深層学習による自然語で記述された要件仕様書に対する深層学習を用いたファンクションポイント計測支援ツールの実装

小田 郁弥　　楠本 真二

大阪大学大学院情報科学研究科
{fumy-oda,kusumoto}@ist.osaka-u.ac.jp

概要

一般にシステム開発プロジェクトの見積もりでは，まず開発規模が見積もられ，それを用いて工数や予算の見積もりが行われる．開発規模の尺度として，ファンクションポイント（FP）が存在する．FP は客観的な見積もりの指標として広く用いられており，幾つかの計測方法は ISO/IEC の国際標準規格として扱われている．一方で，FP の計測は手作業で行われており，その計測の手間は FP の導入における課題となっており，その対策の一つとして FP 計測支援ツールに関する研究が行われている．本稿では，日本語で記述された要件仕様書を対象として，深層学習モデルを用いた FP 計測支援ツールについて報告する．提案ツールの評価にあたっては，デジタル庁が公開している地方公共団体の 33 種類の基幹業務システムの標準仕様書に対して，FP の基本構成要素 5 種類の抽出・分類機能を確認した．その結果，1938 件の要件記述から学習することで，評価用の 216 要件中 180 要件で FP の基本構成要素の分類結果が人間が分類した結果と同様になるという結果を得た．

1　はじめに

一般に，ソフトウェアシステム開発においては初期段階でのシステム開発規模の見積もりが重要である．不正確な見積もりはプロジェクト失敗の原因になることも報告されている．正確な見積もりの実現を目指し，システム開発規模の見積もりに関する研究が盛んに行われている．

システム開発における見積もりの指標として，ファンクションポイント（FP）[4] が存在する．FP は客観的な見積もりの指標として広く用いられており，IFPUG 法 [6] 等の幾つかの計測手法は ISO/IEC の国際標準規格 [1] となっている．通常，FP の計測は計測者が手動で行っている．そのためプロジェクトの規模に比例した計測コストが必要となることが，FP 導入の課題の一つとなっている．文献 [5] によると，2004 年以降，FP 研究の焦点は FP の利点や評価から計測補助へと移行している．また，開発現場のニーズでも，汎用的な自動計測ツールや FP 教育手法の確立が求められている．従って，開発現場では FP 計測にかかるコストの低減と計測者を用意するために必要な教育コストが負担となっていると考えられる．また，自動計測ツールに関する研究も行われているが，特定組織の開発に依存しているものや，現場導入において手間のかかるものも多いという指摘もされている．

本研究では自然語（日本語）で記述されたシステムの要件仕様書に対し，深層学習モデルを用いた FP 計測支援ツールを提案する．ツールの開発に当たっては，深層学習に関する専門知識がなくとも利用できる事を前提し，ユーザーが過去に作成した要件仕様書から深層学習モデルを作成できるようにした．提案手法では，ユーザの所有している要件仕様書を学習することで，語彙や文体などの組織のコンテクストを考慮した深層学習モデルを作成することで，組織のコンテクストに左右されない見積もりを行う．ツールのケーススタディとして，デジタル庁の公開している 33 種の要件仕様書を用いて FP の基本構成要素 5 種類の抽出・分類機能を検証する適用実験を行った．その結果 1938 件の要件記述から学習することで，評価用の 216 要件中 180 要件で FP の基本構成要素の分類結果が人間が分類した結果と同様になるという結果を得た．以降，2 章では研究の背景となる諸用語や関連研究について述べる．3 章では提案手法を用いた計測支援ツールについて述べる．4 章では実際にツールを用いて FP の基本構成要素の抽出・分類機能を確認

