Preparação para representação semântica em Língua Natural

Thais Rodrigues Neubauer[1], **Norton Trevisan Roman**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP – Brazil

{thais.neubauer, norton}@usp.br

Abstract. *To broadly identify studies related to Natural Language semantic representation, a Systematic Review (SR) was conducted on this topic. This article focuses on reporting the reasons for the choice of executing a SR, its characteristics, the way it has been conducted, along with its results.*

Resumo. *Com objetivo de mapear, com a maior abrangência possível, os estudos relacionados à representação semântica em Língua Natural, foi conduzida uma Revisão Sistemática (RS) ao redor desse assunto. Este artigo tem como foco relatar o motivo da escolha pela realização de uma RS, suas características, o modo como foi aqui conduzida e os resultados apresentados com sua execução.*

1. Introdução

A representação semântica de um texto ou diálogo consiste no processo de extrair o significado de suas expressões para apresentá-lo de um modo estruturado e legível a agentes não humanos. Nesse contexto, uma Revisão Bibliográfica Sistemática (RS) se apresenta como a melhor forma de se reunir, avaliar e interpretar todos os estudos disponíveis e relevantes para esse tópico. Segundo Kitchenham (2004), a RS acaba por também identificar as lacunas existentes, ou seja, verificar qual(is) ponto(s) de seu tópico necessita(m) de mais aprofundamento.

A diferença entre uma revisão bibliográfica comum e uma RS é, como o próprio termo sistemática sugere, a sistematização da realização da revisão, pois uma RS deve ser conduzida em uma sequência rigorosa, com passos definidos previamente por um protocolo [Biolchini et al. 2005]. Os estudos utilizados durante a condução de uma RS são denominados primários, enquanto a RS seria um estudo secundário [Kitchenham 2004].

Perante as vantagens da realização de uma RS, neste trabalho desenvolvemos uma com o objetivo de reunir os estudos sobre técnicas e modelos de representação semântica de textos e diálogos em Língua Natural.

2. Revisão Sistemática

A realização de uma RS prevê a definição rigorosa de um protocolo como fase inicial, pois esse deve ser seguido durante toda a revisão para garantir sua replicação e para que a adequação dos padrões escolhidos possa ser julgada adequada ou não para o tópico em questão [Biolchini et al. 2005]. A única tarefa anterior é a realização de uma pesquisa exploratória, que tem como objetivo exatamente auxiliar na definição das questões a serem respondidas e definidas no protocolo, além de verificar a realização de alguma possível RS já existente no assunto. Essa fase exploratória é uma pesquisa bibliográfica comum, na qual o revisor realiza pesquisas iniciais simples a título de verificar a melhor definição possível dos itens presentes no protocolo.

2.1. Protocolo da Revisão Sistemática

Após a pesquisa exploratória, o Protocolo da Revisão Sistemática foi elaborado, baseando-se no modelo proposto por Kitchenham (2004) e Biolchini et al (2005), compreendendo os seguintes pontos:

- **Objetivos:** (i) conhecer quais pesquisas foram feitas ou estão em andamento sobre a representação semântica de textos e diálogos em Língua Natural; (ii) conhecer as técnicas e modelos utilizados nessas pesquisas e os resultados dessas utilizações; (iii) identificar possíveis lacunas nos estudos da área para avançar os estudos de representação semântica adequadamente.
- **Questão de pesquisa:** "Quais técnicas e modelos são utilizados para representação semântica de textos e diálogos em Língua Natural?"
- **Fontes utilizadas:** *Google Scholar*[1]*, ACL Anthology Digital Archive*[2]*, IEEE Xplore Digital Library*[3] *e ACM Digital Library*[4].
- **Línguas dos Artigos:** a RS foi realizada abrangendo palavras-chave e artigos tanto na língua portuguesa quanto na língua inglesa.
- **Critérios de Inclusão:** para serem incluídos, os artigos deveriam: (i) apresentar técnicas ou modelos de representação semântica, ou (ii) avaliar técnicas ou modelos de representação semântica existentes, ou (iii) ao menos estarem inseridos no contexto de extração do significado de expressões em textos e diálogos e sua representação.
- **Critérios de Exclusão:** das pesquisas incluídas pelos critérios de inclusão, foram descartadas as que: (i) não estão disponíveis integralmente nas bases de dados, ou (ii) não estão em uma das línguas pesquisadas, ou (iii) não passaram pelo processo de revisão por pares.
- **Palavras-chave:** após a análise exploratória, foram identificados os seguintes termos: *semântica, significado, representação de conhecimento, representação de informação, representação semântica, análise semântica, computação semântica, processamento semântico, texto semântico e diálogo semântico*, além de seus correspondentes na língua inglesa: *semantics, meaning, knowledge representation, information representation, semantic representation, semantic analysis, semantic parsing, semantic computing, semantic processing, text semantics, dialog semantics e dialogue semantics.*

