Vol.2017-NL-232 No.8
2017/7/20

アラビア語の高粒度な品詞タグ付けのための辞書情報を活用した形態統語的カテゴリの同時予測

井上 剛[1,a)]　進藤 裕之[1,b)]　松本 裕治[1,c)]

概要：アラビア語などの形態的に豊かな言語の品詞タグ付けは，英語など形態的に乏しい言語の品詞タグ付けに比べ，タグセットが膨大になるため，困難な問題である．これは，言語固有の情報を反映した高粒度な品詞タグが，各形態統語的カテゴリごとに定義されたタグの組み合わせによって構成されるためである．既存のアラビア語品詞タグ付けでは，各形態統語的カテゴリを独立に予測しており，各カテゴリを予測する上で有益な情報をカテゴリ間で共有できていなかった．本研究では，マルチタスク学習の枠組みを用いて，各形態統語的カテゴリを同時に予測する手法を提案する．また，入力語に対して各形態統語的カテゴリが取りうるタグを登録した辞書情報をモデルに組み込むことで，さらなる性能向上が得られることを示す．Penn Arabic Treebank を用いた評価実験の結果，これまでに報告されている最高性能の品詞タガーの正解率を上回ることを確認した．

1. はじめに

品詞タグ付けは，自然言語処理において基本的な言語解析タスクのひとつである．言語固有の情報を反映した高粒度な品詞タグセットのサイズは，言語ごとに異なる．例えば，英語などの形態的に乏しい言語では，品詞タグセットのサイズは典型的に 100 以下である一方，アラビア語などの形態的に豊かな言語では，理論上可能なタグが 333,000 種類に上り，そのうち約 2,200 種類のみが実際のコーパスに出現する [9]．タグセットが膨大になる理由として，アラビア語などの形態的に豊かな言語における品詞タグは，形態統語的カテゴリごとに定義されたタグの組み合わせによって構成される点が挙げられる．例えば，*Hb*（“love”）という語[*1]に対する品詞タグは，粗い品詞カテゴリが「名詞」，格カテゴリが「主格」，法カテゴリが「非該当」などのように，各カテゴリごとの値を組み合わせた形として定義される．結果として，膨大な数のタグ候補が存在するため，アラビア語などの形態的に豊かな言語に対する高粒度な品詞タグ付けを困難にする[*2]．

このようなタグセットの枠組みのもとで品詞タグ付けを行うには，ある形態統語的カテゴリのタグを予測する際に，他のカテゴリからの情報を活用することが有益である．例えば，入力語の粗い品詞カテゴリが名詞であった場合，格カテゴリは主格，属格，対格のいずれかを取る一方，法カテゴリは非該当を取らなければならない[*3]．他方，格カテゴリが主格，属格，対格のいずれかを取っている場合，粗い品詞カテゴリは名詞類のタグのいずれかで，法カテゴリは非該当を取らなければならない．

アラビア語の高粒度な品詞タグ付けに関する既存研究では，各形態統語的カテゴリを独立に予測しており，このような情報は十分に活用されてこなかった [9], [22], [24]．そこで本研究では，マルチタスク学習の枠組みを用いて，各形態統語的カテゴリを予測するタスクを同時にモデル化するアプローチを提案する．また，さらなる性能向上のため，入力語に対して各形態統語的カテゴリが取りうるタグを登録した辞書情報をモデルに組み込む手法を提案する．Penn Arabic Treebank を用いた評価実験の結果，これまでに報告されている最高性能の品詞タガー [24] の正解率を上回ることを確認した．

本研究の提案手法を実装したコードは，著者のページにて公開する[*4]．

2. 高粒度な品詞タグ付けの定式化

本研究で取り組むタスクは，入力文の語系列に対し，言

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] inoue.go.ib4@is.naist.jp
[b)] shindo@is.naist.jp
[c)] matsu@is.naist.jp

[*1] 本稿では，アラビア文字を Buckwalter 転記法 [2] に則って転記する．
[*2] アラビア語固有の諸問題については，Habash[8] を参照されたい．
[*3] 名詞類には法という文法範疇が定義されないため．
[*4] https://github.com/go-inoue/FineGrainedArabicPOSTagger

語固有の情報を保持した高粒度な品詞タグを付与することである．品詞タグ付けは，n 語の語系列 $x_{1:n}$ を入力とし，各語に対応するラベル系列 $y_{1:n}$ を出力とする．ここで，x_t は文中における t 番目の語，$y_t \in T$ は，x_t に対するタグとする．英語などの形態的に乏しい言語では，品詞タグがひとつのタグセット T から取られる．対して，アラビア語などの形態的に豊かな言語では，品詞タグが形態統語的カテゴリごとに定義されたタグの組み合わせによって構成される．形式的には，入力語 x_t に対する高粒度な品詞タグ y_t^{fine} は，k 個のタグセット $T^{(1)}, T^{(2)}, ..., T^{(k)}$ から得られるタグの結合 $y_t^{(1)} \wedge y_t^{(2)} \wedge ... \wedge y_t^{(k)}$ として定義される．われわれの目的は，入力語 x_t についてすべての形態統語的カテゴリを予測することであり，これはマルチクラスかつマルチラベルの系列ラベリング問題[*5]として捉えることができる．