した結果と考察について述べる．最後に，5 章では本研究のまとめと今後の課題を述べる．

2 準備

本章では諸用語や関連研究について簡単に述べる．

2.1 ファンクションポイント（FP）

ファンクションポイント（FP）はシステムの規模のメトリクスである [4]．FP 法はユーザーから見たシステムの各機能の量を計測する手法であり，システム開発における客観的な見積もりの指標として扱われている．FP 法は画面や帳票，ファイルなどを通じた情報の入出力に着目し，それらを種類別に数え上げ，種類数を加重合計した値を機能量としている．FP 法は 1979 年に A. J. Albrecht によって提案されて以降，IFPUG 法 [6]，NESMA 法 [7]，SFP 法 [8] 等，様々な FP 法が提案されている．以降，FP 法とは IFPUG 法を意味することとする．IFPUG 法はビジネスアプリケーションソフトウェアを対象に，欧米で広く用いられている．IFPUG 法はシステムの持つ機能を意味のあるデータ集合であるデータファンクション（DF）と，入出力や表示などのデータの流れであるトランザクションファンクション（TF）に分類する．DF は更に内部論理ファイルと外部論理ファイルの 2 種に，TF は外部入力，外部出力，外部照会の 3 種に分類される．これらの 5 つの分類は，FP の基本構成要素と呼称されている [9]．以下に詳細を示す．

1. 内部論理ファイル（ILF）
   見積もり対象のシステム内部の書き換え可能なデータの集合
2. 外部論理ファイル（EIF）
   見積もり対象のシステムによって参照されるが書き換えが行われないデータの集合
3. 外部入力（EI）
   データの書き込みや記録など，システム外部からの入力によって ILF の更新を伴う処理
4. 外部出力（EO）
   システム外部へのデータ出力のうち，出力に合計値や平均値の算出などのデータの加工を含むもの
5. 外部照会（EQ）
   システム外部へのデータ出力のうち，出力にデータの加工を必要とせず，ILF の更新処理を行わないもの

FP の点数の計測には，各基本構成要素の複雑度を計測し，各要点の点数を測定する．それらの点数の合計をシステムの機能の大きさとして評価する．また，NESMA 法や SFP 法では，複雑さが考慮されず，基本構成要素の種類ごとに点数が固定されており，それらの個数から FP の点数を計測することができる．更に，SFP 法では基本構成要素の 5 つの分類ではなく，DF と TF の二種類の分類のみで FP の算出を行うことができ，IFPUG 法に比べて適用が容易であるという利点がある．

2.2 固有表現認識（NER）

固有表現認識（NER）は，自然言語で記述されたテキスト内の固有の表現を識別し，事前に定義された組織名や人名といったラベルを付与する（カテゴリに分類する）自然言語処理のタスクであり，質疑応答 [10]，関係抽出 [11]，対話システム [12] など様々なタスクで応用されている．これにより，テキスト内の要素を自動的に抽出・分類することが可能となる．カテゴリの定義は目的に応じて自由に設定することができ，NER の手法によっては同じ単語でも前後の文脈から意味を判断することができる．例えば，”梅田駅に行くために十三人の客が十三駅で十三時に電車に乗った”のような文章があった場合，”人”の文脈を読み取ることで，最初の”十三”は数詞と判断できる．同様に”駅”の文脈から”梅田”と二つ目の”十三”は地名，”時”の文脈から三つ目の”十三”は時間と判断できる．

本研究では仕様書に記述された要件の文章のうち，FP の基本構成要素を抽出・分類するために使用する．

2.3 Bi-LSTM CRF

Bi-LSTM CRF（Bidirectional Long Short-Term Memory Conditional Random Fields）は，自然言語処理のタスクにおいて，系列ラベリング問題を解くモデルの一つである．この手法は，Bidirectional Long Short-Term Memory（Bi-LSTM）と Conditional Random Fields（CRF）を組み合わせることにより，文脈を考慮した系列タグ付けを実現する．Bi-LSTM は，文脈の前後関係を考慮することができる再帰型ニューラルネットワーク（RNN）の一種である．このモデルは，系列データ内の依存関係をモデリングするのに適している．CRF は，系列データの相互依存関係を考慮してラベルを割り当てる確率的グラフィカルモデルである．

Bi-LSTM CRF は，Bi-LSTM で抽出された特徴量を基に，CRF を使用してラベルの系列に対する最適な推定を行う．固有認識表現の分野では広く用いられており，実例としては生物医学分野でのエンティティ分類を行った Hao らによる研究がある [13].

