文脈化された単語埋め込みの語義への単語頻度と語義分布の影響

Cao Zhihan
東京工業大学　情報理工学院
cao.z.ab@m.titech.ac.jp

徳永健伸
東京工業大学　情報理工学院
take@c.titech.ac.jp

概要

本研究では，事前訓練言語モデルによる文脈化された単語埋め込みについて，異なる粒度の意味を捉える能力の単語頻度とその語義分布との関連を探査子（probe）を用いて調査する．特に，文脈化という過程と文脈化された単語埋め込みの異方性に注目した．結果として，文脈化された単語埋め込みは語義の粒度と関係なく，低エントロピーで高頻度の単語ほどその意味をより捉えている．また，文脈化により単語埋め込みの意味は網羅的になり，意味の網羅性は文脈化された単語埋め込みの形状（異方性）に反映されている．高頻度語ほど文脈化によってきめ細かい意味を正確に取り込める．

1　はじめに

人間は初見の単語でも文脈からのその意味を推測することができる．たとえば，あまり目にすることのない「涵養」という単語は，たとえ単独で意味がわからなくても，「幼少期から読書が好きで，読解力を涵養してきた」という文を見たら，おおよそ「養う」と似ている意味だと推測できる．このように，人間は単語を文脈化して理解している．

しかし，このような理解の仕方だけでは，「涵養する」と「養う」とのニュアンスの相違までは説明しきれない．つまり，人間は見なれない低頻度語を文脈によって理解するときに，その語義を完全に把握できない可能性がある．事前訓練言語モデル（Pre-trained Language Model, PLM）も単語を文脈化して意味空間に埋め込んでいるが，PLM による文脈化された単語埋め込み（Contextualized Word Embeddings, CWE）は低頻度の単語を十分に捉えているだろうか．このような問題意識のもとで，本研究では，語義が付与された SemCor コーパス [1] を利用して，単語頻度が文脈された単語埋め込みの語義への影響を「探査子 (probe)」を用いて調査する．

2　先行研究

静的な単語埋め込みと頻度との関係は，理論的にも実証的に研究されており，高頻度の単語の静的単語埋め込みは単語の意味をよりよく捕まえている傾向が示されている [2, 3, 4]. 静的単語埋め込みに比べると，CWE と頻度に関する先行研究はまだ少ない．Yu らは Transformer 系の PLM の統語的側面を探求した [5]．結論として，これらの PLM は確かに一部の単語の統語的性質をよりよく捉えてはいるが，単語頻度はそれを説明できる要因ではなかった．一方で，Zhou らは，CWE は同じ単語の埋め込みがなす意味部分空間の大きさは，その単語頻度に比例していると主張している [6].

このように，頻度と CWE との関係については見解が一貫していない．また，CWE の語彙意味論的な側面に注目した研究があるが [7]，それを単語頻度と関係づけて論ずるものは見られなかった．本研究ではこの議論上の空白を補填するために，BERT 系 PLM を対象とし，以下の研究課題を設定する．

RQ1：CWE は異なる粒度の語義を捉えているか
RQ2：文脈化前後の単語埋め込みで異なる粒度の語義を捉える能力は変わるか
RQ3：CWE の形状は異なる粒度の語義を捉える能力にどう関与するか
RQ4：RQ1–3 は単語頻度と語義分布との関係するか

RQ2 でいう文脈化以前の単語埋め込みは，BERT などの PLM の埋め込み層の出力のことをさす．その理由は，埋め込み層ではトークンと位置などの埋め込みの和にレイヤー正則化を適用するが，この埋め込み層の出力は文脈が関与しないからである．RQ3 でいう CWE の形状は，本研究で CWE の異方性 (anisotropic) をさす．異方性は BERT 系 CWE の各次元の値のスケールが相違する性質であり，それの

