PURE: Um conjunto de dados de documentos de requisitos públicos

Artigo de conferência · Setembro de 2017

DOI: 10.1109/RE.2017.29

3 autores, incluindo:

Alessio Ferrari

Conselho Nacional de Pesquisa Italiano

147 PUBLICAÇÕES 3.103 CITAÇÕES

VER PERFIL

Stefania Gnesi

Conselho Nacional de Pesquisa Italiano

318 PUBLICAÇÕES 6.001 CITAÇÕES

VER PERFIL

PURE: um conjunto de dados públicos Documentos de Requisitos

Alessio Ferrari
ISTI-CNR
Pisa, Itália E-mail: alessio.ferrari@isti.cnr.it

Giorgio O. Spagnolo ISTI-CNR Pisa,
Itália E-mail: spagnolo@isti.cnr.it

Stefania Gnesi
ISTI-CNR
Pisa, Itália E-mail: stefania.gnesi@isti.cnr.it

Resumo — Este artigo apresenta o PURE (PUblic REquirements dataset), um conjunto de dados de 79 documentos de requisitos de linguagem natural disponíveis publicamente coletados da Web. O conjunto de dados inclui 34.268 frases e pode ser usado para tarefas de processamento de linguagem natural típicas da engenharia de requisitos, como síntese de modelos, identificação de abstrações e avaliação de estrutura de documentos. Ele pode ser ainda mais anotado para funcionar como um benchmark para outras tarefas, como detecção de ambiguidade, categorização de requisitos e identificação de requisitos equivalentes. No artigo, apresentamos o conjunto de dados e comparamos sua linguagem com textos genéricos em inglês, mostrando as peculiaridades do jargão de requisitos, feito de um vocabulário restrito de acrônimos e palavras específicas de domínio e frases longas. Também apresentamos o formato XML comum para o qual portamos manualmente um subconjunto dos documentos, com o objetivo de facilitar a replicação de experimentos de PNL.

I. INTRODUÇÃO

Os requisitos são normalmente expressos com o mais flexível dos códigos de comunicação, que é a linguagem natural (NL) [1]. Vários autores aplicaram técnicas de processamento de linguagem natural (NLP) na engenharia de requisitos (RE) para abordar múltiplas tarefas, incluindo síntese de modelos [2], classificação de requisitos em categorias funcionais/não funcionais [3], classificação de revisões de produtos online [4], rastreabilidade [5], [6], detecção de ambiguidade [7]–[9], avaliação de estrutura [10], detecção de requisitos equivalentes [11], avaliação de completude [12] e extração de informações [13]–[15]. Com algumas exceções, a maioria dos trabalhos usa documentos proprietários ou específicos de domínio como benchmarks, e a replicação dos experimentos, bem como a generalização dos resultados, sempre foram um problema [16]. Este artigo apresenta o PURE (conjunto de dados PUblic REquirements), um conjunto de dados de 79 documentos de requisitos disponíveis publicamente recuperados da Web. O conjunto de dados é orientado para a replicação de experimentos de NLP e generalização de resultados. Os documentos abrangem vários domínios, têm diferentes graus de abstração e variam de padrões de produtos a documentos de empresas públicas e projetos universitários. Também definimos um arquivo de esquema XML geral (XSD) para representar esses diferentes documentos em um formato uniforme. Atualmente, portamos um subconjunto dos documentos para esse formato para facilitar a comparação rigorosa de experimentos de PNL. O artigo estende uma contribuição recente de conferência [17]. Com relação a esse trabalho anterior, fornecemos informações estatísticas sobre o conteúdo NL do conjunto de dados, apresentamos o esquema XSD adotado para formatar os documentos e fornecemos recomendações sobre o uso do conjunto de dados.

O restante do artigo é estruturado da seguinte forma. Na Seção II, apresentamos os documentos recuperados da Web. Na Seção III, descrevemos o arquivo XSD que definimos. Na Seção IV, discutimos o uso do conjunto de dados e suas limitações. Finalmente, a Seção V fornece comentários finais.