本実験では，上述の NER を行うための手法として本モデルを使用する．

2.4 関連研究

自然語で記述された仕様書からの FP 計測を対象とした関連研究について説明する．

KUI ZHANG ら [2] は深層学習モデルを利用することで，中国語で記述された仕様書内の要件からファンクションポイントの基本構成要素を抽出・分類し，システムの規模の推定を支援する研究を行った．同研究では，29 の実プロジェクトの仕様書を用いた実験を行い，深層学習モデルは 94.25% の精度で FP の基本構成要素を抽出・分類することができ，人力での基本構成要素の分類にかかる時間を大きく削減できる可能性があると報告している．但し，この研究に用いられた要件仕様書は公開されておらず，ツールの公開も行われていない．

また，文献 citeyamada では，日本語の構文解析 (係り受け解析) ツールの cabocha を用いて，日本語で記述された要件仕様書に対するファンクションポイントの計測支援を行っている．要件定義書から機能仕様についての文章を提案されているテンプレートに沿って書き換え，その仕様を自動解析ツールを用いて解析し計測に必要な要素を抽出している．しかし，テンプレートへの書き換え部分の手間が大きいという課題がある．

3 提案ツール:FACTUAL

3.1 概要

本研究では日本語で記述されたシステムの要件定義書に対し，深層学習モデルを用いてファンクションポイントの計測支援を行うツール，FACTUAL を提案する．このツールはユーザーの端末上で Bi-LSTM CRF モデルを作成し，ファンクションポイントの基本構成要素を抽出・分類する．図 1 に IFPUG 法での FP 計測の流れと，FACTUAL が補助するタスクについて述べる．現在，FACTUAL が補助するのは要件仕様書内の機能要件の抽出と，抽出した機能要件の分類である．

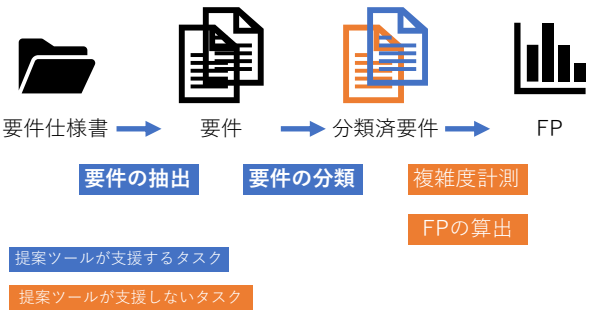

図 1　FP 計測におけるツール補助の概要

3.2 利用の流れ

FACTUAL の利用した処理には，モデルの作成と FP の基本構成要素の抽出・分類の 2 つが存在する．図 2 にツールの利用の流れを示す．まず，学習データとなる要件仕様書の各単語にラベルを付与した分類済仕様書を学習ツールに与えることで学習モデルを作る．次に，作成した学習モデルと FP 計測の対象となる要件が記入された仕様書をタグ付けツールに入力することで，ラベルが付与され要件が出力される．以下に詳細を示す．

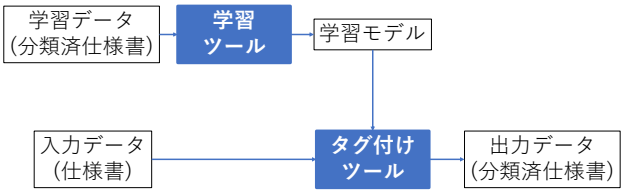

図 2　ツールを用いた処理の流れ

3.2.1 モデルの作成

学習データを基に Bi-LSTM CRF モデルを作成する．学習データは要件の文章に FP の基本構成要素のラベルを付与したものが与えられる．更に，それらの文章に対して形態素解析を行い，単語レベルに分割する．モデルの学習にはこちらの形態素解析を行ったデータを用いる．学習データの例を表 1 に，形態素解析を行い，単語レベルに分割した例を表 2 に記す．なお，形態素解析と，それに伴うラベルの付与はツール内で行われるため，ユーザーが用意するファイルは学習データのファイルの

