

Detecting Speech Act Types in Developer Question/Answer Conversations during Bug Repair

Andrew Wood, Paige Rodeghero
Ameer Armaly, Collin McMillan

Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering
Conference and Symposium on the Foundations of Software Engineering Pages 491-502

背景

- バーチャルアシスタントは人間の会話を理解可能
 - 日常のタスクのみ
 - ソフトウェア工学は？

問題点

- ソフトウェア工学の会話について発表されたデータに関する実験がほとんどない
- ソフトウェアエンジニアが行う言語行為の種類の記載がない
- 発話行為を自動で検出するアルゴリズムがない

言語行為(speech act)

- 発話とそれに伴う行為
 - 質問に答えること
 - 約束すること

アプローチ

- 30人のprofessional programmerを募集
- 2時間バグ修正を行う
 - 質問にはバーチャルアシスタントを使用
 - ▶ Wizard of Oz法[1]

[1]:N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies — why and how.

Knowledge-Based Systems 6, 4 (1993), 258 – 266. [https://doi.org/10.1016/0950-7051\(93\)90017-N](https://doi.org/10.1016/0950-7051(93)90017-N) Special Issue: Intelligent User Interfaces.

使用したバグ

- 解決済みのJavaのバグ
- バグを選んだ基準
 1. 解決に少なくとも20分を要する
 2. 十分に理解できる
 3. ユーザーが簡単にバグを再現
 4. よく知られていないドメイン知識がなくても解決可能

RQ(1/2)

RQ1.別のプログラマーが同じバグについて似た質問をすることがあるか

RQ2. プログラマーはバグ修正中にどのような質問をしたか

RQ3. プログラマーはどのような質問をよくしたか

RQ1:結果

- プログラマーは似た質問をしている

RQ2:結果

- APIなどの様々な質問があった
- 質問のタイプを9のカテゴリに絞る
 - syntax, parameter, documentation, API, clarification, implementation, bug report, confirmation, system

RQ3:結果

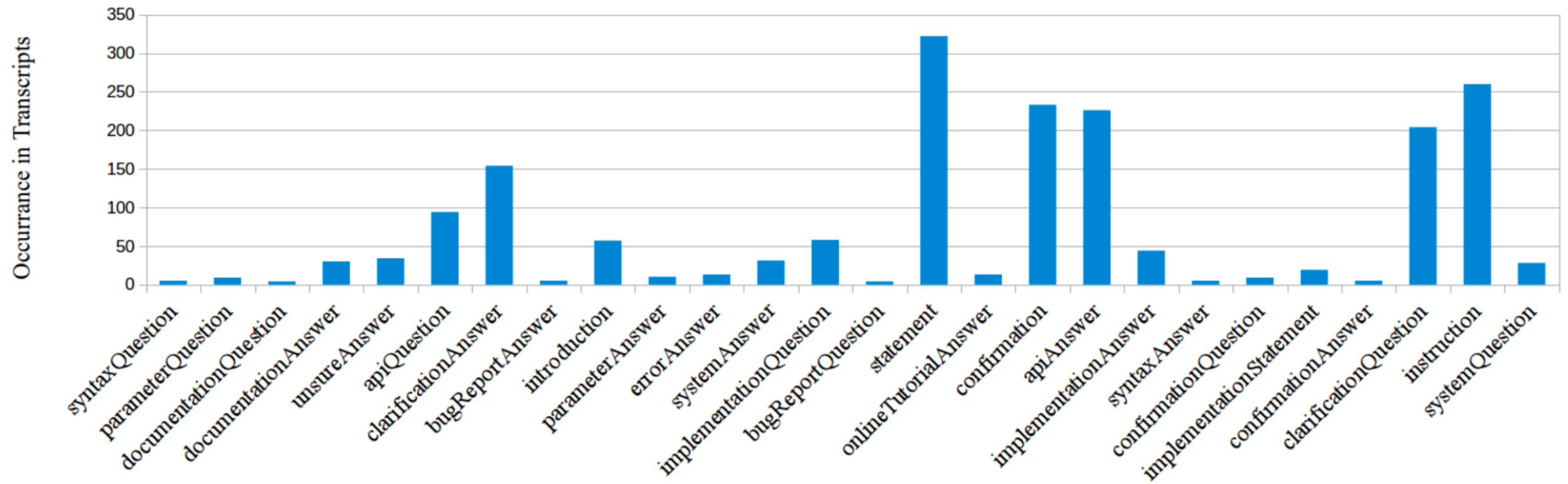


Figure 2: The annotation labels of all 30 transcripts and the occurrences for each label. Each turn can have multiple labels (speech act type).

