lenient はスパン部分一致で評価.

4 実験設定

1 節で述べたような「ユーザにシソーラスから興味のあるクラスを選択してもらう」状況を模倣した実験を行う．本論文では，シソーラスとして UMLS (2021AA 版) [9] を，DS NER の評価には MedMentions コーパス [10] を利用する．UMLS は生物医学分野のシソーラスで，127 のクラス (Semantic Types) と 16,132,274 の用語を持つ．MedMentions は，4,392 の生物医学論文抄録に 352,496 のスパンが UMLS の概念に対応付けられるようにアノテーションされている Entity Linking/NER 用のコーパスである．train/dev/test の文書数はそれぞれ 2,635/878/879 である．train 部分のデータセットは Distant Supervision の擬似アノテーションの対象として，dev/test の分割は正解アノテーションまま利用する．ただし，評価時にユーザが興味のあるクラスのみを対象とするため，ユーザが興味のない (と実験上仮定する) クラスは dev/test データから除外している．

5 結果

表 1 の実験では MedMentions [10] が指定する 21 のクラスをユーザーが UMLS から選択したと仮定し評価を行った[8]．具体的には Strict/Lenient NER P./R./F. を文字列マッチ，Distant Supervision のベースライン，提案法で比較した．Strict/Lenient NER P./R./F. は，スパン完全一致 (Strict) またはスパン部分一致 (Lenient) のどちらかで計算された Precision/Recall/F1 スコアである．Chunker Match の行は，擬似データセット構築に利用した NP chunker を利用した文字列マッチ手法のスコアを示している．Span Classif. w/ N.U. の行は，Li ら [5] のように擬似負例をアンダーサンプリングすることで Unlabeled Entity Problem に対処したスパン分類モデルにおけるスコアを示す．また，+Thesaurus Negatives の行は，Span Classif. w/ N.U. に追加してシソーラスに基づいたより信頼できる負例[9]を利用した場合のスコアを示している．

Baseline モデル (Span Classif.) を文字列マッチ手法と比べた時 Recall が向上していることがわかる．これは擬似負例のアンダーサンプリングにより，擬似負例のスパンに対して正のラベルを予測できているためであると考えられる．提案手法は，ベースラインと比較して，Strict 設定では F1 の差が小さく，Precision が若干低下するが， Lenient 設定では Precision， Recall ともに向上することが示された．Distant Supervsion によるアノテーションでは，表 1 の Chunker Match の行にあるように，完全なスパン一致を得ることは非常に困難である．したがって，Lenient 設定に着目すると、この結果はシソーラスを用いた負例カテゴリの利用の有効性を示していると言える．

6 結論

本論文においてユーザーがシソーラスから選択した固有表現を認識するフレームワークを示した．本研究では私達は擬似正例だけでなく，擬似負例に着目し，より負例として信用でき，難しい擬似負例の重要性を明らかにした．また本実験は関心のある固有表現だけではなく，それらの固有表現と排他的なカテゴリ活用の重要性を明らかにした．

今後の課題としては，より具体的で粒度の細かい固有表現を取得したり DBPedia [11] のような巨大シソーラスで見られるシソーラスのノイズに対応するなどが考えられる．また，Mention レベルと Entity レベルの間にはスパン情報の大きな違いがあり、文字列マッチのみでは限界がある。そのため，少量のアノテーションを組み合わせていくのことも有用だと考えられる [12, 13].

謝辞

本研究は JSPS 科研費 JP19K20351 の助成を受けたものである．

8） 選択したクラスによる精度のばらつきを見るために、10 個のクラスから 1 つずつクラスを選択し, それぞれのクラスを認識するように学習させた 10 個の NER モデルのスコアを評価する実験も追加実験として行った (付録 A).

9） シソーラスを利用した擬似負例の信用性の改善は実際に確認された (付録 A).

参考文献

[1] Mahboob Alam Khalid, Valentin Jijkoun, and Maarten de Rijke. The impact of named entity normalization on information retrieval for question answering. In **Advances in Information Retrieval**, pp. 705–710. Springer Berlin Heidelberg, 2008.

[2] Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. Neural relation extraction via Inner-Sentence noise reduction and transfer learning. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2195–2204, Brussels, Belgium, 2018. Association for Computational Linguistics.

[3] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Z Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. **IJCAI**, 2015.

[4] Kevin Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker. SlugNERDS: A named entity recognition tool for open domain dialogue systems. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[5] Yangming Li, Lemao Liu, and Shuming Shi. Empirical analysis of unlabeled entity problem in named entity recognition. September 2020.

[6] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 729–734, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. BOND: BERT-Assisted Open-Domain named entity recognition with distant supervision. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '20, pp. 1054–1064, New York, NY, USA, August 2020. Association for Computing Machinery.

[8] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. Distantly supervised named entity recognition using Positive-Unlabeled learning. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2409–2419, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic Acids Res.**, Vol. 32, No. Database issue, pp. D267–270, January 2004.

[10] Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with UMLS concepts. p. 13, 2018.

[11] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Others. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. **Semantic web**, Vol. 6, No. 2, pp. 167–195, 2015.