みである．ファイルの形式は txt ファイルと csv ファイルを予定している．

ラベルは 5 つの基本構成要素 (EI, EO, EQ, ILF, EIF) に対して，文頭を意味する I を付与したもの（例えば，I-ILF) と，前の単語から継続する事を表す B を付与したもの（例えば，B-ILF）の計 10 種類に加え，基本構成要素に属さない事を示す O の 11 種類のラベルが存在する．加えて，必要に応じてラベルを追加することで，FP の基本構成要素以外の情報の抽出も可能となる．

表 1 と表 2 の例では，”データベース”が ILF，”保存”が EI にあたる．形態素解析を行うことで，”データベース”は”データ”と”ベース”の 2 語に分割される．このとき，ILF の付与された文章の文頭にあたる”データ”の側には I-ILF のタグが付与され，そのあとに続く語である”ベース”には B-ILF のタグが付与される．”保存”は一語で EI を表しているが，このような場合は I-EI が付与される．

また, 必要に際して他のラベルを追加することで, ユーザーにとって必要な情報の抽出にも用いることができるほか，基本構成要素のラベルを別のものに置き換えてモデルを作成することもできる．例えば，シンプルファンクションポイント（SFP）の計測においては，DF と TF の二つで十分である．この際はラベルに基本構成要素ではなく，DF と TF，それ以外を示す O のラベルを用いることで学習データを作成することができる．

モデルの作成は，ユーザーが保有している要件仕様書などの過去の開発資料を利用することを想定している．これにより，ユーザーの文章の書式や使用する語彙の傾向をモデルに反映することができるため，推測の精度向上を見込むことができる．

表 1　作成する学習データの例

文章	ラベル
入力されたテキストを	O
データベース	ILF
に	O
保存	EI
する	O

表 2　形態素解析を行った学習データの例

単語	ラベル
入力	O
された	O
テキスト	O
を	O
データ	I-ILF
ベース	B-ILF
に	O
保存	I-EI
する	O

3.2.2　FP の基本構成要素の抽出・分類

前項で作成したモデルに，仕様書の要件の文章を入力として与える．入力された文章は形態素解析により単語ごとに分割され，各単語にラベルが付与されたものが出力される．

ラベルの個数を計測することで，NESMA 法や SFP 法による FP の点数の算出が可能になる．また，複雑度の算出を行う際にも，FP の基本構成要素の対象となる単語が明らかになっていることから，見積もり作業のコストの低減が期待できる．

4　実験

4.1　概要

提案手法の精度を調査するために，実プロジェクトの仕様書に提案ツールを用いて実験を行った．本実験では学習によって FP の基本構成要素の抽出・分類が行えることの確認を目的とし，学習の経過と，学習の結果の評価を示す．

以下に実験の詳細を示す．

4.2　実験設定

実験用のデータとして，デジタル庁の提供するデータ要件・連携要件の標準仕様[*1]に含まれる 33 プロジェクト 2154 要件を用いた．この仕様書の各要件に表 1 の形で筆者が手動でラベル付を行い，実験用のデータを作成した．要件全体の 90% を学習データ，残りを試験データとして 150 エポックの学習を行う．学習は学習データを読み込み，モデルのパラメーターの更新を行うことで

[*1] https://www.digital.go.jp/policies/local_governments/specification/

行われる．学習データ全てを読み込み，パラメーターを更新することで 1 エポックの学習が完了する．学習データと試験データに含まれる各ラベルの単語数を表 3 に記す．実験に用いたデータはシステム間の連携に関する仕様書であったため，調査対象のシステムであることを意味するラベルの TS と，連携先のシステムである事を意味するラベルの ES を追加し，調査を行った．これにより，ユーザーが FP の基本構成要素以外の情報の抽出を行うことができるかどうかの確認を行うことができる．