II. O CONJUNTO DE DADOS PURO

O conjunto de dados PURE é composto de documentos de requisitos públicos recuperados da Web. Para recuperar os documentos, consultamos o Google com as palavras-chave vinculadas a OR Requirements Documents, Requirements Specification, System Specification, Software Specification, SRS, e selecionamos os links que apontavam para documentos de requisitos (em .doc, .pdf, .html, .rtf).

Além disso, navegamos pelos sites de origem nos quais os documentos estavam localizados para procurar documentos de requisitos adicionais. Nossa busca levou à identificação de 79 documentos. O conjunto de dados completo — junto com o arquivo XSD e os arquivos XML atualmente portados — pode ser baixado do nosso site (http://fmt.isti.cnr.it/nlreqdataset/). Não reivindicamos uma cobertura sistemática de todos os documentos de requisitos públicos disponíveis na Web, também dada a natureza dinâmica das pesquisas na Internet — por exemplo, alguns dos links de origem do documento não puderam ser recuperados no momento da escrita, outros documentos podem ser recuperados em pesquisas futuras. Em vez disso, o PURE deve ser considerado como uma amostra dos requisitos que podem ser recuperados da Web. Inspecionamos informalmente cada documento e o rotulamos de acordo com os seguintes campos, além de outros adicionais, que fornecem algumas informações qualitativas de primeiro estágio.

- **Nome do Doc:** um ID alfanumérico que identifica o documento.
- **Páginas:** o número de páginas do documento.
- **Estrutura:** uma letra, ou combinações de letras, indicando como os requisitos são estruturalmente expressos. Pode ser: S = estruturado: se os requisitos são expressos em um formato estruturado, como, por exemplo, casos de uso; U = não estruturado: se os requisitos são expressos como descrições NL não estruturadas; O = uma declaração: se cada requisito é expresso em uma única declaração NL. Se formas mistas de expressar requisitos foram usadas – por exemplo, se no mesmo documento, encontramos requisitos estruturados (S) e

não estruturados (U) –, combinamos as letras com as + operador (ou seja, S + U).

- **Nível:** letra que indica o grau de abstração do requisitos. Pode ser H = requisitos de alto nível, ou L = requisitos de baixo nível. O julgamento foi subjetivo dado de acordo com a seguinte justificativa. Se mais foi necessário um refinamento do documento antes da sistema poderia ser implementado, rotulamos o documento com H. Se o conteúdo do documento estava pronto para implementação, nós a rotulamos com L. Análise sobre o o nível de requisitos individuais é deixado como trabalho futuro.
- **Fonte:** uma letra indicando se a fonte dos requisitos é uma Universidade (U) ou uma Organização Pública/Privada (I). Os documentos marcados com U normalmente incluem o caso estudos ou tarefas realizadas por estudantes universitários. Os documentos marcados com I incluem requisitos de força industrial, tanto provenientes de empresas privadas quanto de órgãos públicos, incluindo grupos de padronização.

Uma tabela completa que resume todos os requisitos documentos, e todos os campos estão disponíveis em nosso site. Aqui, mostramos algumas estatísticas relativas a parte do campos e estatísticas agregadas sobre o conteúdo NL automaticamente extraído dos documentos.

a) Páginas: Na Fig. 1, mostramos o número de páginas para cada documento. Temos um máximo de 288 páginas, um mínimo de 7 páginas, uma média de 47 páginas, com um padrão bastante elevado desvio de 45 páginas. Isso indica uma forte variabilidade de o conjunto de dados em termos de comprimento. Indicadores mais precisos de o comprimento dos documentos (por exemplo, número de requisitos) será fornecido quando todos os documentos serão formatados em XML.

