

Automated Extraction of Non-functional Requirements in Available Documentation

John Slankas

Laurie Williams

2013 1st international Workshop on Natural Language Analysis in
Software Engineering(NaturaLiSE)

目次

- 背景
- 目的
- RQ
- アプローチ
- 実験結果
- まとめ
- 所感

背景

- **NFR(非機能要件)**は要件書内に書かれていても、見過ごされることが多く、見過ごされた**NFR**を特定することはコストを要する
- 開発者は**NFR**を素早く理解し、分類できる必要がある

目的

- 自然言語処理を用いて，自然言語の文書内のNFRを特定し，抽出することで開発者の手助けをする

RQ

- RQ1: 文書の種類によって出現するNFRに違いはあるか
- RQ2: 特定のキーワード(タイムピリオド, パーセンテージ)はどのカテゴリに割り振られるか
- RQ3: NFRの分類において, どの分類器の精度が一番良いか
- RQ4: どういった特徴語が精度に影響を及ぼすか

アプローチ

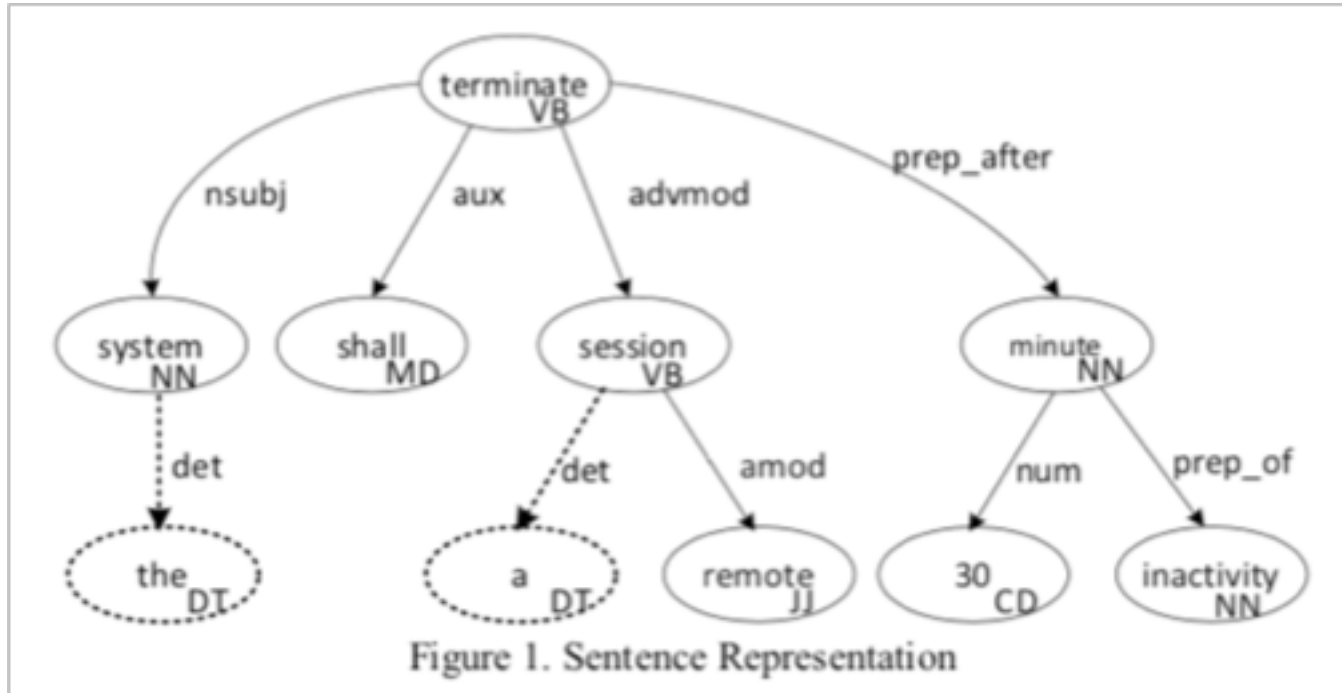
- **NFR**分類を行うツール, 「**NFR Locator**」を提案する
- 入力は構文解析を行なった自然言語
- 該当する**NFR**カテゴリを返す

構文解析

- 自然言語をStanford Natural Language Parser(NLP)を用いて解析し,
Stanford Type Dependency Representation(STDR)でグラフ化する

構文解析

(例) The system shall terminate a remote session after 30 minutes of inactivity.