本稿では，Pasha ら [22] で用いられている 14 種類の形態統語的カテゴリ[*6]を用いる．この枠組みは，アラビア語の自然言語処理ツールで広く用いられている [11], [22], [24]．

3. 高粒度な品詞タグ付けモデル

本節では，アラビア語の高粒度な品詞タグ付けのための双方向 LSTM リカレントニューラルネットワークを用いたモデルについて説明する．はじめに，ベースラインとして各形態統語的カテゴリを独立に予測するモデルについて述べる．次に，提案手法として各カテゴリを同時に予測するモデルについて述べる．また，入力語に対して各カテゴリが取りうるタグを登録した辞書情報をモデルに組み込む手法について説明する．

3.1 独立予測モデル

本研究ではベースラインとして，双方向 LSTM リカレントニューラルネットワーク [6] を用いて，各形態統語的カテゴリを独立に予測するモデルを用いる．このモデルは，Plank ら [23] による双方向 LSTM モデルに基づいている．図 1 に，格カテゴリを予測するネットワーク構造の概要を示す．n 語からなる入力語系列 $x_{1:n}$ を， 単語ベクトル $\mathbf{w}_t$ と語を構成する文字列の特徴ベクトル $\mathbf{c}_t$ を連結した特徴ベクトル $\mathbf{r}_t = [\mathbf{w}_t; \mathbf{c}_t]$ へと変換する．文字列の特徴ベクトルは，文字レベルの順方向 LSTM の隠れ状態ベクトルと逆方向の隠れ状態ベクトルを連結することで計算される．図 2 に，文字列の特徴ベクトルを計算するネットワーク構造の概要を示す．

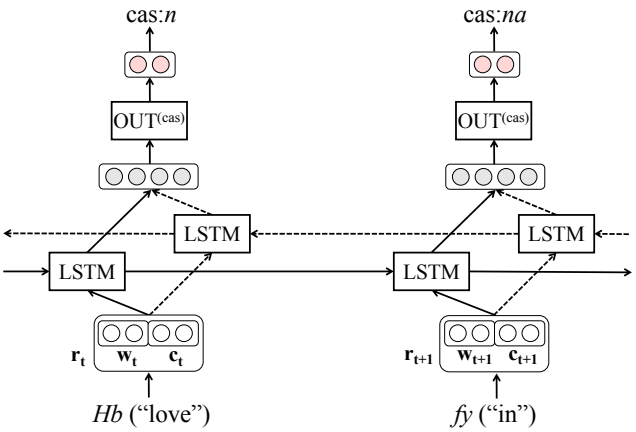

図 1 格カテゴリを予測する独立予測モデルのネットワーク構造．

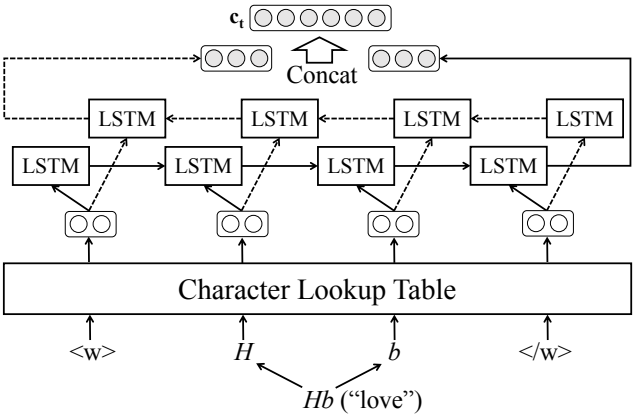

図 2 文字列の特徴ベクトルの計算例．<w> は語頭を </w> は語末を意味する．

次に，入力となる特徴ベクトル $\mathbf{r}_t$ に対して，順方向の LSTM による隠れ状態ベクトル $\overrightarrow{\mathbf{h}}_t$ と逆方向の LSTM による隠れ状態ベクトル $\overleftarrow{\mathbf{h}}_t$ を求める．これらの隠れ状態ベクトルを連結することで，双方向の LSTM による隠れ状態ベクトル $\mathbf{v}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$ を得る．このベクトルを出力層へ入力し，最終的に，出力層に対してソフトマックス関数を適用することで出力ラベル y_t を得る．独立予測モデルでは，予測する形態統語的カテゴリごとにモデルを学習するため，形態統語的カテゴリが 14 種類ある場合，合計 14 個のモデルが得られる．

3.2 同時予測モデル

独立予測モデルは，各形態統語的カテゴリを予測するタスクを独立にモデル化しているため，各タスクに共通する有益な情報が共有されていない．しかし，各カテゴリ間には依存関係があるため，あるカテゴリのタグを予測する際，他のカテゴリからの情報を活用することができれば，解析性能の向上が期待できる．そこで本研究では，マルチタスク学習の枠組み [1], [3], [18], [25], [26] を用いてこれを実現する．具体的には，双方向 LSTM モデルの隠れ層においてパラメータを共有することで，各タスクに有益な情報を保

[*5] 英語などの形態的に乏しい言語に対する品詞タグ付けは，各語に対するラベルが 1 つであるため，この問題の特別な形式として捉えることができる．

[*6] 14 種類のカテゴリは次の通り：粗い品詞 (pos)，性 (gen)，数 (num)，格 (cas)，法 (mod)，アスペクト (asp)，人称 (per)，態 (vox)，state (stt)，4 種類の前接語 (prc0, prc1, prc2, prc3)，後接語 (enc)．

