FSI: FrameNet Semantic Infrastructure

Paulo Victor Hauck da Encarnação

paulovhe@gmail.com

Orientadora

Regina Maria Maciel Braga

regina@acessa.com

Área de Pesquisa

Engenharia de Software/Banco de Dados

Ano de ingresso no programa: *2012*

Época prevista de conclusão: *2014*

Resumo. *O projeto FrameNet busca documentar os frames linguísticos da língua Inglesa, possibilitando que esta informação seja utilizada por outras pessoas em diversas outras aplicações, como por exemplo em aplicações de PLN. Entretanto os dados são mantidos de forma que seja apresentável apenas a humanos, e caso seja desejado desenvolver algum projeto ou sistema utilizando estes dados é necessário solicitar uma cópia da base de dados aos mantenedores do projeto. Afim de promover facilidade ao acesso aos dados FrameNet tanto para humanos quanto para clientes de software, apresentaremos a FrameNet Semantic Infrastructure, uma proposta de infraestrutura apoiada no uso de ontologias, dados ligados e serviços web.*

Palavras-chave: *FrameNet, Ontologia, Dados ligados, Serviços, Semântica, Insfraestrutura, Frame.*

FSI: FrameNet Semantic Infrastructure

Paulo V. H. Encarnação, Regina M. M. Braga

Instituto de Ciências Exatas – Universidade Federal de Juiz de Fora (UFJF)

paulovhe@gmail.com, regina@acessa.com

Abstract. *The FrameNet project seeks to document the linguistic frames of the English language, allowing this information to be used by others in several other applications, as an example the use in PLN applications. However the data is maintained in a way that only humans are capable of reading, and if you want to develop a project or a system using these data is necessary to request a database copy to the project maintainers. In order to promote ease of data access from FrameNet for both humans and softwares clients, we present the "FrameNet Semantic Infrastructure" (FSI). The FSI infrastructure considers the use of ontologies, linked data and web services.*

Resumo. *O projeto FrameNet busca documentar os frames linguísticos da língua Inglesa, possibilitando que esta informação seja utilizada por outras pessoas em diversas outras aplicações, como por exemplo em aplicações de PLN. Entretanto os dados são mantidos de forma que seja apresentável apenas a humanos, e caso seja desejado desenvolver algum projeto ou sistema utilizando estes dados é necessário solicitar uma cópia da base de dados aos mantenedores do projeto. Afim de promover facilidade ao acesso aos dados do FrameNet tanto para humanos quanto para clientes de software, apresentaremos a FrameNet Semantic Infrastructure, uma proposta de infraestrutura apoiada no uso de ontologias, dados ligados e serviços web.*

1. Caracterização do Problema

O Projeto FrameNet (ICSI, 2012) é um esforço para documentação de frames linguísticos liderado por Charles Fillmore no International Computer Science Institute, em Berkeley - USA. Existem também outros projetos para documentação dos frames de outros idiomas, um deles é o FrameNet-Brasil (Fn-BR) desenvolvido pela Universidade Federal de Juiz de Fora (UFJF) com o foco na documentação dos frames da língua portuguesa (Gamonal, 2011). A base de dados mantida pela projeto tem diversas aplicações, se destacando entre estas aplicações o uso em Processamento de Linguagem Natural. Entretanto, a forma de acesso a estes dados atualmente está voltada ao uso por humanos, dificultando o acesso por ferramentas computacionais. Neste caso, uma cópia da base de dados do projeto pode ser solicitada ao ICSI (ICSI, 2012). Buscando facilitar o acesso as informações tanto para humanos quanto para maquinas, neste trabalho é proposto a FSI (FrameNet Semantic Infrastructure) uma infraestrutura voltada para o armazenamento das informações do projeto FrameNet baseados na tecnologia de Web Semântica.

2. Fundamentação Teórica

Buscando melhorar a estruturação dos dados na Web, Berners-Lee(2001) descreve um novo modelo da Web, denominado Web Semântica. Este modelo busca aplicar semântica as relações entre documentos e dados. Um dos fatores considerados essenciais para a concretização da Web Semântic é o emprego de ontologias. Apesar de ser definida de maneiras diferentes por diversos autores, uma ontologia é considerada essencialmente um conjunto de informações que definem a conceitualização dos termos que podem ser empregados a respeito de um certo domínio, definindo formalmente e de maneira explícita os elementos deste domínio (Noy e McGuinness, 2001). Ontologias normalmente são utilizadas para promover o compartilhamento de informações de maneira precisa, possibilitando assim aplicações como, reuso de conhecimento, explicitações de suposições sob um determinado domínio, descoberta de informações explícitas a partir de inferência, etc.

Outra proposta para aprimoramento da estruturação das informações disposta na Web foi a abordagem de Dados Ligados (Linked Data) apresentada em Berners-Lee (2006). Utilizando este padrão é possível construir uma teia semântica entre os objetos publicados. Como muitos provedores de dados passaram a adotar este padrão, acabaram surgindo muitos conjuntos de dados (*datasets*) ligados sobre diversos domínios. Sendo assim, a W3C iniciou um projeto chamado "Linked Open Data" (LOD) com o objetivo integrar estes *datasets* criando uma rede semântica.