2.2. Condução da Revisão Sistemática

Posteriormente à elaboração do protocolo, iniciou-se a RS em si. Basicamente, seu desenvolvimento diz respeito à consulta de cada um dos termos selecionados pela pesquisa exploratória em cada uma das bases propostas no protocolo, verificando, através de análise e julgamento do revisor, se os resultados retornados respondem ou não a questão de pesquisa, sempre respeitando os critérios de inclusão e exclusão.

Para atingir uma amplitude significativa e, ao mesmo tempo, alcançável, considerando-se o volume de dados a serem analisados por conta da quantidade de

[1]http://scholar.google.com.br/
[2]http://www.aclweb.org/anthology/
[3]http://ieeexplore.ieee.org/
[4]http://dl.acm.org/

palavras-chaves selecionadas, foram analisados somente os 100 primeiros resultados para cada termo de busca. Garantindo a propriedade de replicação de uma legítima RS e seguindo os métodos constantes em sua proposta de execução, todos os resultados obtidos foram documentados, sendo classificados como relacionados ou não.

Primeiramente, foi realizada uma classificação prévia pelo título dos trabalhos. Caso algum desses fosse considerado irrelevante em relação à questão de pesquisa ou aos critérios de inclusão/exclusão, sua classificação já era definida como não relacionado e o estudo assim era reportado. Se, contudo, o trabalho foi considerado relevante, analisou-se seu abstract e, a partir dessa análise, novamente, conforme a adequação do trabalho junto à questão de pesquisa e aos critérios de inclusão/exclusão, decidiu-se por sua análise mais profunda ou não.

A partir da leitura dos trabalhos classificados como relacionados na fase de análise de título e abstract, foram identificadas as técnicas de representação semântica de textos e diálogos em Língua Natural utilizadas nesses artigos, com o objetivo de se chegar à resposta da pergunta inicial da RS. O processo de identificação das técnicas foi feito durante o período compreendido entre os meses maio e junho de 2014.

3. Resultados

Durante todo o processo de condução da RS foram analisados 4.508 artigos, recolhidos a partir das quatro bases especificadas no protocolo. A base com mais resultados para as buscas feitas foi a mais geral delas, a *Google Scholar* (36% dos artigos). A base *ACM Digital Library* foi a segunda a retornar mais trabalhos: 27% dos artigos analisados, seguida pela *ACL Anthology Digital Archive*, com 19%, e pela *IEEE Xplore Digital Library*, com 18%.

Durante o estágio em que foram classificados os artigos, a partir da análise de seu título e *abstract* - *Fase 1* - foram selecionados 477 trabalhos (10% do total), segundo os critérios definidos no protocolo. A partir desse momento, foram analisados os textos na íntegra desses 477 artigos, identificando as técnicas neles reportadas para representação semântica em Língua Natural. Nesse estágio, que aqui trataremos como Fase 2, ainda foram descartados outros 164 estudos, por não tangerem à questão de pesquisa (144) ou por suas versões integrais não estarem disponíveis (20).

Dos 213 artigos classificados como relacionados ou incluídos na Fase 2, 50 citaram a utilização de *Ontologias*; 39, a utilização de *Frames*; e 19, a utilização de *TreeBanks* (Figura 1). Vale mencionar que, no total, 75 técnicas foram identificadas, das quais 33 eram citadas em um único artigo e, por isso, não são apresentadas na Figura 1. Além disso, também é importante dizer que cada artigo podia identificar mais de uma técnica, distanciando então o número total de técnicas citadas do número de artigos pesquisados.