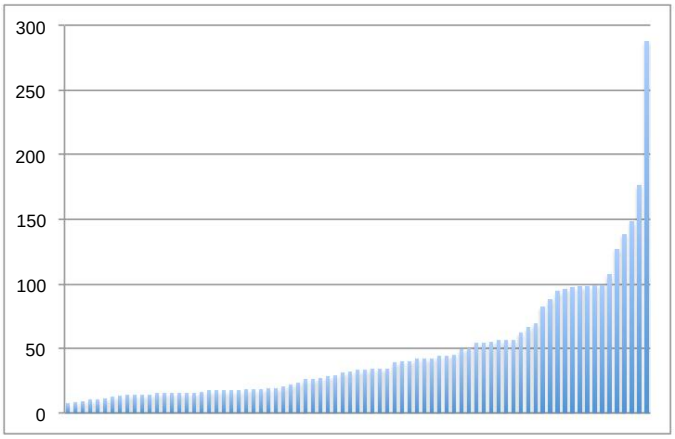

Fig. 1: Comprimento dos diferentes documentos em termos de páginas.

b) Estrutura: Na Fig. 2, mostramos as distribuições de as diferentes classes de estrutura. Vemos que a maioria de os documentos incluem uma combinação de conteúdo não estruturado e requisitos expressos em uma frase (U + O, 38%). Documentos com formatos uniformes – ou seja, U, S ou O – são igualmente distribuídos, com cerca de 15% dos documentos de cada turma. Menos representadas são as outras classes compostas. No entanto, a conjunto de dados já parece bastante geral e equilibrado para o que diz respeito à estrutura dos requisitos.

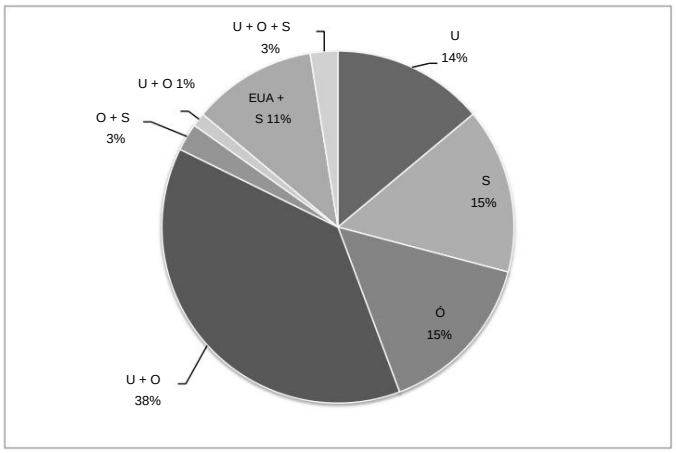

Fig. 2: Distribuição dos diferentes documentos em termos de Estrutura.

c) Nível e Fonte: Quanto ao nível dos requisitos – não mostrados nas imagens – temos uma dominância de requisitos de alto nível (H), com 71% dos documentos classificados com H, e 29% deles classificados com L. No geral, mais documentos de requisitos de baixo nível serão adicionados ao conjunto de dados para aumentar o equilíbrio. O conjunto de dados também é ligeiramente desequilibrado no que diz respeito à origem dos requisitos. Na verdade, temos 62% dos requisitos provenientes de Organizações Públicas/Privadas (I), e 38% de Universidades (U). São necessários requisitos industriais adicionais, uma vez que cada a empresa tem seu jargão específico [16] e, embora o conjunto de dados inclui documentos de empresas, não abrange todos os possíveis estilos de escrita.

d) Conteúdo NL: Extraímos automaticamente o texto de os diferentes documentos, e na Tabela I fornecemos alguns estatísticas agregadas sobre o conteúdo NL dos documentos. Para dar algum significado às nossas estatísticas, comparamos nosso corpus com o corpus clássico de Brown, que inclui 500 amostras de textos em língua inglesa [18]. Consideramos o corpus Brown como referência do inglês genérico, pois inclui texto de 15 gêneros, que vão do jornalismo à narrativa. No tabela, diferenciamos entre tokens (ou seja, todos os textos separados itens incluindo palavras e números, mas excluindo pontuação marcas) e palavras lexicais (ou seja, todos os termos, com exceção de números, sinais de pontuação e stopwords, como artigos, pronomes, etc.).

