Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Uma Proposta Metodológica para a Categorização Automatizada de Atrações Turísticas a partir de Comentários de Usuários em Plataformas Online

Vanessa Maria Ramos Lopes Paiva[1], Tiago Timponi Torrent[1]

[1]FrameNet Brasil – Programa de Pós-Graduação em Linguística
Universidade Federal de Juiz de Fora (UFJF)
Rua José Lourenço Kelmer s/nº - Campus Universitário
36036-900 – Juiz de Fora – Minas Gerais – Brasil

vanessaletrasufjf@gmail.com, tiago.torrent@ufjf.edu.br

Abstract. *Some useful information for planning a trip should go beyond what is available in some travel guides, that is, they should bring specific information to the tourist. Considering this need, the present work aims to present a recommendation system for tourist attractions based on semantic information extracted from tourist comments on online platforms, which will go beyond the basic information (museum, restaurant), presenting specific information (if it is a suitable place for children, for example). The system performs a semantic analysis of tourists' comments on the Internet, using a knowledge base from relevant semantic frames and associated lexical items.*

Resumo. *Algumas informações úteis para planejar uma viagem devem ir além do que está disponível em alguns guias de viagem, ou seja, devem trazer informações específicas para o turista. Considerando essa necessidade, o presente trabalho tem como objetivo apresentar um sistema de recomendação para atrações turísticas, o qual irá além das informações básicas (museu, restaurante), apresentando informações específicas (se é um local adequado para crianças, é acessível). O sistema executa uma análise semântica dos comentários de turistas na Internet, usando uma base de conhecimentos a partir de frames semânticos relevantes e itens lexicais associados.*

1. Introdução

Planejar uma viagem ou atividade de lazer requer diferentes tipos de informações sobre uma atração turística. Muitos guias de viagem podem auxiliar trazendo informações sobre os locais, como chegar, o que fazer ou até mesmo a temperatura em determinada época do ano. Do mesmo modo, essas ferramentas costumam focar em atrações de destaque ou informações mais gerais que auxiliam um planejamento básico de uma viagem. Entretanto, os guias de viagem não trazem informações específicas que muitos turistas podem precisar ao planejar uma viagem, tais como qual atração é melhor para um dia chuvoso ou que museu é interessante para crianças. Essas informações são subjetivas e são sujeitas a alterações.

Embora muitas informações estejam disponíveis em plataformas online na forma de comentários e avaliações postadas por usuários, ler todos eles é tarefa incompatível com o dinamismo de uma viagem. Considerando esse contexto, uma análise automática desses comentários poderia gerar informações mais úteis ao turista, em especial, se

disponibilizadas em uma plataforma interativa e dinâmica. Não se trata apenas de extrair se a impressão geral sobre uma determinada atração é positiva ou negativa, tarefa já clássica em Processamento de Língua Natural (PLN), mas de ir além delas, trazendo informações mais específicas que auxiliem o usuário a tomar decisões.

O trabalho se desenvolve no âmbito do projeto m.knob do Laboratório FrameNet Brasil de Linguística Computacional da Universidade Federal de Juiz de Fora. Tal projeto está desenvolvendo um assistente pessoal de viagem na forma de um *chatbot* com o qual os turistas podem interagir usando língua natural com vias a obter recomendações de atrações e atividades.

Nesse contexto, este trabalho tem como objetivo propor uma metodologia de categorização automatizada para atrações turísticas baseada na informação semântica extraída de comentários de turistas em plataformas online. Tal metodologia prevê a existência de um analisador que extrairá a informação semântica dos comentários e a traduzirá em um *cluster* de frames. O sistema também irá gerar *clusters* a partir dos inputs do usuário e, posteriormente, mapeará as semelhanças entre os *clusters*, sugerindo atrações e atividades turísticas que possam aderir aos interesses do usuário.

Na próxima seção, apresentamos os conceitos fundamentais da Semântica de Frames, teoria que dá sustentação ao desenvolvimento da FrameNet Brasil e da base m.knob, enquanto as seções seguintes estão assim distribuídas: a seção 3 traz a proposta metodológica, a seção 4 apresenta uma análise de exemplo em Português e a seção 5 apresenta as considerações finais.