形状的特徴と考えられる．BERT 系 CWE の異方性を正則化すれば，CWE を用いた情報検索の性能を上げられ，異方性と語義との関連があると言われている [8]．本研究ではそれに基づき，RQ3 では CWE の形と異なる粒度の語義との関連を探る．RQ4 で語義の分布を考慮する理由は，高頻度の単語ほど多義となり，語義分布のエントロピーが高くなる可能性があるので，仮に頻度と語義との関連が見られても，それは果たして頻度だけによるものかはわからないからである．語義分布のエントロピーを一定の数値に固定すれば，語義と頻度との関係をより正確に把握できる．なお，本研究でいう単語頻度はレンマの頻度で，エントロピーは以下で定義する．

$$H(S \mid W = w) = \sum_{s \in S} P(s \mid W = w) \log P(s \mid W = w) \tag{1}$$

ただし，S は語 w の全ての可能な語義の集合で，$P(s \mid W = w)$ は語 w がコーパス中で語義 s として使われた Laplace 平滑化後の頻度である．

3 方法

本研究では，上述の研究課題に対して探査子を用いてアプローチする．この手法では，PLM が生み出した単語 w の CWE $\mathrm{cwe}(w)$ を，分類器 $\mathrm{probe}(\cdot)$ の入力として与え，対象とする言語学的性質 l を予測する．この予測の性能によって，性質 l が埋め込みにどの程度反映されているかを評価する．本研究では探査タスクを 3 種類設定した．これらのタスクを 3.1 で詳述する．3.2 ではデータセットの構成方法を，3.3 では評価方法を説明する．

3.1 語義の探査タスク

本節では，RQ1 に答えるために CWE に粒度の異なる語義の情報がの埋め込みにどの程度反映されているかを調査する実験について述べる．

多義語は周囲の文脈がによって，文中での意味が確定できる．したがって，もし CWE にうまく語義が埋め込めていれば，埋め込みから正しい語義に分類できるはずである．このような仮説を踏まえて，WordNet における Synset，Supersense，Hypernyms の異なる粒度の語義を同定するタスクを設定した．

Synset と Supersense 探査タスク (以下，syp と sup) では語の CWE から，その語の synset，supersense を探査子で予測する．ここでいう synset と supersense はそれぞれ WordNet [9] で定義されているものであり，両者とも単語の意味を示すものであるが，前者が比較的具体的な意味を示すのに対し，後者は品詞と意味カテゴリだけに対応する抽象度が高いものである．したがって，syp と sup では CWE の異なる粒度の語義を埋め込む性質を調査できると考えられる．syp と sup では精度（syp-acc と sup-acc）を性能の指標とする．

表 1 実験で使用する性能指標

指標	意味
syp-acc	単体の抽象的語義の正解率
sup-acc	単体の具体的語義の正解率
hyp-pre	マルチレベルの抽象度での語義の精度
hyp-rec	マルチレベルの抽象度での語義の再現率

Hypernyms 探査タスク (以下，hyp) では syp と同じように synset を予測するが，1 つの synset だけでなく，hypernym 階層において根までの経路中のすべての synset を予測するマルチラベル分類をおこなう．hyp を取り入れたひとつの理由は，syp ではすべての synset を同等に扱い 1 つの synset のみを出力するが，synset 同士は hyponymy の関係をなしたりして必ずしも同等ではないからである．もうひとつの理由は，supersense と synset では意味の抽象度の隔りがかなり大きいが，その中間の粒度の意味の違いを調査するためである．hyp を取り入れることで，CWE の意味の抽象度の違いをより細かく評価できると考えられる．hyp では精度と再現率（hyp-rec，hyp-pre）を性能の指標とする．本研究で採用するすべての性能指標を表 1 にまとめた．線形的な探査子は $\mathrm{cwe}(w_i)$ と l_i との真のパターンをより捉えられると報告されて，また，L2 正則化はこの能力を高めると報告される [10]．その結果を受けて，本研究の全ての探査子は L2 正則化された線形的モデルを使用する．ただし，syp と sup の探査子では出力層で softmax 関数を，hyp では sigmoid 関数を使用する．