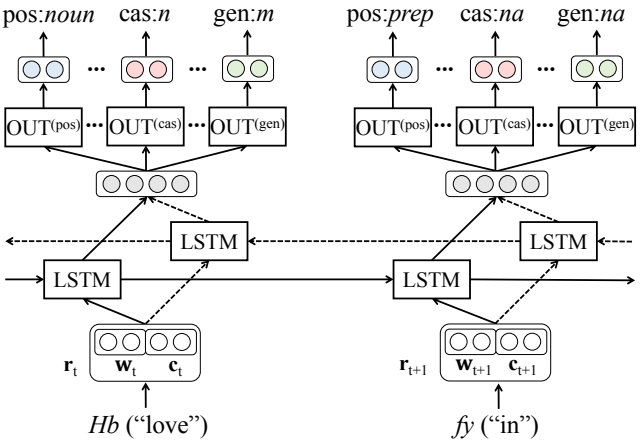

図 3　同時予測モデルのネットワーク構造.

持する統一的なモデルを得る．

図 3 に同時予測モデルの概要を示す．独立予測モデルとの違いは，出力層が予測するカテゴリごとにあるという点である．すなわち，双方向の LSTM による隠れ状態ベクトルを各出力層に入力し，それらの出力に対してソフトマックス関数を適用することで，カテゴリごとの出力ラベルを得る．モデルの学習は，入力文の各語における各カテゴリの予測に対する交差エントロピー誤差を計算し，それらの平均を取った上で，入力文全体における和を最小化するようにパラメータの最適化を行う．ここで，各入力語に対する損失関数を次のように定義する．

$$L(\hat{y}^{fine}, y^{fine}) = \frac{1}{|M|} \sum_{m \in M} L(\hat{y}_m, y_m)$$

ただし，$M = \{pos, cas, gen, ...\}$ は形態統語的カテゴリを予測するタスクの集合，$L(\hat{y}_m, y_m)$ はカテゴリ m に対する交差エントロピー誤差である．

3.3　辞書情報のモデルへの組み込み

タグの辞書情報を用いた既存研究 [9], [22], [24] では，辞書の出力結果から正解となる高粒度なタグを選択するという点で，辞書情報を強い制約（hard constraints）として利用している．既存手法の欠点は，辞書に入力語のエントリが登録されていなかった場合，正解が得られないという点にある．実際に，Habash ら [10] が行ったエラー分析では，解析誤りの 31.3%が前述の原因によると報告している．

この問題に対処するため，本研究では，入力語に対して各形態統語的カテゴリが取りうるタグを登録した辞書情報を特徴ベクトルとしてモデルに組み込むで，辞書情報を弱い制約（soft constraints）として用いる手法を提案する．また既存手法では，辞書を構築する際のタグセットと出力となるタグセットは同一でなければばならなかったが，提案手法では辞書情報をモデルへ素性として入力するため，出力となるタグセットは任意に設定できるという利点がある．

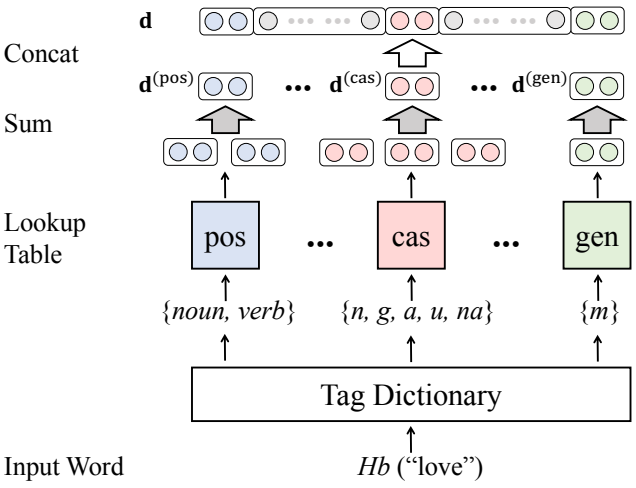

図 4　辞書情報の特徴ベクトルの計算例.

図 4 に，*Hb* ("love") という語に対する辞書情報の特徴ベクトルの計算法を示す．まず，入力語を辞書に与え，形態統語的カテゴリごとに取りうるタグの集合を生成する．これらのタグに対してベクトルを用意し，カテゴリごとにベクトルの和を取る．これらのベクトルは，各カテゴリについてその語が取りうるタグの可能性を表現している．最終的に，これらのベクトルをすべて連結し，辞書情報の特徴ベクトルとする．

形式的には，入力語 x_t に対する特徴ベクトル $\mathbf{d}_t$ は，各カテゴリ m に定義されるサブベクトルすべてを連結したベクトルである．

$$\mathbf{d}_t = [\mathbf{d}_t^{(pos)}; ... ; \mathbf{d}_t^{(cas)}; ... ; \mathbf{d}_t^{(gen)}]$$

サブベクトル $\mathbf{d}_t^{(m)}$ は，次の式で計算される．

$$\mathbf{d}_t^{(m)} = \sum_{d \in D_t^{(m)}} \mathbf{W}^{(m)} \mathbf{e}_d^{(m)}$$