- clarificationQuestionが最も多い

RQ(2/2)

RQ4. マルチラベル構成の全体的な予測モデルのパフォーマンスはどのくらいか

RQ5. 予測モデルはどの言語行為に対して最も高い性能を有するか

RQ6. 最も有益な特性はどれか

評価実験

- ロジスティック回帰
- 5分割交差検証
 - トレーニングデータにSMOTE[2]を適用

[2]:Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002.

SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), 321–357.

RQ4:結果

Table 1: Performance metrics calculated for each speech act type (some speech act types have been abbreviated). Recall that the averages are a weighted average based on the support for each speech type, see Section 8.3.

	precision	recall	f-measure	support
apiAnswer	0.93	0.76	0.83	24.6
apiQuestion	0.81	0.66	0.71	17.2
clarifAnswer	0.13	0.07	0.09	6.0
clarifQuestion	0.59	0.41	0.48	32.6
confirmation	0.88	0.8	0.83	27.0
docAnswer	0.25	0.2	0.22	3.2
implQuestion	0.52	0.21	0.28	10.6
implStatement	0.0	0.0	0.0	3.0
introduction	0.76	0.6	0.63	4.0
stmnt	0.69	0.4	0.51	49.8
systemQuestion	0.37	0.22	0.27	4.8
avg / total	0.69	0.5	0.57	16.62

RQ5:結果

- 高いもので, precision90%を超えrecall75%を超えた
- 低いもので, precision,recall共に10%を下回る

RQ6:結果

Table 2: The top 10 most-informative features for each speech act type, calculated by f-score. Most features are words, but features with the suffix `_sf` are shallow features (see Section 7.2). See Section 9.3 for a deeper discussion of this table.

	10	9	8	7	6	5	4	3	2	1
apiAnswer	node	if	onfinished	keyframes	constructor	values	time	timeline	keyvalue	keyframe
apiQuestion	size	method	have	how	pane	class	object	an	does	what
clarifAnswer	compilation	configurations	word	trigger	supply	green	appear	box	clicking	bottom
clarifQuestion	need	the	or	other	wc_sf	you	this	fix	prime bug	
confirmation	of	yes	is	to	thanks	slen_sf	the	thank	wc_sf	ok
documentAnswer	byte	marks	later	reading	bytes	joptionpane	external	input	audio	stream
implementQuestion	face	why	mark	eratosthenes	occurs	gets	arraycopy	reason	button	clicked
implementStatement	signature	widget	funtion	hidden	drawing	waitfor	throwing	paint	timeout	jcomponent
introduction	supervised	programmers	today	hello	human	am	start	hi	study	ready
statement	seems	what	slen_sf	looks	fixed	but	works	was	think	it
systemQuestion	password	there	permitted	lang	running	way	programs	kill	eclipse	how

- shallow feature[3, 4, 5]は有用
- 特有の単語がある

[3]: Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 773–782.

[4]: Sarah Rastkar, Gail C Murphy, and Gabriel Murray. 2014. Automatic summarization of bug reports. IEEE Transactions on Software Engineering 40, 4 (2014), 366–380.

[5]: Paige Rodeghero, Siyuan Jiang, Ameer Armaly, and Collin McMillan. 2017. Detecting User Story Information in Developer-client Conversations to Generate Extractive Summaries. In Proceedings of the 39th International Conference on Software Engineering (ICSE ’17). IEEE Press, Piscataway, NJ, USA, 49–59. <https://doi.org/10.1109/ICSE.2017.13>

まとめと所感

- まとめ
 - ソフトウェア工学の会話の提供
 - 言語行為の分類システムの作成
 - バーチャルアシスタントの基礎を築く
- 所感
 - ラベルの説明が欲しかった