表 3　データに含まれる各ラベルの単語数

ラベル	学習データ	試験データ
I-TS	1936	217
B-TS	6196	681
I-ES	1969	220
B-ES	3743	586
I-ILF	498	45
B-ILF	2187	199
I-EIF	1471	177
B-EIF	6423	835
I-EI	262	25
B-EI	0	0
I-EO	709	82
B-EO	36	1
I-EQ	968	109
B-EQ	34	0
O	18128	2093
合計	46560	5270

4.3　評価方法

本実験では以下の 4 つの指標を基に評価する．

4.3.1　各ラベルごとの F1 値

各単語のラベルの予測の正答率を，学習データ，試験データそれぞれについて，各ラベル毎に F1 値を算出することで評価する．算出は 150 エポックの学習が終了した時点で行う．また，学習データの B-EI タグのように，データが存在しないものについては評価しない．F1 値は 0 以上 1 以下の値であり，精度が高ければ大きな値を取る．

4.3.2　単語ごとのラベル一致数

文章内の各単語に対して，設定したラベルと予測したラベルが一致する単語数を各エポックごとに評価する．正確な予測ができているならばラベルの一致数が増加する．

4.3.3　要件ごとのラベルの完全一致数

各要件のうち，含まれる単語全ての予測したラベルが設定したラベルと一致している要件の個数を各エポックごとに評価する．正確な予想ができているならばラベルの一致数が増加する．

4.3.4　要件ごとのラベルの個数一致数

各要件のうち，要件内の予測した基本構成要素のラベルが設定した基本構成要素のラベルと一致しているかを各エポックごとに評価する．具体的には両者の各ラベルの先頭要素である”I-（ラベル名）”の個数を比較する．これは NESMA 法や SFP 法の見積もりでは基本構成要素の個数が重要であること，また，ラベルの予測の際に見積もりには影響しないような軽微な付与の揺れを許容するためにこれを調査する．

完全一致ではないが個数一致である要件について，表 4 に記す．この例のように，修飾する語を含んでラベルを付与するか否かなどの軽微な揺れをこの評価指標で吸収することができる．

表 4　個数一致となった要件の例

単語	正解ラベル	予測ラベル
国民	I-TS	I-TS
健康	B-TS	B-TS
保険	B-TS	B-TS
システム	B-TS	B-TS
が	O	O
次期	I-ES	I-ES
国保	B-ES	B-ES
総合	B-ES	B-ES
システム	B-ES	B-ES
に	O	O
給付	I-EIF	O
管理	B-EIF	O
に	B-EIF	O
必要な	B-EIF	O
各種	B-EIF	I-EIF
情報	B-EIF	B-EIF
を	O	O
照会	I-EQ	I-EQ
する	O	O

4.4　結果と考察

実験結果のグラフを図 3, 4, 5, 6 に示す．いずれも 60 エポック程度までは大きな変動はないが，その後，徐々に右肩上がりに精度が向上していることが見て取れる．このことから，提案手法によって要件仕様書を学習し，要件の抽出・分類が可能であると考えられる．しかし，周期的に精度が大きく低下している点や，最初の 60 エポックまで精度に変動がないことなど，現状説明が不能な点がいくつか存在しており，そのことについては更なる調査が必要と考える．

150 エポックの学習終了時点での各評価指標の数値と各ラベル毎の F1 スコアを表 5 と表 6 に記す．各指標について，学習データでは約 97% の要件に対して，完全に正しくラベルを付与できていることがわかる．試験データでは約 83% の要件について完全に正しくラベルを付与することができており，約 88% の要件に対してラベルの個数が正しいという結果になった．学習データの各ラベルの F1 スコアは全て 0.9 を上回っており，正しく学習できていることが読み取れる．試験データに関しては B-EO のラベルの F1 スコアが大幅に低くなっている．これは B-EO が正解になる単語が 1 つしか存在せず，学習データでも 36 単語しか存在しないことなど，データの不足によるものが原因と考えられる．