Os dois corpora têm um número total de tokens semelhante e palavras lexicais, o que implica que nossa análise é realizada entre corpora de tamanho comparável – também a distribuição de tokens, não mostrados aqui, são comparáveis. Por outro lado, o tamanho dos seus vocabulários diferem severamente tanto em termos de palavras lexicais (21.791 vs 46.018) e radicais, ou seja, raízes morfológicas (16.011 vs 29.846). Em particular, podemos dizer que, em documentos de requisitos, o vocabulário é cerca de metade de o vocabulário usado em textos genéricos. Isso também é indicado por o valor da diversidade lexical [19], calculado como o número de diferentes radicais de palavras únicas e o número total de palavras (0,031 vs 0,054). Se compararmos as palavras usadas nos dois

TABELA I: Estatísticas agregadas sobre o conteúdo NL dos documentos.

Indicador	Castanho PURO
Número de Tokens 865.551 1.034.378 Número de Palavras Lexicais 522.444 542.924	
Tamanho do Vocabulário (Palavras Lexicais) 21.791 46.018 Tamanho do	
Vocabulário (Raízes) 16.011 29.846 Número de Frases 34.268 57.340	
Comprimento Médio da Frase (Tokens) 25 18 Comprimento Médio da Frase	
(Palavras Lexicais) 15 10 Diversidade Lexical 0,031 0,054	

corpora, vemos que **62%** das palavras lexicais usadas em PURE não aparecem em Brown (valores não relatados na tabela).
Isso confirma que os documentos de requisitos usam um vocabulário específico [13], com siglas e termos específicos de domínio, o que difere muito do vocabulário comum em inglês.
Observando o comprimento das frases, vemos que as frases em PURE são sete tokens mais longas que as frases em Brown.
No geral, podemos dizer que os requisitos do corpus têm um vocabulário mais restrito e específico, mas frases mais longas, em relação aos textos genéricos.

Na Fig. 3, relatamos a frequência das palavras lexicais mais comuns no corpus. As palavras típicas relacionadas a requisitos, por exemplo, sistema, deve, dados, requisitos, usuário, software, especificação, etc. aparecem no topo desta lista. Também vemos que algumas siglas específicas de domínio, como npac (Number Portability Administration Center) e tcs (Train Control System, Telescope Control System, Tactical Control System) também aparecem na lista. A primeira sigla ocorre no maior documento do corpus (288 páginas), enquanto a última é uma sigla ambígua, ocorrendo em quatro grandes documentos de diferentes domínios. Essas observações indicam que (a) os documentos de requisitos usam uma terminologia peculiar que é comum a diferentes documentos, mas (b) também são caracterizados por expressões específicas de domínio, que são altamente frequentes nos documentos individuais. Esta análise sugere que as ferramentas de PNL treinadas em textos genéricos em inglês (por exemplo, analisadores estatísticos [20]) podem não ser adequadas para análise de requisitos, e personalizações adequadas podem ser necessárias.