A Semântica de frames descreve que o conhecimento humano não é construído a partir de um conceito isolado, mas sim a partir de um todo compartilhado. Buscando representar a maneira em que o conhecimento humano é estruturado, surge o frame. Os frames constituem um sistema complexo de conceitos relacionados de maneira que para se compreender um deles é necessário entender toda a estrutura na qual o frame se encaixa (Fillmore, 1982). O FrameNet é um projeto desenvolvido pelo International Conputer Sciences Institute, que trata-se do desenvolvimento de uma base de dados. Atualmente o banco de dados do FrameNet conta com aproximadamento 1160 frames contendo 1675 relações entre estes, com 12600 unidades lexicais (ISCI, 2012). Estes dados podem ser utilizados em diversas aplicações, como por exemplo na rotulação semântica automática de textos, possibilitando desenvolvimento em vários setores de Processamento de Linguagem Natural (PLN), entre eles a tradução automática, extração de informações, sumariamento de texto, etc. Além disso o projeto FrameNet também impulsionou a criação de vários outros projetos baseado na semântica de frames. Temos entre eles o FrameNet-BR, que trata de uma adaptação do projeto FrameNet entretanto com base na língua Portuguesa do Brasil. Temos também o *Kicktionary* (Schmidt, 2009) que busca construir um dicionário temático no domínio do futebol nas línguas Inglesa, Alemã e Francesa. E de maneira semelhante ao *Kicktionary* temos o "Copa 2014 FrameNet Brasil" (Gamonal, 2011) que busca construir um dicionário trilingue (português, inglês e espanhol) no domínio do Futebol e do Turismo.

3. Caracterização da contribuição

Apresentamos a seguir a "FrameNet Semantic Infrastructure" (FSI), uma infraestrutura que tem como objetivo facilitar o acesso às informações mantidas pelo FrameNet, possibilitando o seu uso de diversas formas, principalmente em aplicações em PLN sem

a necessidade de solicitar e manter uma cópia da base de dados do FrameNet. Além disto, a adoção do formato de dados ligados para representação das anotações dos frames, possibilita o uso de informações de outras bases de Dados Ligados como anotações. Utilizaremos na FSI serviços com suas descrições em metadados, permitindo assim que seja feita a descoberta dos serviços disponíveis pelas ferramentas sem a necessidade de intervenção humana. Na Figura 1 temos uma representação geral da FSI e seus elementos.

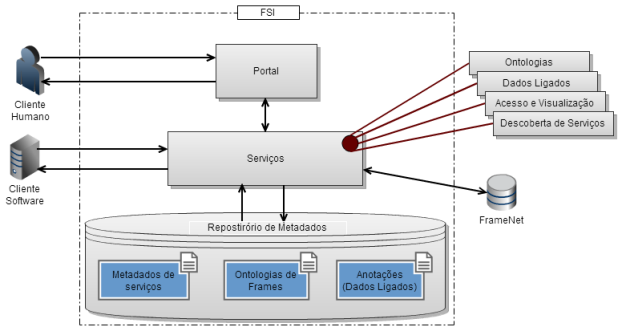

Figure 1. Visão geral da infraestrutura Semântica de Frames.

De maneira semelhante ao trabalho de Scheffczyk et. al. (2008) é utilizada uma ontologia de Frame para definir um esquema de transformação dos dados nativos da base do FrameNet para uma ontologia. A partir da importação dos dados do frame para a ontologia suas anotações serão transformadas em Dados Ligados e também armazenadas no repositório de metadados.

4. Estado atual do trabalho

Foram estudados uso de ontologias como a que foi usada por Scheffczyk et. al.(2008), e a ontologia de Nuzzolese et. al.(2011). Entretanto após análise destas ontologias foram detectadas estruturas que não correspondem a realidade dos arquivos do FrameNet e/ou algumas divergências a respeito da interpretação do seu autor, sendo assim foi desenvolvida uma ontologia própria a partir de uma análise dos arquivos da base do FrameNet.

Após o desenvolvimento desta ontologia, foi iniciado o desenvolvimento dos serviços web fazendo uso da linguagem JAVA para codificação. O primeiro serviço desenvolvido foi o serviço que faz a importação dos dados da base do FrameNet para a ontologia de Frames, utilizando o Apache Jena como API para manuseio de ontologias, e o JDOM para leitura dos arquivos XML. A fim de tornar o processo de importação o mais automatizado possível, um arquivo de esquema dos XML presente na base do FrameNet é lido inicialmente para definir uma equivalência entre as estruturas das informações com os elementos da ontologia de frames. Este serviço encontra-se atualmente em desenvolvimento.