ここで，$D_t^{(m)}$ は入力語 x_t が取りうるタグ集合，$\mathbf{W}^{(m)}$ はカテゴリ m に対する埋め込み行列，$\mathbf{e}_d^{(m)}$ はカテゴリ m に対するタグ d の one-hot 表現を表す．最終的にモデルへ入力される特徴ベクトル $\mathbf{r}_t = [\mathbf{w}_t; \mathbf{c}_t; \mathbf{d}_t]$ は，単語ベクトル $\mathbf{w}_t$，文字列の特徴ベクトル $\mathbf{c}_t$，辞書情報の特徴ベクトル $\mathbf{d}_t$ を連結したベクトルである．図 5 に辞書情報を組み込んだ同時予測モデルの概要を示す．

4.　実験

4.1　実験設定

PATB データセット

提案手法の有効性を先行研究と同等の条件で比較検証するため，Penn Arabic Treebank (PATB, parts 1, 2 and 3) [14], [15], [16] を用いて実験を行う．表 1 にデータセットの詳細を示す．訓練，開発，検証データの分割は，Diab ら [4] に従っている．データは Pasha ら [22] と同様の手法で，アノテーションの不一致を取り除く前処理を行う．ま

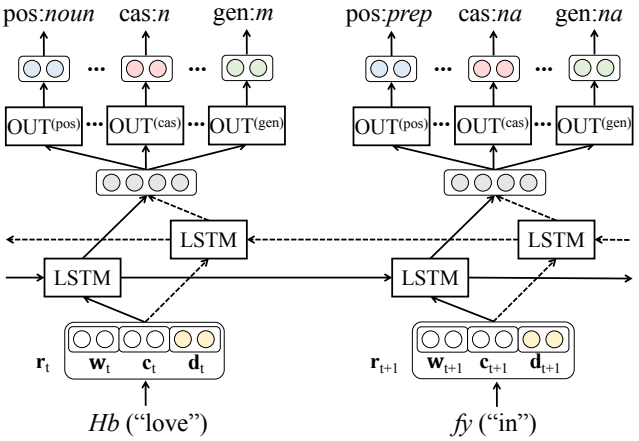

図 5 辞書情報を組み込んだ同時予測モデルのネットワーク構造.

た，数字は 0 に置換する．

比較対象には，これまで報告されているなかで最高性能のタガー CamelParser[24] を用いる．CamelParser は，入力語が取りうるタグが登録された辞書の出力に対し，形態統語的カテゴリごとに学習された SVM 分類器を用いてランキングを行った上で，依存構造解析器の出力結果と人手の規則から得た統語情報を用いてリランキングを行う．CamelParser で用いられているタグセットは，本研究で用いるタグセットと互換性がある．評価は，14 種類の形態統語的カテゴリとそれらの組み合わせ（高粒度なタグ）に対して行う．

	訓練	開発	検証
文数	15789	1986	1963
語数	502991	63136	63168
高粒度なタグ数	2028	1034	1069

表 1 PATB データセットにおける文数，語数，高粒度なタグ数.

UD Arabic データセット

提案手法の頑健性を検証するため，PATB データセットとは異なるデータ，異なるタグセットが用いられているアラビア語版の Universal Dependencies[*7] Version 1.4 を用いて実験を行う．表 2 にデータセットの詳細を示す．なお，UD Arabic データセットにおける入力単位は，接語分割がなされたトークンとする．前処理として，数字は 0 に置換する．

性能比較は，辞書情報の有無それぞれについて，独立予測モデルと同時予測モデル間で行う．評価は，17 種類の形態統語的カテゴリ[*8]とそれらの組み合わせ（高粒度なタグ）に対して行う．

[*7] http://universaldependencies.org/
[*8] UD Arabic データセットで定義されている universal POS タグと 16 種類の morphological feature タグを用いる．

	訓練	開発	検証
文数	6174	786	704
トークン数	225853	28263	28268
高粒度なタグ数	327	214	213

表 2 UD Arabic データセットにおける文数，トークン数，高粒度なタグ数.

4.2 実装詳細

提案モデルの実装には，深層学習ライブラリ DyNet[21] を用いる．独立予測モデルと同時予測モデルには，同じハイパーパラメータを適用する．すなわち，単語ベクトルの次元数は 100 次元，文字列の特徴ベクトルは 50 次元，形態統語的カテゴリごとの辞書情報の特徴ベクトルは 10 次元，LSTM の隠れ状態ベクトルは 500 次元，出力層は 100 次元とする．各埋込みベクトルは，既存研究と同等のデータ環境での実験を実現するため，外部資源を用いた事前学習は行わず，ランダムに初期化する．パラメータの最適化は文単位で行い，Adam[12] を用いる．学習は 10 エポックまで行い，開発データでの正解率が最大となるエポックにおけるモデルを選択する．辞書には，MADAMIRA[22] から入手可能な ALMOR[7] のデータに SAMA[17] のデータを追加した辞書を用いる．