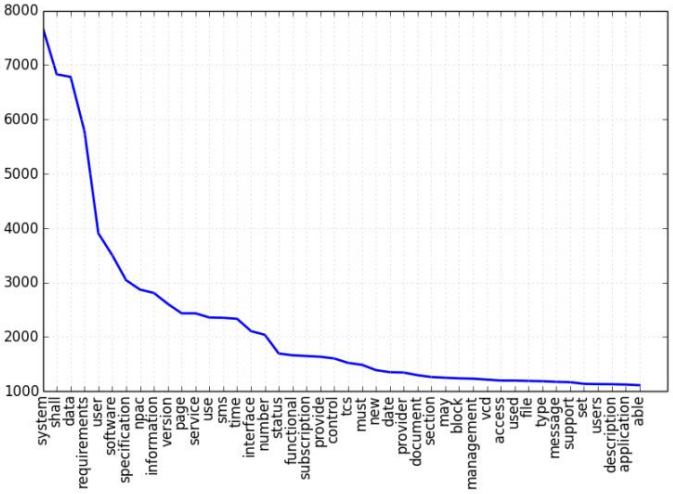

Fig. 3: Palavras mais frequentes no conjunto de dados geral.

```
< nome do elemento="p">
  <complexType misto="falso">
    <sequência>
      <elemento ref="title" minOccurs="0" maxOccurs="1"/> <choice maxOccurs="ilimitado">
      <elemento ref="p" minOccurs="0" maxOccurs="ilimitado"/
        > <elemento ref="req" minOccurs="0" maxOccurs="ilimitado"/> <elemento ref="b" minOccurs="0"
        maxOccurs="ilimitado"/> <elemento ref="glossário" minOccurs="0" maxOccurs="1"/> </choice> </
        sequence> <attribute name="id" type="string"/> </complexType> </element> [...] <elemento
        name="req"> <complexType mixed="false"> <sequence> <elemento ref="b"/> <elemento
      ref="modificador"
    minOccurs="0"
    maxOccurs="1"/>

      </sequence> < nome
      do atributo="id" tipo="string" uso="obrigatório"/> </complexType> </elemento> [...] < nome do
  elemento="req_document">
<complexType

misto="verdadeiro">

    <sequência>
      <elemento ref="title" minOccurs="1" maxOccurs="1"/> <elemento ref="version" minOccurs="1"
      maxOccurs="1"/> <elemento ref="issue_date" minOccurs="0" maxOccurs="1"/> <elemento
      ref="file_number" minOccurs="0" maxOccurs="1"/> <elemento ref="source" minOccurs="0"
      maxOccurs="1"/> <elemento ref="change_log" minOccurs="0" maxOccurs="1"/> <choice
      maxOccurs="unbounded"> <elemento ref="p"/> <elemento ref="req"/> </choice> </sequence>
      </complexType> </elemento>
```

Fig. 4: Trecho do arquivo XSD para representar documentos de requisitos.

III. ARQUIVOS XSD E XML

Definimos uma versão preliminar de um arquivo de esquema XML, ou seja, um arquivo XSD, com base em uma revisão dos documentos de requisitos reunidos. Um trecho simplificado do arquivo XSD é mostrado na Fig. 4. O elemento req_document (na parte inferior da figura) é o elemento raiz. Ele inclui vários elementos obrigatórios (minOccurs = “1”) e opcionais (minOccurs = “0”) que podem aparecer no início do documento de requisitos. Além disso, ele inclui um conjunto de parágrafos (p) e elementos de requisitos (req) (Fig. 4, superior).

Os parágrafos p são tipos complexos, que podem ter um título e podem incluir outros elementos em sequência. Esses elementos são outros parágrafos, elementos de glossário e requisitos ou elementos de corpo de texto – elementos complexos identificados com b na figura, mas chamados text_body no arquivo original.
Os requisitos req são sequências de elementos do corpo do texto b com modificadores (por exemplo, M = requisito obrigatório, O = re-requisito opcional). Tanto os parágrafos quanto os requisitos têm identificadores em formato de string (por exemplo, 1.1.a, 2.7.i, etc.). Todos os outros elementos que não são definidos aqui, mas aparecem na Fig. 4, são relatados em nosso arquivo original, junto com a definição de outros elementos úteis, como, por exemplo, listas, referências cruzadas e itens de glossário.

O XSD foi definido com base em uma revisão informal dos documentos coletados. Ele foi projetado para ser simples, claro e

suficientemente abrangente, mas, por outro lado, não podemos afirmar que todas as peculiaridades dos diferentes documentos são levadas em conta. Modificações no arquivo XSD serão realizadas junto com a portabilidade dos documentos. Representamos manualmente 12 documentos de acordo com o formato XSD.

Fomos fiéis à fonte, ou seja, não corrigimos formatação ou outros erros nos documentos. Para exemplos representativos, recomendamos olhar para 2007-ertms.xml e 2007-eirene_fun_7-2.xml, dois documentos de requisitos do domínio ferroviário disponíveis em nosso site.