[12] Lukas Lange, Michael A Hedderich, and Dietrich Klakow. Feature-dependent confusion matrices for low-resource NER labeling with noisy labels. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[13] Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. Named entity recognition with small strongly labeled and large weakly labeled data. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

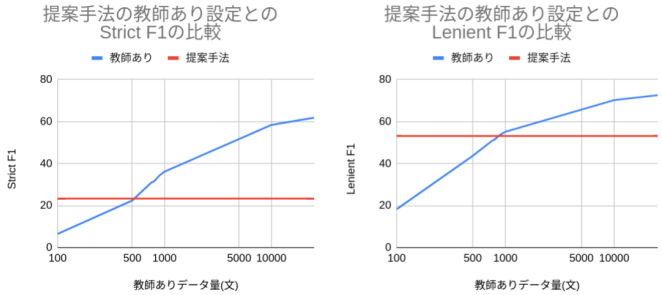

図 3 教師あり設定との比較: 提案手法とデータセット量を変えたときの教師ありモデルに対する完全マッチ (Strict) と部分マッチ (Lenient) に基づく NER F1 スコア

	T005 Virus			T007 Bacterium			T017 Anatomical Structure			T022 Body System			T031 Body Substance		
Method	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.
Chunker Match	94.23	34.00	49.97	88.08	34.76	49.85	45.68	48.80	47.19	21.67	15.29	17.93	29.23	51.52	37.29
Span Classif. w/ N.U.	64.08	45.33	53.10	20.29	77.58	32.17	22.75	68.81	34.20	02.16	23.53	03.96	07.10	78.79	13.03
+Thesaurus Negatives	85.14	41.33	55.65	78.68	67.00	72.38	45.57	60.24	51.89	45.00	10.59	17.14	33.15	64.14	43.71

	T033 Finding			T037 Injury or Poisoning			T038 Biologic Function			T058 Health Care Activity			T062 Research Activity		
	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.
Chunker Match	29.30	30.76	30.01	44.24	60.00	50.93	55.78	52.41	54.04	44.33	50.49	47.21	61.66	63.04	62.34
Span Classif. w/ N.U.	13.62	60.65	22.24	06.27	88.75	11.72	40.02	70.32	51.01	28.46	68.15	40.15	31.43	73.97	44.12
+Thesaurus Negatives	35.82	27.08	30.84	66.10	25.00	36.28	67.08	62.93	64.94	68.20	42.39	52.28	67.31	30.31	41.80

表 2 カテゴリを一つずつ選んだときの Lenient P./R./F.: それぞれの行は文字列マッチ，Distant Supervision におけるベースライン，提案手法を示す (表 1 と同様)．列はそれぞれの UMLS カテゴリ (T***) を関心のあるカテゴリとして選択したときの lenient NER P./R./F. スコアを示す．

A 追加実験

MedMentions [10] で指定された 21 クラスを着目クラスとして選択した際の追加実験として，擬似負例の負例としての正しさを Negative NP P./R. で確認した．Negative NP P./R. とは Gold で固有表現とアノテーションされていない NP[10] を負例の正解としたとき，作成された擬似負例がこれに一致する割合としての Precision/Recall である．辞書マッチしなかった NP スパンの擬似負例としての Negative NP P./R. は 58.00/59.57 となった．一方でシソーラスに基づいた負例カテゴリを用いた擬似負例の Negative NP P./R. は 78.25/29.91 となった．我々の提案したシソーラスに基づく擬似負例は Negative NP Recall では劣るが，Negative NP Precision では改善している．

このことから確かに，我々のシソーラスに基づく擬似負例はより信頼性の高いものになっているといえる．更にシソーラスを利用した擬似負例は NP Chunker を利用しているので句や NP になっていない，明らかに固有表現でない擬似負例を含まない．負例として信頼性が高くより識別が難しい擬似負例が，表 1 でみられた精度改善をもたらしていると考えられる

また，図 3 における実験では，教師ありデータ量を変化させて教師ありモデルの精度を確認し，擬似負例のアンダーサンプリングとシソーラスに基づいた擬似負例を用いた提案のスパン分類モデルと比較した．その結果，完全教師ありの設定と比較して，Lenient f1 では 20%近くスコアが減少した．しかし，提案手法では，800〜900 文の人手アノテーションと同等の Lenient f1 が得られることが分かった．また， Strict f1 では，完全教師ありモデルとの間で 48%近い差が生じた．この結果は， Distant Supervsion モデルでは，スパン範囲を正確に捉えることが困難であることを示唆している．この困難は Entity レベルと Mention レベルの間の違いに原因があると考えられる．例えば，“water” が化学物質として登録されていて，擬似アノテーションの対象文に “purified water” があった場合，“water” までを取るべきか “purified water” までを取るのが良いのかが分からないというような困難である．このような Mention と Entity の間の違いに対処するには少量のアノテーションを組み合わせることも必要だと考えられる．

表 2 は，シソーラス上の 10 個の概念から 1 つずつ着目する概念を選択し Distant Supervision を行った際の，lenient P./R./F. の比較を行ったものである．擬似負例のアンダーサンプリングだけのベースラインでは，Precision を犠牲に Recall を増加させている．しかし，Unlabeled Entity に対して固有表現ラベルを予測するため，過剰に固有表現を予測しやすく，f1 スコアが低下することが多い．擬似負例のアンダーサンプリングだけのベースラインと比較すると，シソーラスに基づいた擬似負例を利用したモデルはより頑健であり，T062 を除いた全てで f1 を向上させることができた．しかし，T022，T033，T037，T058，T062 では，Recall が文字列マッチの水準よりも低下している．その結果，T022，T037，T062 の f1 スコアが減少した．これは負例クラスの追加により，Unlabeled Entity に対する固有表現のラベル予測が過剰に妨げられているためだと考えられる．

10) より正確には，固有表現と NP のスパンが完全一致するとは限らないため，固有表現と部分一致しない NP を評価対象にした．