4.3 結果

PATB データセット：提案モデル vs CamelParser

表 3 は PATB データセットに対するタグ付けの正解率を示している．表 3 に示されているように，辞書情報を用いた同時予測モデルが最も高い解析性能を達成していることがわかる．高粒度なタグ（**All**）による評価では，CamelParser の解析性能を 2.11 ポイント上回り，正解率 91.38%を記録している．このモデルは，すべての形態統語的カテゴリの予測性能において CamelParser を上回っており，提案手法の有効性を示している．独立予測モデルと同時予測モデルを比較すると，同時予測モデルが高い解析性能を記録している．これは，各形態統語的カテゴリを同時にモデル化することで，カテゴリ間の依存関係を考慮できるようになったため，性能向上につながったと考えられる．また，辞書埋め込みベクトルの有無を比較すると，独立予測モデルと同時予測モデル共に解析性能が向上している．これは，辞書埋め込みベクトルが入力語の形態統語的振る舞いについての有益な特徴量を学習できており，タグの候補を生成する辞書さえあれば，容易に解析性能を向上させられることを示している．

UD Arabic データセット

表 4 は UD Arabic データセットに対するタグ付けの正解率を示している．PATB データセット同様に，辞書情報を用いた同時予測モデルが最も高い解析性能を達成してい

	pos	gen	num	cas	mod	asp	per	vox	stt	prc0	prc1	prc2	prc3	enc	**All**
CamelParser	96.78	99.41	99.43	92.68	99.13	99.27	99.23	99.08	97.54	99.67	99.63	99.59	99.90	99.61	89.27
Independent	96.31	99.05	99.26	93.17	99.07	99.08	99.10	98.80	97.23	99.62	99.64	99.73	**99.97**	99.44	87.74
+Dict	97.07	99.33	99.51	94.70	99.31	99.34	99.35	99.18	98.11	99.48	99.78	**99.78**	**99.97**	99.68	90.17
Joint	96.24	99.27	99.16	93.48	99.18	99.19	99.20	98.91	97.70	99.66	99.64	99.68	**99.97**	99.58	89.49
+Dict	**97.21**	**99.50**	**99.59**	**94.76**	**99.41**	**99.44**	**99.47**	**99.25**	**98.24**	**99.71**	**99.81**	99.73	99.96	**99.71**	**91.38**

表 3 PATB データセットに対する正解率.

	POS	Gender	Number	Case	Mood	Aspect	Person	Voice	Definite
Independent	95.15	97.28	96.38	93.76	99.56	99.35	99.37	99.14	96.40
+Dict	96.08	98.06	97.23	94.86	99.68	99.51	99.47	99.16	97.09
Joint	95.92	97.96	96.69	94.60	99.67	99.50	99.45	99.21	96.67
+Dict	**96.64**	**98.32**	**97.47**	**95.43**	**99.69**	**99.58**	**99.59**	**99.32**	**97.35**

	Abbr	AdpType	Foreign	Negative	NumForm	NumValue	PronType	VerbForm	**All**
Independent	99.88	99.75	99.16	99.99	99.88	99.80	99.76	99.69	86.45
+Dict	**100.00**	99.84	99.58	99.99	**99.90**	99.80	99.79	99.73	89.17
Joint	99.99	99.85	99.47	99.99	**99.90**	**99.98**	99.81	99.78	90.36
+Dict	99.99	**99.86**	**99.66**	99.99	99.89	**99.98**	**99.84**	**99.84**	**91.68**

表 4 UD Arabic データセットに対する正解率.

る．また，独立予測モデルと同時予測モデルを比較すると，同時予測モデルが高い解析性能を記録している．このことから，提案手法は，異なるデータセットに対しても頑健に性能向上をもたらしていることがわかる．また，辞書埋め込みベクトルの有無を比較すると，独立予測モデルと同時予測モデル共に，辞書情報を使用した場合の解析性能が向上している．このことから，出力のタグセットが異なる場合であっても，辞書情報をモデルへ素性として入力することで，解析性能が向上することがわかる．

4.4 辞書情報の特徴ベクトルにおける各形態統語的カテゴリの影響度

辞書情報の特徴ベクトルは，形態統語的カテゴリごとに定義されたサブベクトルすべてを連結した特徴ベクトルである．このうち，どのカテゴリがどれほど解析性能に影響を与えているかを調査するために，これらのサブベクトルのうち 1 つだけを特徴ベクトルとして追加する実験を PATB データセットに対して行った．すなわち，辞書情報の特徴ベクトルを計算する際に，各カテゴリのサブベクトルを連結せずに，あるカテゴリに対する辞書情報の特徴ベクトルのみを特徴ベクトルとして追加した．これによって，あるカテゴリについての辞書情報が，予測する各カテゴリの解析性能に対してどれほど影響を与えるのかがわかる．

表 5 は，各形態統語的カテゴリの辞書情報を用いた同時予測モデルの解析性能を示している．高粒度なタグによる評価では，粗い品詞（+pos）カテゴリの辞書情報が解析性能に最も貢献していることがわかる．これに続いて，格カテゴリ（+cas）と state カテゴリ[*9]（+stt）が同率で解析性能に貢献している．このことから，数や性などの語内で完結する形態的なカテゴリに比べて，語の外部との関係を表す統語的なカテゴリについての辞書情報が，解析性能の向上により貢献していることがわかる．

予測する各カテゴリへの影響を個別にみると，5 つのカテゴリ（pos，gen，cas，mod，vox）において，粗い品詞（+pos）カテゴリの辞書情報が性能向上に最も貢献している．このことから，粗い品詞カテゴリの辞書情報が，その他の形態統語的カテゴリを予測する上で，中心的な役割を果たしていることがわかる．