5. Comparação com trabalhos relacionados

Alguns trabalhos na literatura foram identificados como relacionados ao tema deste trabalho. Entre eles temos o trabalho de Scheffczyk et. al.(2008) que propõe a construção de ontologias em OWL-DL a partir da transcição das informações

expressadas pelos frames do FrameNet com o objetivo de permitir, por exemplo, o refinamento das restrições de elementos de frame a partir de seus tipos semânticos. Como exemplo, a restrição para o EF "turista" do frame de Turismo, sendo restrito a este EF a associação de equivalência a membros classificados como um ser consciente em uma ontologia que defina seres vivos, ou seja, o ser humano é um ser consciente assim, qualquer instancia de sua classe pode ser um turista na anotação do frame de Turismo. A partir destas ontologias é possível também a criação de uma ontologia de anotações, a partir da qual o autor aponta a possibilidade de execução de um *reasoner* para identificação e extração de informações.

No trabalho de Nuzzolese et al.(2011), primeiramente cada frame é transformado em um conjunto simples de triplas RDF pela ferramenta Semion fazendo uso de uma ontologia que defina a estrutura forma de um frame, e em seguida é possível que esta base RDF gerada pela primeira etapa, seja alinhada a uma ontologia de domínio também pela ferramenta Semion, transformando estes dados em uma base RDF que expressa o conhecimento que está representado na base de origem na visão do domínio a qual foi aplicado. Segundo o autor, o uso da ferramenta requer um conhecimento prévio sobre o domínio ao qual a base RDF será alinhada, uma vez que é necessário que o usuário defina regras de transformações para que o alinhamento seja feito. Como resultado deste processo pode-se obter dois formatos para a base resultante, o de dados ligados ou de Knowledge Patterns.

Além destes temos a APRI (Vegi et.al., 2011) que é uma infraestrutura que faz uso de serviços Web e uma representação de metadados para a especificação dos padrões de análise e de seus serviços. O objetivo da APRI é promover o reuso de padrões de análise criando um ambiente compartilhado para armazenamento e recuperação dos padrões de análise por humanos ou maquinas.

Utilizando os conceitos utilizados por Vegi et. al.(2011) a FSI será construída utilizando a arquitetura baseada em serviços e a descrição destes por metadados a fim de facilitar o acesso destes também por clientes de software. Entretando a FSI é voltada para o armazenamento dos dados do FrameNet de uma maneira semântica, com uso de uma ontologia de estruturação formal do frame para transformação dos dados. Assim como apontados nos trabalhos de Scheffczyk et. al.(2008) e Nuzzolese et al.(2011) buscamos prover uma interface de acesso a estes dados permitindo a interpretabilidade por humanos e maquinas.

Entretanto uma das principais diferenças da FSI é a criação de um espaço único de acesso e manutenção dos dados de frames eliminando a necessidade de se manter uma cópia da base para cada ferramenta que faz uso destes dados, sendo assim eliminamos também a possibilidade de divergência entre as informações captadas por ferramentas distintas uma vez que possibilitamos a captura de dados de uma fonte única. Outra importante diferença é a adoção de Dados Ligados como formato para as anotações, o que permite que outros *datasets* de dados ligados sejam incorporados ou utilizados como fontes para novas anotações.

References

Berners-Lee, T.; Hendler, J. ; Lassila, O. (2001). The semantic web. In *Scientific American*, v.284, n.5, pages 34–43.

Berners-Lee, T. (2006). Linked Data - Design Issues. http://www.w3.org/DesignIssues/LinkedData.html. Acessado em: 15/09/2012.

Fillmore, C. J. (1982). Frame semantics, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Gamonal, M. A. (2011). Copa 2014 Framenet Brasil: Análise da unidade lexical "visitar" do frame de turismo. In *Anais do SILEL*.

ICSI (2011). The framenet project. https://framenet.icsi.berkeley.edu/fndrupal/, 2012. Acessado em: 15/09/2012.

Lage, L. M. (2011). Frames e construãções: Um estudo de caso da construção. In: *Anais do SILEI*.

Noy, N. F.; McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Development, v.32, n.1, pages 1–25.

Nuzzolese A. G., Gangemi A., and Presutti V. (2011). Gathering lexical linked data and knowledge patterns from FrameNet. In *Proceedings of the sixth international conference on Knowledge capture (K-CAP '11)*. ACM, New York, NY, USA, p 41-48.

Scheffczyk, J.; Baker, C. F. ; Narayanan, S. (2008). Ontology-Based reasoning about lexical resources. In *Ontologies and Lexical Resources for Natural Language Processing, Cambridge Studies in Natural Language Processin*g. Cambridge University Press, Cambridge, MA.

Schmidt, T. (2009). The Kicktionary - a multilingual lexical resource of football language. In: BOAS, Hans. (Ed.). *Multilingual FrameNets - Methods and Applications*. Berlin/New York: Mouton de Gruyter.

Vegi, L. F.; Peixoto, D. A.; Soares, L. S.; Filho, J. L. ; Oliveira, A. P. (2011). An infrastructure oriented for cataloging services and reuse of analysis patterns. In *Lecture Notes in Business Information Processing*, pages 338–343. Springer.