3.2 データセット

SemCor コーパスから単語埋め込みと正解語義のペアを抽出し，実験用のデータセット $\{(\mathrm{cwe}(w_i; m), l_i)\}_{i \in [1,N]}$ を構築した．ただし，N はコーパス中の総単語出現数，$\mathrm{cwe}(w_i; m)$ は PLM m が生み出した単語 w_i の CWE である．w_i が複数の wordpiece に分割された時には，それらの CWE の平均を $\mathrm{cwe}(w_i; m)$ とする．対象モデルは BERT-base-uncased，BERT-base-cased と RoBERTa-base の 3 つである．CWE は各モデルの隠れ層の出力の平均と

する．

訓練セットをランダム抽出により生成すると，高頻度の単語を含む実例を多く含んでしまう．探査子自体もひとつのモデルなので，訓練セットに頻出のパターンをよりよく捉えられると予想されるが，これは頻度と CWE の関係を探求する上で好ましくない．したがって，本研究では訓練セットを抽出する際に，2 つの抽出方法をとった．1 つは通常のランダム抽出で，もう 1 つは層化抽出である．層化抽出ではデータセット $\{(\mathrm{cwe}(w_i;m), l_i)\}$ を w_i のレンマの頻度により 10 個の部分集合に分割してから，それぞれから固定の割合でランダム抽出する．いずれの抽出法でも，データセットは 6 割を訓練セットに，4 割をテストセットにした．異なる調査対象モデルと探査子用の訓練セットを抽出する際にはランダムシードを固定した．

3.3 評価方法

RQ2, 3 にアプローチするために, 文脈化利得 (Contextualization Gains, CN) と正規化利得 (Normalzation Gains, NG) を定義する．

CG は RQ2 のために用意した．CG を計算するため，データセット中の事例 $(\mathrm{cwe}(w_i;m), l_i)$ の $\mathrm{cwe}(w_i;m)$ を，単語 w_i の同じモデルの埋め込み層の出力 $\mathrm{static}(w_i;m)$ に置き換えて統制分類器を学習する．CG は探査子と統制分類器の性能の差として定義するので，文脈化が有効であれば CG が大きくなる．

NG は RQ3 のために用意した．Su らの bert-whitening [8] に基づき，NG では下式のとおり訓練データを書き換えて統制分類器を学習する．

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{c}_j \\ \boldsymbol{\Sigma} &= \frac{1}{N}(\boldsymbol{C} - \mathbf{1}\boldsymbol{\mu}^T)^T(\boldsymbol{C} - \mathbf{1}\boldsymbol{\mu}^T) \\ norm(\boldsymbol{c}_j;\boldsymbol{\mu},\boldsymbol{\Sigma}^{-\frac{1}{2}}) &= (\boldsymbol{c}_j - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-\frac{1}{2}} \end{aligned} \tag{2}$$

ただし, $\boldsymbol{c}_j \in \mathbb{R}^d$ は d 次元の CWE で, $\boldsymbol{C} = [\boldsymbol{c}_1, ..., \boldsymbol{c}_N]^T$ でサイズ N の訓練セット中の CWE による行列である．$\mathbf{1} \in \mathbb{R}^d$ は各要素が 1 の d 次元のベクトルである．$\boldsymbol{\Sigma}$ は実に $\boldsymbol{C}$ の各次元についてのサンプル共分散行列である．$norm$ のパラメータは訓練セットで計算する．この書き換え方は分散を共分散に置き換えた Z-socre と見なせる．訓練セットの CWE による行列について正規化をして CWE の異方性を削減する．NG は統制分類器からの探査子の性能の差分であるが，探査子の性能について引き算の向きを注意されたい．もし異方性が CWE の語義と関わるのなら，NG の絶対値は大きいと予想される．

RQ4 に答えるために，各性能指標と 2 つの利得を batch_{ef} で計算する．batch_{ef} はテストセットの部分集合で，同バッチ内の事例のエントロピーと頻度も同等である．e と f はそれぞれの語義分布のエントロピーと頻度のランクである[1)]．エントロピー e を固定して，f が大きくなるに連れて，各 batch_{ef} で計算された指標の値の変化を見れば，RQ4 に答えることができる．