4. Conclusão

Este trabalho apresentou uma Revisão Sistemática acerca das técnicas utilizadas na representação semântica em Língua Natural. Com ela, foi possível determinar técnicas úteis para a identificação de vertentes não exploradas, ou mesmo para a identificação das técnicas de maior aceitação, caso se opte pela construção de sistemas que façam uso de representação semântica, como os que tratam da Web Semântica, por exemplo.

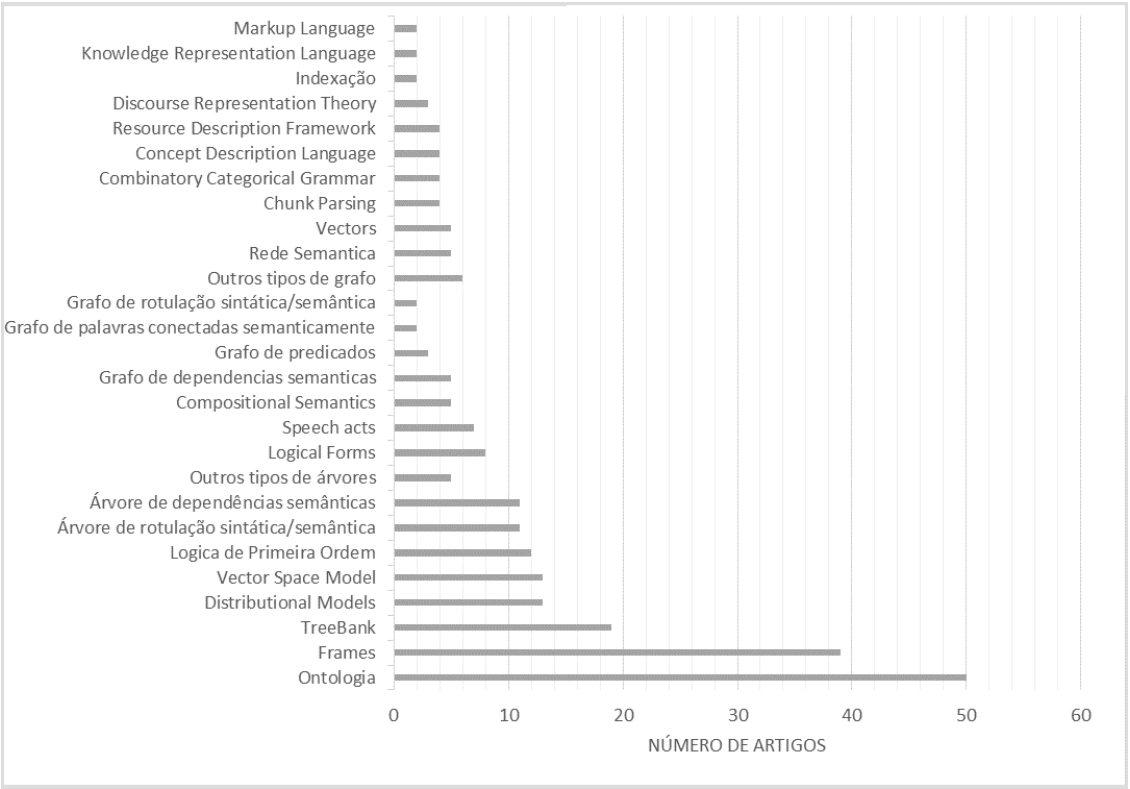

Figura 1. Técnicas identificadas por número de artigos que as citam.

5. Agradecimentos

Esta pesquisa contou com o apoio do Programa de Educação Tutorial (PET) – MEC/SESu e com a Pró-Reitoria de Graduação da Universidade de São Paulo.

Referências

Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering. Technical Report TR – ES 679 / 05, Systems Engineering and Computer Science Department, UFRJ, Rio de Janeiro.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE – 0401, Keele University, Keele, Staffs, UK.