IV. TEMAS DE PESQUISA HABILITADOS E LIMITAÇÕES

Nesta seção, discutimos as tarefas de PNL que podem ser executadas nos documentos atuais (ou seja, os tópicos de pesquisa habilitados pelo PURE), bem como as tarefas que exigem trabalho adicional no PURE antes que ele possa ser usado como referência.

Os documentos atuais podem ser usados como estão para pesquisa sobre identificação de abstração, avaliação de estrutura de documento e síntese de modelo. Pesquisa sobre identificação de abstração e extração de informação em geral pode alavancar os glossários disponíveis para parte dos documentos e comparar abstrações extraídas automaticamente com os termos dos glossários, como realizado, por exemplo, por Gacitua et al. [13]. Pesquisa sobre avaliação de estrutura pode comparar técnicas de extração de estrutura automatizada com a estrutura XML dos documentos, como em Ferrari et al. [10]. Pesquisa de síntese de modelo pode usar os diferentes documentos para produzir modelos de resumo gráfico a serem avaliados posteriormente por avaliadores humanos, como realizado, entre outros, por Robeer et al. [2].

Outras tarefas, como categorização de requisitos [21], detecção de ambiguidade [7] e identificação de requisitos equivalentes [11], exigem que anotações adicionais sejam realizadas pela comunidade de RE interessada em PNL. Para cada tarefa de RE específica, anotações manuais devem ser fornecidas para os requisitos, a fim de usar os documentos como conjuntos de treinamento, teste e validação, para algoritmos de aprendizado de máquina supervisionados ou como padrões ouro para algoritmos não supervisionados [20]. Na prática, categorias de requisitos funcionais/não funcionais devem ser fornecidas, bem como anotações de termos e frases percebidos como ambíguos e requisitos considerados equivalentes.

Para a tarefa de rastreabilidade, o conjunto de dados não é adequado no momento. Na verdade, para identificar traços de requisitos entre requisitos em diferentes níveis de abstração, como realizado, por exemplo, por Gervasi e Zowghi [22], precisamos de requisitos de alto e baixo nível pertencentes ao mesmo projeto, com links de rastreabilidade.
Portanto, para a tarefa de rastreabilidade, sugerimos consultar os benchmarks usados por outros autores (por exemplo, NASA CM1 [5], [22]).

V. CONCLUSÃO

Este artigo apresenta o PURE, um conjunto de dados para processamento de requisitos de linguagem natural. Ele é orientado para permitir a replicação de experimentos de PNL e generalização de resultados em RE. O conjunto de dados é atualmente composto por 79 documentos em vários formatos e 12 documentos que foram portados para um formato XML comum. Como trabalho futuro, estamos comprometidos em portar todo o conjunto de dados para XML, para enriquecer o conjunto de dados com outros recursos públicos.

documentos e compartilhá-los posteriormente em sites de hospedagem pública. Também esperamos que os pesquisadores de RE interessados em PNL anotem os requisitos para tarefas específicas e compartilhem os esquemas de anotação adotados, para que possam ser reutilizados. Conforme destacado em nosso trabalho anterior [17], APIs para manipular os arquivos XML estão em desenvolvimento, e a comunidade de RE também é encorajada a contribuir para o PURE com documentos adicionais.

REFERÊNCIAS

[1] M. Kassab, C. Neill e P. Laplante, “Estado da prática em engenharia de requisitos: dados contemporâneos”, Innovations in Systems and Software Engineering, vol. 10, no. 4, pp. 235–241, 2014.

[2] M. Robeer, G. Lucassen, JME van der Werf, F. Dalpiaz e S. Brinkkemper, “Extração automatizada de modelos conceituais de histórias de usuários via PNL”, em RE'16. IEEE, 2016, pp.

[3] A. Casamayor, D. Godoy e M. Campo, “Agrupamento funcional de requisitos de linguagem natural para assistência em projeto de software arquitetônico”, KBS, vol. 30, pp. 78–86, 2012.