4 実験結果

4.1 概要

表 2 層化抽出法の場合の結果一覧

指標	Static	CG	Static+CG	NG	Static+CG+NG
syp-acc	54.9	+8.9	63.8	+6.3	70.1
hyp-pre	**82.9**	-6.9	76.0	+15.4	**91.4**
hyp-rec	40.8	**+23.8**	64.6	**-24.1**	40.5
sup-acc	72.1	+4.6	**76.7**	+3.2	79.9

表 2 に，BERT-base-uncased による CWE を用いた，層化抽出法の場合の実験結果をまとめた[2)]．CG と NG の列は異なる性能指標における文脈化利得と正規化利得である．Static，Static+CG と Static+CG+NG の列は，それぞれ非文脈化単語埋め込み，CWE，正規化後の CWE での探査子の性能である．太字は該当列の絶対値としての最大値を意味する．全体的に CWE での性能は hyp-pre と sup-acc が大きく，NG は絶対値として hyp の全ての評価指標に比較的に大きく，CG は hyp-rec だけ大きい．Static+CG+NG と Static を比較すれば，hyp-rec と sup-acc だけ増加しなかったことがわかる．

syp-acc と sup-acc では，CG と NG は両方ともに正であることから，BERT 系 PLM の非文脈化単語埋め込みは，文脈化と正規化によって抽象度の違いによらず語義を単体で捉える能力が向上する．抽象的語義の場合，向上の幅はより大きい．しかし，hyp-rec は文脈化で増加して正規化で減少する傾向を示す．

1) 事例 $(\mathrm{cwe}(w_i;m), l_i)$ について，w_i の語義分布のエントロピーが 0.04 以下，つまり実質上の一義語ならば，エントロピーランクを 0 とする．エントロピーランク 1 から 4 の分割は，データセットの単語のエントロピー分布の四分位点による．

2) 抽出法や PLM のモデルが結果に影響をしない．すべての抽出法と PLM の結果は付録の表 3 を参照されたい

つまり，CWE は文脈化でマルチレベルの抽象度の意味を同時に捉えるようになり，正規化したらそれができなくなる．

一方，hyp-pre は hyp-rec と逆の傾向である．hyp-pre は低くなるのは，hypernym 階層において根までの経路上にない無関係の語義を多く誤同定してしまった場合である．ここでいう無関係な語義は，そもそも真の語義とは完全に別種の語義と，ある抽象度では同義であるが細部が違う語義の 2 種類を含む．これと hyp-pre の傾向を踏まえて，CWE 中の無関係な語義は文脈化の前で少なく，文脈化後で増えるが，正規化で取捨選択されると考えられる．このことから，CWE の異方性は語義の網羅性と関係することを示唆している．つまり，ある単語の BERT 系 CWE は周囲の文脈から語義を豊かにするが，同時に語義の余剰も生み出す．これ過程の結果は CWE の異方性に反映される．

4.2 詳細

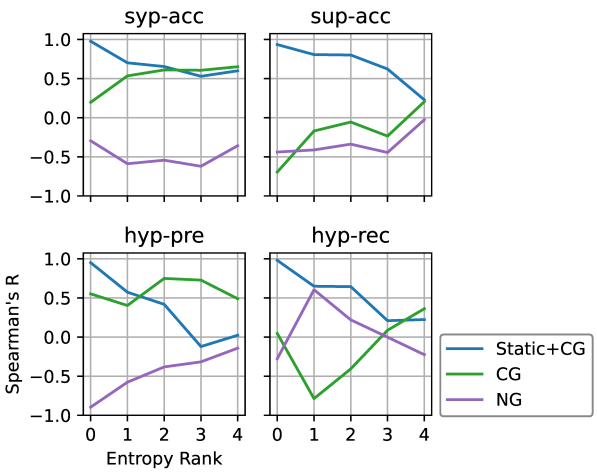

図 1 各評価指標の単語頻度との相関 (Spearman の R)

図 1 は，エントロピーランク e を固定し，batch$_{ef}$ で計算した各評価指標と頻度ランク f とのスピアマン相関係数を示す．抽出方法は結果に影響しないので，ここでは層化抽出の結果だけを示す．大多数の場合，エントロピーの増加が Static+CG，CG と NG と頻度との相関程度を軽減するとみられる．