距離の計算

- K近傍法を用いるための距離(0 ~ 1)を計算する

```
computeVertexDistance(Vertex a, Vertex b)
1: if a = NULL or b = NULL return 1
2: if a.partOfSpeech <> b.partOfSpeech return 1
3: if a.parentCount <> b.parentCount return 1
4: for each parent in a.parents
5:     if not b.parents.contains(parent) return 1
6: if a.lemma = b.lemma return 0
7: if a and b are numbers, return 0
8: if ner classes match, return 0
9: wnValue = wordNetSynonyms(a.lemma,b.lemma)
10: if wnValue > 0 return wnValue
11: return 1
```

Figure 2. Compute Vertex Distance Logic

9,10行目が返す値が主に分類に用いられる

使用するデータセット

Document	Document Type
CCHIT Ambulatory Requirements	Requirement
iTrust	Requirement, Use Case
PromiseData	Requirement
Open EMR Install Manual	Installation Manual
Open EMR User Manual	User Manual
NC Public Health DUA	DUA
US Medicare/Medicaid DUA	DUA
California Correctional Health Care	RFP
Los Angeles County EHR	RFP
HIPAA Combined Rule	CFR
Meaningful Use Criteria	CFR
Health IT Standards	CFR

Requirement：要件書

DUA：データ使用契約書

RFP：提案依頼書

CFR：米国連邦規則

分類のカテゴリ

アクセス制御(AC)	操作性(OP)
審査, 監査(AU)	プライバシー(PR)
可用性(AV)	回復性(RC)
性能(PS)	信頼性(RL)
法律, 正当性(LG)	セキュリティ(SE)
見た目と雰囲気(LF)	使用性(US)
保守性(MT)	その他(OT)

実験結果

- RQ1: 文書の種類によって出現するNFRに違いはあるか

TABLE 1. CLASSIFIED DOCUMENTS BY REQUIREMENTS CATEGORY

Document	Document Type	Size	AC	AU	AV	LG	LF	MT	OP	PR	PS	RC	RL	SC	US	OT	FN	NA
CCHIT Ambulatory Requirements	Requirement	306	12	27	1	2	0	10	0	0	1	5	2	28	4	8	228	6
iTrust	Requirement, Use Case	1165	439	44	0	2	2	18	2	9	0	9	9	55	2	0	734	376
PromiseData	Requirement	792	164	20	36	10	50	26	89	7	75	4	12	71	101	19	340	0
Open EMR Install Manual	Installation Manual	225	3	0	0	0	0	0	5	1	0	6	1	25	0	0	2	184
Open EMR User Manual	User Manual	473	169	0	0	0	14	0	0	0	0	0	0	8	4	0	286	95
NC Public Health DUA	DUA	62	1	0	0	20	0	0	0	4	0	0	0	1	0	0	0	41
US Medicare/Medicaid DUA	DUA	140	1	0	0	26	0	0	0	17	0	0	0	0	0	5	2	108
California Correctional Health Care	RFP	1893	94	120	9	85	0	133	94	52	13	16	13	193	14	38	987	409
Los Angeles County EHR	RFP	1268	58	37	8	3	2	28	19	3	11	8	13	108	21	10	639	380
HIPAA Combined Rule	CFR	2642	28	8	3	0	0	78	0	213	0	9	0	41	1	0	317	2018
Meaningful Use Criteria	CFR	1435	0	0	0	0	0	0	0	0	0	0	0	8	0	0	116	1311
Health IT Standards	CFR	1475	10	20	0	0	0	119	0	1	0	2	2	71	1	2	164	1146
Total		11876	979	276	57	152	68	413	207	300	100	50	43	563	148	82	3568	6076

実験結果

- RQ1: 文書の種類によって出現するNFRに違いはあるか

TABLE 1. CLASSIFIED DOCUMENTS BY REQUIREMENTS CATEGORY

Document	Document Type	Size	AC	AU	AV	LG	LF	MT	OP	PR	PS	RC	RL	SC	US	OT	FN	NA
CCHIT Ambulatory Requirements	Requirement	306	12	27	1	2	0	10	0	0	1	5	2	28	4	8	228	6
iTrust	Requirement, Use Case	1165	439	44	0	2	2	18	2	9	0	9	9	55	2	0	734	376
PromiseData	Requirement	792	164	20	36	10	50	26	89	7	75	4	12	71	101	19	340	0
Open EMR Install Manual	Installation Manual	225	3	0	0	0	0	0	5	1	0	6	1	25	0	0	2	184
Open EMR User Manual	User Manual	473	169	0	0	0	14	0	0	0	0	0	0	8	4	0	286	95
NC Public Health DUA	DUA	62	1	0	0	20	0	0	0	4	0	0	0	1	0	0	0	41
US Medicare/Medicaid DUA	DUA	140	1	0	0	26	0	0	0	17	0	0	0	0	0	5	2	108
California Correctional Health Care	RFP	1893	94	20	9	85	0	133	94	52	13	16	13	193	14	38	987	409
Los Angeles County EHR	RFP	1268	58	37	8	3	2	28	19	3	11	8	13	108	21	10	639	380
HIPAA Combined Rule	CFR	2642	28	8	3	0	0	78	0	213	0	9	0	41	1	0	317	2018
Meaningful Use Criteria	CFR	1435	0	0	0	0	0	0	0	0	0	0	0	8	0	0	116	1311
Health IT Standards	CFR	1475	10	20	0	0	0	119	0	1	0	2	2	71	1	2	164	1146
Total		11876	979	276	57	152	68	413	207	300	100	50	43	563	148	82	3568	6076