また，試験データの F1 スコアのうち，DF である ILF と EIF の数値が他と比べてやや低くなっている．これに関しては文章の前後関係を踏まえて学習を行う Bi-LSTM CRF の影響と考えられる．今回の実験に用いた文章の多くは，”対象システムが外部システムからデータを受信する”のような形で記述されている．そのため，TS や ES のタグは順番が固定化されており，前後関係を基に判断がしやすい．TF にあたる EI，EO，EQ はそれぞれ”入力”，”出力”，”照会”等，各ラベルごとの語彙が少なく，また，他のラベルと競合することがないため，高い精度で分類できると考えられる．他方，DF は同じ位置に存在し，同じような語彙が異なるラベルに混在することから，ILF と EIF の混同が見受けられた．その結果，F1 スコアが他のラベルよりも低くなったと考えられる．

また，本研究では複雑度の判定を行っていないため，IFPUG 法による FP の算出にはここから複雑度の算出を行う必要がある．しかし，NESMA 法や SFP 法では複雑度の算出が不要であるため，本ツールの結果を用いることで要件の抽出から FP の算出までの流れを自動化できると考えられる．

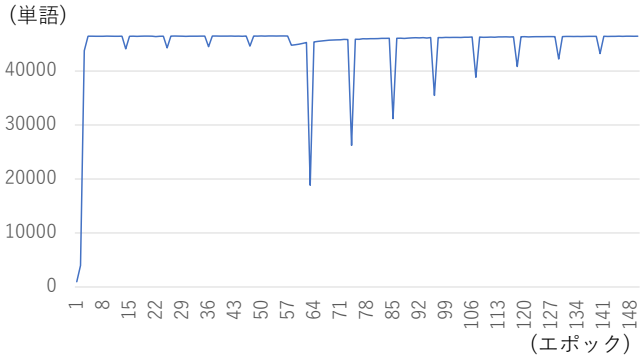

図 3　試験データの単語ごとのラベル一致数

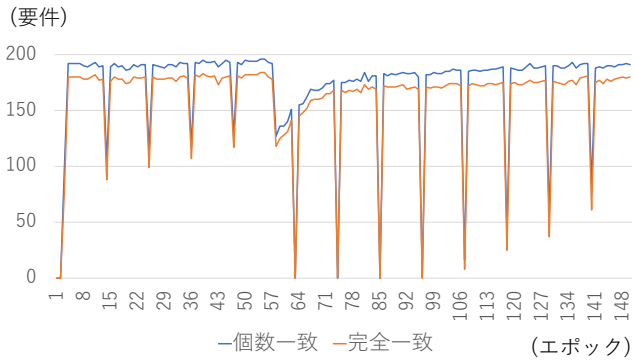

図 4　試験データの要件の個数一致数と完全一致数

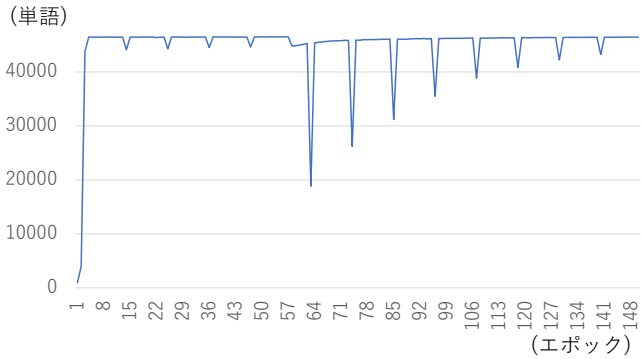

図 5　学習データの単語ごとのラベル一致数

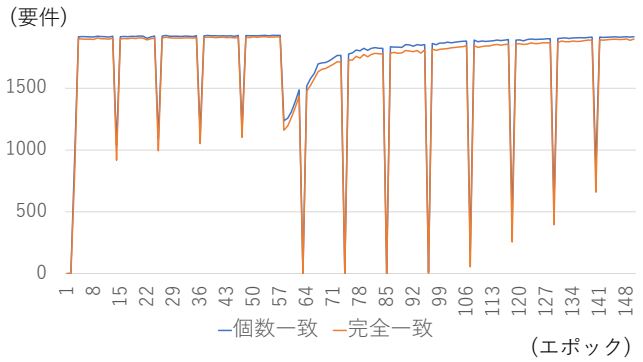

図 6　学習データの要件の個数一致数と完全一致数

表 5　学習データと試験データの評価指標

評価指標	学習データ	試験データ
単語ごとの一致数	46453/46560	5112/5270
要件ごとの完全一致数	1899/1938	180/216
要件ごとの個数一致数	1918/1938	191/216