Static+CG 各指標において，Static+CG と頻度との相関係数はほとんど非負であるが，値がエントロピーの増加につれて減少する．つまり，BERT 系 CWE は，抽象度と関係なく，低エントロピーで高頻度の単語ほど意味をより捉えている．

CG CG は syp-acc で値も頻度との相関係数も正で，syp-acc で値が正で相関係数が負である．これは CWE は文脈から，低頻度語ほど抽象的な語義を，頻出語ほど具体的語義の情報を有効に獲得している．つまり，PLM は低頻度語の大雑把な意味を文脈から推測し，高頻度語ならよりきめ細かい意味も取りこんでいると言える．hyp-pre の結果から，文脈化は CWE に意味の余剰をもたらすが，CG が頻度と相関係数が正なので，頻出語ほどこの余剰が少ない[3]．

NG NG は sup-acc と syp-acc とも値が正で相関係数が負である．これは CWE の異方性を正すことによる利得は，粒度と関係なく低頻度語ほど大きい．つまり，低頻度語の CWE の異方性には，無関係な語義と関連すると思われる．NG は hyp-rec で値であるが相関係数の絶対値がエントロピーランク 1 以外に小さい．実際，正規化は CWE の文脈化による意味の網羅性を軽減するが，この効果は頻度と正の二次関数的な関係をなしており，エントロピーが小さいほどその傾向が顕著である[4]．その極値は比較的低い頻度ランクに出現する．つまり，比較的に低エントロピーで低頻度の単語の CWE の異方性がなくなったら，意味の網羅性が小さくなる．

5 結論

以上より，本研究の RQ1 から RQ4 に即して結論を述べる．（RQ1）CWE は異なる粒度を捉える能力があるが，（RQ2&3）文脈化前後でも，さらに異方性を修正するかどうかによっても，その能力が変化する．文脈化は語義を豊富に埋め込めるが，同時に無関係の語義も埋め込んでしまい，語義の余剰を招く．しかし，余剰語義は CWE の異方性と関係し，正則化で解消可能である．（RQ4）1)BERT 系 CWE の意味と単語頻度および語義分布との関係について，以下 3 点が確認された．

1. BERT 系 CWE は意味の粒度と関係ぜずに，高頻度で低エントロピーの単語ほど意味をより捉えられる
2. 高頻度語ほど具体的な意味を正確に文脈から取り込み，語義の余剰が少ない
3. 比較的に低エントロピー語の BERT 系 CWE は正則化による余剰語義の解消は頻度と相関し，その中でも比較的低頻度の単語は解消程度が大きい．

3) 実際，高エントロピー・高頻度の単語に対して，CG が正になることが見られる．詳細は付録の図 2 を参照されたい．
4) その詳細は付録の図 3 を参照されたい．

参考文献

[1] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In **Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993**, 1993.

[2] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z Ghahramani, M Welling, C Cortes, N Lawrence, and K Q Weinberger, editors, **Advances in Neural Information Processing Systems**, 2014.

[3] Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: Frequency-agnostic word representation. 2018.

[4] Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. Factors influencing the surprising instability of word embeddings. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2092–2102, 2018.

[5] Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. Word frequency does not predict grammatical knowledge in language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4040–4054, 2020.

[6] Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. Frequency-based distortions in contextualized word embeddings. 2021.

[7] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7222–7240, 2020.

[8] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. 2021.

[9] Christiane Fellbaum, editor. **WordNet: An Electronic Lexical Database**. The MIT Press, 1998.

[10] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. 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)**, pp. 2733–2743, 2019.