一方，8 つのカテゴリ（pos，num，per，stt，prc0，prc1，prc2，enc）において，予測するカテゴリとモデルが用いた辞書情報のカテゴリが同じ場合に，解析性能が最も向上している．これは，あるカテゴリに対する辞書情報が，そのカテゴリを予測する上で弱い制約として機能していることを示している．

5. 関連研究

Diab ら [5] は，接語分割を行った上で各トークンに対し 24 種類の品詞タグを付与する SVM を用いた手法を提案している．Mohamed ら [19] は，接語分割は行わずに，各入力語に対して 993 種類の品詞タグを付与するメモリーベース学習を用いた手法を提案している．Zhang ら [27] は，接語分割，12 種類からなる品詞タグ付け，依存構造解析を同時にモデル化する手法を提案している．これらの研究では，粒度の低いタグセット[*10]を用いた品詞タグ付けを行っ

[*9] 名詞類について限定，非限定，連結が定義されている．連結とは，属格の名詞を従えて名詞句を構成することを指す．

[*10] Mohamed らのタグセットは，Diab らや Zhang らに比べ粒度が高いが，語の統語関係を示す主要なカテゴリである格についてのタグが除外されている．また 993 種類のタグをタグに分解せずに 1 つのタグセットとして扱っているため，訓練データに現れていないタグは予測できないという問題がある．

	pos	gen	num	cas	mod	asp	per	vox	stt	prc0	prc1	prc2	prc3	enc	**All**
Joint	96.24	99.27	99.16	93.48	99.18	99.19	99.20	98.91	97.70	99.66	99.64	99.68	99.97	99.58	89.49
+pos	**+0.96**	**+0.25**	+0.27	**+1.00**	**+0.25**	+0.21	+0.23	**+0.38**	+0.46	+0.04	+0.09	+0.09	0.00	+0.08	**+1.48**
+gen	+0.35	+0.10	+0.18	+0.34	+0.12	+0.12	+0.09	+0.21	+0.19	0.00	-0.06	0.00	-0.01	+0.02	+0.33
+num	+0.36	+0.10	**+0.43**	+0.45	+0.06	+0.07	+0.08	+0.17	+0.13	+0.03	-0.02	+0.02	-0.01	+0.01	+0.63
+cas	+0.51	+0.13	+0.25	+0.82	**+0.25**	+0.22	+0.23	+0.32	+0.41	-0.01	+0.08	+0.04	0.00	+0.06	+0.99
+mod	+0.38	+0.10	+0.14	+0.77	+0.23	+0.23	+0.21	+0.31	+0.39	-0.01	+0.04	+0.05	-0.01	+0.06	+0.82
+asp	+0.47	+0.12	+0.22	+0.48	+0.22	+0.22	+0.24	+0.33	+0.33	+0.02	+0.06	+0.03	0.00	+0.03	+0.68
+per	+0.26	+0.16	+0.18	+0.72	+0.24	**+0.28**	**+0.29**	+0.36	+0.32	+0.01	+0.08	+0.06	0.00	+0.07	+0.78
+vox	+0.27	+0.13	+0.15	+0.65	+0.21	+0.21	+0.19	+0.31	+0.29	+0.01	-0.07	-0.01	-0.01	+0.04	+0.60
+stt	+0.60	+0.12	+0.20	+0.87	+0.23	+0.23	+0.22	+0.35	**+0.47**	+0.03	+0.07	+0.05	-0.01	+0.05	+0.99
+prc0	+0.31	+0.10	+0.16	+0.56	+0.06	+0.08	+0.08	+0.16	+0.16	**+0.06**	+0.06	+0.05	0.00	0.00	+0.56
+prc1	+0.40	+0.09	+0.21	+0.50	+0.06	-0.02	+0.06	+0.14	+0.11	+0.02	**+0.15**	+0.02	0.00	0.00	+0.69
+prc2	+0.23	+0.04	+0.16	+0.23	0.00	-0.01	+0.04	+0.12	+0.05	+0.04	-0.09	**+0.10**	-0.01	-0.02	+0.35
+prc3	+0.14	+0.05	+0.16	+0.33	+0.07	+0.04	+0.04	+0.15	+0.09	+0.01	-0.05	+0.05	-0.01	+0.01	+0.28
+enc	+0.26	+0.02	+0.12	+0.53	+0.09	+0.07	+0.07	+0.21	+0.22	+0.02	0.00	+0.04	-0.01	**+0.12**	+0.63
+all	+0.97	+0.23	+0.43	+1.28	+0.23	+0.25	+0.27	+0.34	+0.54	+0.05	+0.17	+0.05	-0.01	+0.13	+1.89

表 5 辞書情報の特徴ベクトルにおける各形態統語的カテゴリの影響度.