表 6　学習データと試験データの F1 スコア

ラベル	学習データ	試験データ
I-TS	0.9994	0.9953
B-TS	0.9997	0.9955
I-ES	0.9982	0.9816
B-ES	0.9982	0.9837
I-ILF	0.9889	0.8571
B-ILF	0.9940	0.8733
I-EIF	0.9935	0.8965
B-EIF	0.9968	0.9402
I-EI	0.9924	1.0000
B-EI	-	-
I-EO	0.9964	0.9878
B-EO	0.9315	0.5000
I-EQ	1.0000	1.0000
B-EQ	0.9855	-
O	0.9980	0.9825

5　おわりに

日本語で記述された要件仕様書を対象とした深層学習モデルを用いた FP 計測支援ツールの提案を行った．ツールの精度を確認するため，デジタル庁の公開している 33 プロジェクトの要件仕様書を用いて実証実験を行い，十分な精度であることを確認した．本研究の今後の予定としては以下のことが挙げられる．

1. 追加実験の実施
   新たな要件仕様書に対する実験を予定している．総務省などの公開している情報システムの標準化に関する仕様書に対して，同様の実験を行う．また，加えて交叉実験を行うことで，使用する語彙や表現の近い仕様書を学習に用いる事が有効に働くことを確認する予定である．
   追加実験の際に改良点が見つかった場合，ツールに反映し，より有益なツールになる様に改良に努める．
2. ツールの公開
   上記の追加実験で発見された改良点を基に，現在のツールに改善を加えたのち，ツールの公開を予定している．OSS として広く公開することで，本研究及びツールの有益性を高める．

謝辞

本研究は一部 JSPS 科研費 21K11829 の助成を受けている.

参考文献

[1] International Organization for Standardization (ISO)”ISO/IEC 20926:2009 - Software and systems engineering — Software measurement — IFPUG functional size measurement method 2009”https://www.iso.org/standard/51717.html

[2] K. Zhang, X. Wang, J. Ren and C. Liu, ”Efficiency Improvement of Function Point-Based Software Size Estimation With Deep Learning Model,” in IEEE Access, vol. 9, pp. 107124-107136, 2021, doi: 10.1109/ACCESS.2020.2998581.

[3] 山田涼太, 山田悠斗, 楠本真二, 柗本真佑, 肥後芳樹 ”要件定義書からのファンクションポイント自動計測の試み”ソフトウェア・シンポジウム 2017

[4] A. J. Albrecht. Function point analysis. Encyclopedia of Software Engineering, Vol. 1, pp.518-524,Addison-Wesley Professional, 1994.

[5] 山田涼太, 柗本真佑, 肥後芳樹, 楠本真二 ”ファンクションポイントについての研究論文の系統的レビュー”ソフトウェア信頼性研究会第 14 回ワークショップ

[6] IFPUG: Function Point Counting Practices Manual, Release 4. 3. International Function Point Users Group, 2010

[7] NESMA（National Electrical Switchboard Manufacturers Association：オランダソフトウェア計測協会）,”Early Function Point Analysis”, https://nesma.org/themes/sizing/function-point-analysis/early-function-point-counting/

[8] IFPUG（International Function Point USERS GROUP）,Introducing Simple Function Points (SFP),https://ifpug.org/2021/10/26/introducing-simple-function-points-sfp

[9] 柏本隆志, 楠本真二, 井上克郎, 鈴木文音, 湯浦克彦, 津田道夫. イベントトレース図に基づく要求仕様書からのファンクションポイント計測手法. Vol.41,No. 6, pp. 1895-1904, 2000.

[10] 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.

[11] 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.

[12] 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).

[13] Wei, Hao, et al. ”Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF.” IEEE Access 7 (2019): 73627-73636.