[4] W. Maalej e H. Nabil, “Relatório de bug, solicitação de recurso ou simplesmente elogio? sobre a classificação automática de avaliações de aplicativos”, em RE'15. IEEE, 2015, pp. 116–125.

[5] H. Sultanov e JH Hayes, “Aplicação de aprendizagem por reforço à engenharia de requisitos: rastreamento de requisitos”, em RE'13. IEEE, 2013, pp. 52–61.

[6] J. Cleland-Huang, A. Czauderna, M. Gibiec e J. Emenecker, “Uma abordagem de aprendizagem de máquina para rastrear códigos regulatórios para requisitos específicos de produtos”, em ICSE (1). ACM, 2010, pp. 155–164.

[7] SF Tjong e DM Berry, “O design do SREE: um protótipo de localizador de ambiguidade potencial para especificações de requisitos e lições aprendidas”, em REFSQ'13. Springer, 2013, pp. 80–95.

[8] H. Femmer, DM Fernandez, S. Wagner e S. Eder, “Garantia rápida de qualidade com cheiros de requisitos”, JSS, vol. 123, pp. 190–213, 2017.

[9] C. Arora, M. Sabetzadeh, L. Briand e F. Zimmer, “Verificação automatizada de conformidade com modelos de requisitos usando processamento de linguagem natural”, IEEE TSE, vol. 41, no. 10, pp. 944–968, 2015.

[10] A. Ferrari, S. Gnesi e G. Tolomei, “Usando agrupamento para melhorar a estrutura de documentos de requisitos de linguagem natural”, em REFSQ'13. Português Springer, 2013, págs. 34–49.

[11] D. Falessi, G. Cantone e G. Canfora, “Princípios empíricos e um estudo de caso industrial na recuperação de requisitos equivalentes por meio de técnicas de processamento de linguagem natural”, IEEE TSE, vol. 39, no. 1, pp. 18–44, 2013.

[12] A. Ferrari, F. dellOrletta, GO Spagnolo e S. Gnesi, “Medindo e melhorando a completude dos requisitos de linguagem natural”, em REFSQ'14. Springer, 2014, pp. 23–38.

[13] R. Gacitua, P. Sawyer e V. Gervasi, “Sobre a eficácia da identificação de abstração na engenharia de requisitos”, em RE'10. IEEE, 2010, pp. 5–14.

[14] T. Quirchmayr, B. Paech, R. Kohl e H. Karey, “Extração de informações relevantes para recursos de software semiautomático de manuais de usuário em linguagem natural”, em REFS'17. Springer, 2017, pp. 255–272.

[15] X. Lian, M. Rahimi, J. Cleland-Huang, L. Zhang, R. Ferrari e M. Smith, “Mineração de conhecimento de requisitos de coleções de documentos de domínio”, em RE'16. IEEE, 2016, pp. 156–165.

[16] A. Ferrari, F. Dell'Orletta, A. Esuli, V. Gervasi e S. Gnesi, “Processamento de requisitos de linguagem natural: uma visão 4D”, IEEE Software (a ser publicado), 2017.

[17] A. Ferrari, GO Spagnolo e S. Gnesi, “Rumo a um conjunto de dados para processamento de requisitos de linguagem natural”, em REFSQ'17 Workshops, 2017.

[18] WN Francis e H. Kucera, “Manual do corpus Brown”, Brown University, vol. 15, 1979.

[19] D. Malvern, B. Richards, N. Chipere e P. Duran, “Diversidade lexical e desenvolvimento da linguagem”, Quantificação e Avaliação. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2004.

[20] CD Manning e H. Schutze, ¨ Fundamentos da estatística natural processamento de linguagem. MIT Press, 2003.

[21] A. Casamayor, D. Godoy e M. Campo, “Identificação de requisitos não funcionais em especificações textuais: uma abordagem de aprendizagem semi-supervisionada”, IST, vol. 52, no. 4, pp. 436–445, 2010.

[22] V. Gervasi e D. Zowghi, “Apoiando a rastreabilidade por meio da afinidade mineração”, em RE'14. IEEE, 2014, pp. 143–152.