ているが，本研究では，訓練データに約 2,000 種類のタグが出現する最も粒度の高いタグセットの枠組みの 1 つを用いてタグ付けを行う．

Müeller ら [20] は，接語分割が所与のもと，各トークンに対し高粒度な品詞タグを付与する高次の条件付き確率場を用いた手法を提案している．Pasha ら [22] は，入力語が取りうるタグが登録された辞書の出力に対し，形態統語的カテゴリごとに学習された SVM 分類器を用いてランキングを行う手法を提案している．Shahrour ら [24] は，Pasha らの手法で生成された出力結果に対し，依存構造解析器の出力結果と人手の規則から得た統語情報を用いてリランキングを行っている．本研究では，接語分割は形態統語的カテゴリの一部としてモデル化されるため，Müeller らのように所与の接語分割を前提とせず，また Shahrour らのように依存構造解析器の出力結果と人手の規則から得た統語情報を必要とせずに，最高性能を達成している．

6. おわりに

本研究では，アラビア語の高粒度な品詞タグ付けのために，マルチタスク学習の枠組みを用いて，各形態統語的カテゴリを同時に予測する手法を提案した．また，さらなる性能向上のため，入力語に対して各形態統語的カテゴリが取りうるタグを登録した辞書情報をモデルに組み込む手法を提案した．Penn Arabic Treebank を用いた評価実験の結果，これまでに報告されている最高性能の品詞タガー [24] の正解率を上回ることを確認した．また UD Arabic データセットを用いた評価実験の結果，提案手法は，異なるデータセット，タグセットに対しても頑健に性能向上をもたらすことが示された．

今後は，アラビア語以外の形態的に豊かな言語へ提案手法を適用する予定である．特に提案手法は，任意のタグセットで構築された辞書があれば，容易に解析性能を向上させることができる．課題となるのは辞書の構築コストだが，Kirov ら [13] に代表される Wiktionary から構築したオープンソースの辞書を用いれば，この課題を解決することができる．また本研究では，カテゴリ間の依存関係をパラメータ共有によって捉える手法を提案したが，出力ラベル間の直接的な依存関係は扱っていない．今後は，出力ラベル間の依存関係を直接的に最適化し，かつ系列としても最適化するモデルを考案する予定である．

謝辞 New York University Abu Dhabi の Nizar Habash 氏，Salam Khalifa 氏には，前処理済みのデータならびに CamelParser を用いた実験結果を提供していただいた．また大内啓樹氏，澤井裕一郎氏，池田大志氏，真鍋陽俊氏には，数々の助言をいただいた．ここに深く感謝の意を表する．

参考文献

[1] Bingel, J. and Søgaard, A.: Identifying beneficial task relations for multi-task learning in deep neural networks, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 164–169 (2017).

[2] Buckwalter, T.: Buckwalter Arabic Morphological Analyzer Version 1.0 LDC2002L49, Linguistic Data Consortium (LDC, Philadelphia US) (2002).

[3] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P.: Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research*, Vol. 12, No. Aug, pp. 2493–2537 (2011).

[4] Diab, M., Habash, N., Rambow, O. and Roth, R.: LDC Arabic Treebanks and Associated Corpora: Data Divisions Manual, *arXiv preprint arXiv:1309.5652* (2013).

[5] Diab, M., Hacioglu, K. and Jurafsky, D.: Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks, *Proceedings of the 2004 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, Boston, Massachusetts, USA, Association for Computational Linguistics, pp. 149–152 (2004).

[6] Graves, A. and Schmidhuber, J.: Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures, *Neural Networks*, Vol. 18, No. 5, pp. 602–610 (2005).

[7] Habash, N.: Arabic Morphological Representations for Machine Translation, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pp. 263–285 (2007).

[8] Habash, N.: *Introduction to Arabic Natural Language Processing*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010).

[9] Habash, N. and Rambow, O.: Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, USA, Association for Computational Linguistics, pp. 573–580 (2005).

[10] Habash, N., Shahrour, A. and Al-Khalil, M.: Exploiting Arabic Diacritization for High Quality Automatic Annotation, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4298–4303 (2016).

[11] Khalifa, S., Zalmout, N. and Habash, N.: YAMAMA: Yet Another Multi-Dialect Arabic Morphological Analyzer, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, The COLING 2016 Organizing Committee, pp. 223–227 (2016).

[12] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[13] Kirov, C., Sylak-Glassman, J., Que, R. and Yarowsky, D.: Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S., eds.), Paris, France, European Language Resources Association (ELRA) (2016).

[14] Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., Mekki, W., Krouna, S., Bouziri, B. and Zaghouani, W.: Arabic Treebank: Part 1 v 4.1, Linguistic Data Consortium (LDC, Philadelphia US) (2010).

[15] Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., Mekki, W., Krouna, S., Bouziri, B. and Zaghouani, W.: Arabic Treebank: Part 2 v 3.1, Linguistic Data Consortium (LDC, Philadelphia US) (2011).

[16] Maamouri, M., Bies, A., Kulick, S., Krouna, S., Gaddeche, F. and Zaghouani, W.: Arabic Treebank: Part 3 v 3.2, Linguistic Data Consortium (LDC, Philadelphia US) (2010).

[17] Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A. and Kulick, S.: LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1 LDC2010L01, Linguistic Data Consortium (LDC, Philadelphia US) (2010).

[18] Martínez Alonso, H. and Plank, B.: When is multitask learning effective? Semantic sequence prediction under varying data conditions, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 44–53 (2017).

[19] Mohamed, E. and Kübler, S.: Is Arabic Part of Speech Tagging Feasible Without Word Segmentation?, *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, USA, Association for Computational Linguistics, pp. 705–708 (2010).

[20] Müeller, T., Schmid, H. and Schütze, H.: Efficient Higher-Order CRFs for Morphological Tagging, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, Association for Computational Linguistics, pp. 322–332 (2013).