A 付録 (Appendix)

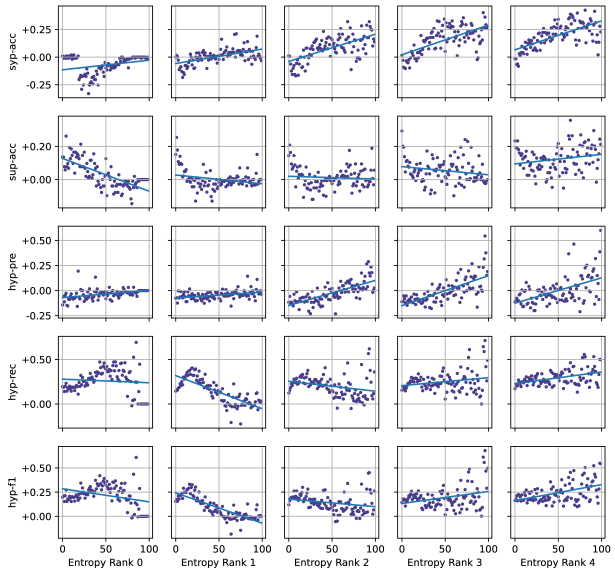

図 2 CG の詳細/Layered Sampling/BERT-base-uncased

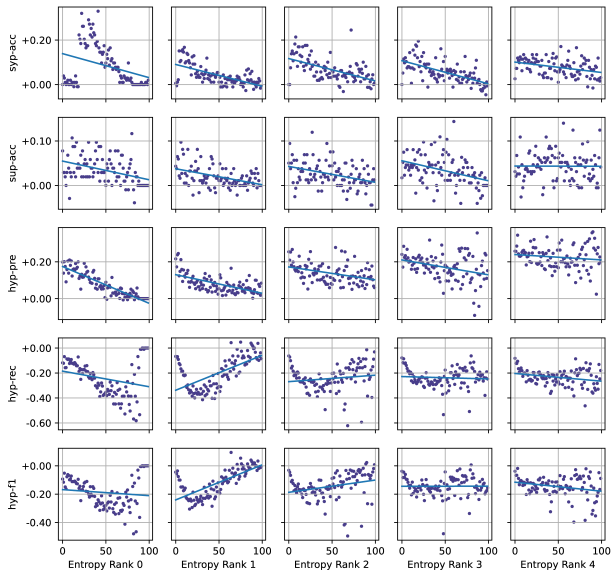

図 3 NG の詳細/Layered Sampling/BERT-base-uncased

図 2 と図 3 はそれぞれ，各評価指標とエントロピーランクにおける CG と NG の詳細な結果を示している．たとえば，図 2 の syp-acc の列の一番右にあるサブグラフは，エントロピーランク 4 の場合の，Synset 探査子の精度について計算した batch_{ef} の文脈化利得と f とのプロット図である．サブグラフにある青い線は，batch_{ef} の文脈化利得と f との回帰直線である．

表 4.1 はすべての PLM の探査結果である．hyp-F1 の列は順に，Hypernyms 探査タスクの F1 尺度の値とその文脈化利得・正規化利得である．

表 3 すべての PLM の結果一覧

		random			layered		
PLM	指標	Static	CG	NG	Static	CG	NG
BERT-base-uncased	syp-acc	54.9	+8.9	+6.3	55.1	+8.8	+6.4
	hyp-pre	82.9	-6.9	+15.3	82.4	-5.6	+14.8
	hyp-rec	40.8	+23.8	-24.1	41.0	+22.6	-23.3
	hyp-F1	54.7	+15.1	-13.7	54.7	+14.9	-13.6
	sup-acc	72.1	+4.6	+3.2	72.0	+4.7	+3.1
BERT-base-cased	syp-acc	54.3	+9.1	+6.5	54.5	+9.2	+6.3
	hyp-pre	81.8	-6.3	+16.0	81.7	-6.4	+16.2
	hyp-rec	42.6	+21.9	-24.4	42.5	+21.8	-24.1
	hyp-F1	56.0	+13.6	-13.8	55.9	+13.4	-13.4
	sup-acc	71.6	+4.8	+3.4	71.5	+4.7	+3.5
RoBERTa-base	syp-acc	56.2	+7.4	+5.9	56.0	+6.9	+6.7
	hyp-pre	85.8	-7.0	+12.7	85.0	-5.9	+12.4
	hyp-rec	33.0	+28.7	-21.5	33.4	+27.5	-20.9
	hyp-F1	47.7	+21.5	-13.3	48.0	+20.8	-13.1
	sup-acc	72.5	+4.7	+2.6	72.3	+4.9	+2.4