実験結果

- RQ2:特定のキーワード(タイムピリオド, パーセンテージ)はどのカテゴリに割り振られるか

TABLE 2. TOP 20 KEYWORDS BY CATEGORY

NFR Category	Keywords
Access Control (AC)	choose, lhcp, bcp, visit, privilege, read, office, add, representative, sort, name, administrator, personal, dlhcp, view, status, accessor, edit, role, list
Audit (AU)	authorship, trail, arise, worksheet, auditable, exclusion, reduction, deletion, examine, editing, stamp, non-repudiation, inclusion, id, alteration, finalize, disable, summarize, attestation, log
Availability (AV)	achieve, 24, availability, 98, addition, available, 99, hour, day, online, schedule, confidentiality, resource, technical, year, transmit, integrity, maintenance, %, period
Legal (LG)	infeasible, custodian, hipaa, breach, dua, discovery, iihus, publication, iih, recipient, delay, secretary, definition, harm, scope, jurisdictional, affect, derive, vocabulary, reuse
Look & Feel (LF)	appearance, scheme, tree, radio, appeal, color, look, navigation, sound, feel, ship, left, shot, menu, ccr, button, corporate, page, openemr, employer
Maintenance (MT)	4010, washington, ibr, x12n, asc, 2002, addenda, 837, september, 1999, 1.1, tele-communication, 5.1, astm, draft, february, january, 2010, context-aware, infobutton
Operational (OP)	mysql, microsoft, euhr, soms, letter, infrastructure, interoperability, connect, cches, machine, browser, platform, cardmember, central, cdcd, extraction, cche, model, registry, interchange
Privacy (PR)	health, protected, entity, disclose, covered, use, disclosure, individual, such, purpose, law, permit, other, section, require, plan, person, paragraph, care, request
Recoverability (RC)	restore, credentials, backup, back, recovery, disaster, previous, emergency, establish, copy, state, need, implement, loss, plan, event, failure, organization, business, hour
Performance & Scalability (PS)	fast, simultaneous, 0, second, scale, capable, increase, peark, longer, average, acceptable, lead, handle, flow, response, capacity, 10, maximum, cycle, distribution
Reliability (RL)	reliable, dependent, validate, validation, input, query, accept, loss, failure, operate, alert, laboratory, prevent, database, product, appropriate, event, application, capability, ability
Security (SC)	cookie, encrypted, ephi, http, predetermined, strong, vulnerability, username, inactivity, portal, ssl, deficiency, uc3, authenticate, certificate, session, path, string, password, incentive
Usability (US)	easy, enterer, wrong, learn, word, community, drop, realtor, help, symbol, voice, collision, training, conference, easily, successfully, let, map, estimator, intuitive

実験結果

- RQ3:NFRの分類において, どの分類器の精度が一番良いか

TABLE 3. STRATIFIED 10-FOLD CROSS VALIDATION

Classifier	Precision	Recall	F_1	F_1 SD
Weighted Random	.047	.060	.053	.0042
50% Random	.044	.502	.081	.0016
Naïve Bayes	.227	.347	.274	.0043
SMO	.728	.544	.623	.0132
NFR Locator k -NN	.691	.456	.549	.0047

K-NNよりSMOの方が精度が高くなった

実験結果

- RQ4: こういった特徴語が精度に影響を及ぼすか

TABLE 4. WORD TYPE AND STOP WORDS

Model	Word Form	Stop Words	F_1	F_1 SD
Naïve Bayes	Original	Determiners	.291	.0022
Naïve Bayes	Porter	Determiners	.287	.0021
Naïve Bayes	Lemma	Determiners	.292	.0032
Naïve Bayes	Lemma	Frakes	.297	.0021
Naïve Bayes	Casamayor	Glasgow	.327	.0018
SMO	Original	Determiners	.603	.0044
SMO	Lemma	Determiners	.584	.0039
SMO	Lemma	Frakes	.586	.0042

Original: 文に現れた単語

Porter: ポーターステミングアルゴリズムから生成されたステム

Determiners: a, theなどの冠詞

実験結果

- RQ4: こういった特徴語が精度に影響を及ぼすか

TABLE 4. WORD TYPE AND STOP WORDS

Model	Word Form	Stop Words	F_1	F_1 SD
Naïve Bayes	Original	Determiners	.291	.0022
Naïve Bayes	Porter	Determiners	.287	.0021
Naïve Bayes	Lemma	Determiners	.292	.0032
Naïve Bayes	Lemma	Frakes	.297	.0021
Naïve Bayes	Casamayor	Glasgow	.327	.0018
SMO	Original	Determiners	.603	.0044
SMO	Lemma	Determiners	.584	.0039
SMO	Lemma	Frakes	.586	.0042

冠詞を除外することで、F値は0.027向上した

まとめ

- NFRを効果的に抽出するツールを紹介した
- NFRは多種多様な文書に存在した
- 分類精度はSMOが最も高かった

所感

- 分類器に関する勉強が必要だと感じた
- K近傍法より**SMO**分類器の方が良い？
- 日本語のテキストでのストップワードをどうするか考慮する必要あり