[21] Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., Duh, K., Faruqui, M., Gan, C., Garrette, D., Ji, Y., Kong, L., Kuncoro, A., Kumar, G., Malaviya, C., Michel, P., Oda, Y., Richardson, M., Saphra, N., Swayamdipta, S. and Yin, P.: DyNet: The Dynamic Neural Network Toolkit, *arXiv preprint arXiv:1701.03980* (2017).

[22] Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R.: MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Vol. 14, Reykjavik, Iceland, pp. 1094–1101 (2014).

[23] Plank, B., Søgaard, A. and Goldberg, Y.: Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 412–418 (2016).

[24] Shahrour, A., Khalifa, S. and Habash, N.: Improving Arabic Diacritization through Syntactic Analysis, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 1309–1315 (2015).

[25] Søgaard, A. and Goldberg, Y.: Deep multi-task learning with low level tasks supervised at lower layers, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 231–235 (2016).

[26] Yang, Z., Salakhutdinov, R. and Cohen, W.: Multi-Task Cross-Lingual Sequence Tagging from Scratch, *arXiv preprint arXiv:1603.06270* (2016).

[27] Zhang, Y., Li, C., Barzilay, R. and Darwish, K.: Randomized Greedy Inference for Joint Segmentation, POS Tagging and Dependency Parsing, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, Association for Computational Linguistics, pp. 42–52 (2015).

付　　録

A.1 PATB データセットのタグセット

pos ($n = 35$)	noun, noun_num, noun_quant, noun_prop, adj, adj_comp, adj_num, adv, adv_interrog, adv_rel, pron, pron_dem, pron_exclam, pron_interrog, pron_rel, verb, verb_pseudo, part, part_dem, part_det, part_focus, part_fut, part_interrog, part_neg, part_restrict, part_verb, part_voc, prep, abbrev, punc, conj, conj_sub, interj, digit, latin
gen ($n = 3$)	m (masculine), f (feminine), na (not applicable)
num ($n = 5$)	s (singular), d (dual), p (plural), u (undefined), na
cas ($n = 5$)	n (nominative), a (accusative), g (genitive), u, na
mod ($n = 5$)	i (indicative), j (jussive), s (subjunctive), u, na
asp ($n = 4$)	i (imperfective), p (perfective), c (command), na
per ($n = 4$)	1, 2, 3, na
vox ($n = 4$)	a (active), p (passive), u, na
stt ($n = 5$)	i (indefinite), d (definite), c (constructive/poss/idafa), u, na
prc0 ($n = 10$)	0, na, Aa_prondem, AlmA_detneg, lA_neg, mA_neg, mA_part, mA_rel
prc1 ($n = 27$)	0, na,<i$_interrog, bi_part, bi_prep, bi_prog, Ea_prep, EalaY_prep, fiy_prep, hA_dem, Ha_fut, ka_prep, la_emph, la_prep, la_rc, libi_prep laHa_emphfut, laHa_rcfut, li_jus, li_prep, min_prep, sa_fut, ta_prep, wa_part, wa_prep, wA_voc, yA_voc
prc2 ($n = 9$)	0, na, fa_conj, fa_conn, fa_rc, fa_sub, wa_conj, wa_part, wa_sub
prc3 ($n = 3$)	0, na, >a_ques
enc ($n = 54$)	0, na, 1p_dobj, 1p_poss, 1p_pron, 1s_dobj, 1s_poss, 1s_pron, 2d_dobj, 2d_poss, 2d_pron, 2p_dobj, 2p_poss, 2p_pron, 2fp_dobj, 2fp_poss, 2fp_pron, 2fs_dobj, 2fs_poss, 2fs_pron, 2mp_dobj, 2mp_poss, 2mp_pron, 2ms_dobj, 2ms_poss, 2ms_pron, 3d_dobj, 3d_poss, 3d_pron, 3p_dobj, 3p_poss, 3p_pron, 3fp_dobj, 3fp_poss, 3fp_pron, 3fs_dobj, 3fs_poss, 3fs_pron, 3mp_dobj, 3mp_poss, 3mp_pron, 3ms_dobj, 3ms_poss, 3ms_pron, Ah_voc, lA_neg, ma_interrog, mA_interrog, man_interrog, man_rel, ma_rel, mA_rel, ma_sub, mA_sub

A.2 UD Arabic データセットのタグセット

POS ($n = 17$)	ADJ, ADP, ADV, AUX, CONJ, DET, INTEJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X
Gender ($n = 3$)	Fem, Masc, EMPTY
Number ($n = 4$)	Dual, Plur, Sing, EMPTY
Case ($n = 4$)	Acc, Gen, Nom, EMPTY
Mood ($n = 5$)	Imp, Ind, Jus, Sub, EMPTY
Aspect ($n = 3$)	Imp, Perf, EMPTY
Person ($n = 4$)	1, 2, 3, EMPTY
Voice ($n = 3$)	Act, Pass, EMPTY
Definite ($n = 5$)	Com, Cons, Def, Ind, EMPTY
Abbr ($n = 2$)	Yes, EMPTY
AdpType ($n = 2$)	Prep, EMPTY
Foreign ($n = 2$)	Yes, EMPTY
Negative ($n = 2$)	Negative, EMPTY
NumForm ($n = 3$)	Digit, Word, EMPTY
NumValue ($n = 4$)	1, 2, 3, EMPTY
PronType ($n = 4$)	Dem, Prs, Rel, EMPTY
VerbForm ($n = 2$)	Fin, EMPTY