2. Semântica de Frames

Para a Linguística Cognitiva, o significado das palavras associa-se às experiências sociais e culturais, uma vez que essa perspectiva teórica considera a linguagem como parte da cognição. Desse modo, em suas interações, o falante/ouvinte constrói o significado através de conceptualizações, com base em suas experiências. Nesse contexto, Fillmore (1982) propõe um modelo segundo o qual o significado das expressões é perspectivizado, ou seja, o significado é construído através de pontos de vistas diferentes e não de uma única maneira objetiva. Partindo desse princípio, ao se propor uma pesquisa a partir da Semântica de Frames, compreende-se que a dimensão do significado é expressa a partir de estruturas cognitivas – frames – os quais os falantes utilizam para expressar o entendimento de sua língua[Fillmore & Baker 2010].

A Semântica de Frames é o estudo de como as formas linguísticas evocam ou ativam frames e de como os frames ativados podem ser integrados no entendimento de sentenças [Fillmore & Baker 2010]. Desse modo, a Semântica de Frames oferece um entendimento de como o significado emerge a partir de cenas. Um exemplo são os verbos *comprar* e *vender* que evocam cenas com perspectivas diferentes, isto é, em *comprar*, a cena é perspectivizada pelo comprador, enquanto que em *vender*, o é pelo vendedor. Percebe-se que, a partir dessas diferentes perspectivas, um frame é construído cognitivamente na interação, contribuindo para a compreensão de uma expressão.

A partir da Semântica de Frames, surgiu a FrameNet, um projeto lexicográfico computacional, que extrai informações sobre propriedades semânticas e sintáticas de palavras do inglês, através de um grande corpus eletrônico [Fillmore 2003a].A FrameNet identifica e analisa os frames evocados nas sentenças anotadas para o

sistema, buscando estudar que propriedades – sintáticas e semânticas – se instanciam nelas.

2.1.FrameNet Brasil

A partir da FrameNet de Berkeley, outras FrameNets surgiram ao redor do mundo e uma delas é a FrameNet Brasil. Ela vem se desenvolvendo desde 2007 e se baseia na Semântica de Frames para a análise de sentenças no Português [Torrent & Ellsworth 2013]. Considerando a Semântica de Frames como a base teórica da FrameNet Brasil, os principais pilares analíticos dessa iniciativa são frames, elementos de frame (EFs), unidades lexicais (ULs) e anotação lexicográfica.

Assim como em outras FrameNets, a FrameNet Brasil considera em sua análise: (a) um conjunto de frames, compostos por elementos de frame (EFs) e (b) as unidades lexicais (ULs), palavras que evocam frames. Essas categorias são consideradas na anotação lexicográfica de sentenças do Português. Esse tipo de anotação considera a unidade lexical (UL) como ponto central no processo de anotação, porém, outros constituintes da sentença também são considerados nessa tarefa [Fillmore, et al. 2003]. Esses constituintes, tais como nomes, verbos e advérbios são anotados em camadas separadas, as quais identificam os EFs, que podem ser nucleares e não-nucleares, a Função Gramatical (FG) e o Tipo Sintagmátco (TS). A título de exemplo, considere-se o frame `Chegar`, reproduzido na Figura 1.

Figura 1. Exemplo de frame na FrameNet Brasil

A partir da definição do frame, sentenças contendo as ULs que o evocam, nesse caso, *aportar.v, chegar.v* entre outras, podem ser anotadas e o produto dessa anotação gera padrões de valência nos quais cada UL pode se instanciar. Para a composição de tais padrões, são atribuídas, para cada EF, etiquetas relativas à FG – tais como "Ext" (Externo), "Obj" (Objeto), "Dep" (Dependente) e "Quant" (Quantificador) – e ao TS –

NP (Sintagma Nominal), PP (Sintagma Preposicionado), VInf (Verbo Infinitivo), entre outros.

2.2. Multilingual Knowledge Base (m.knob)

A FrameNet Brasil vem desenvolvendo um repositório de frames multilínguesde domínio específico chamado Multilingual Knowledge Base (m.knob). A primeira versão do banco de dados do m.knob, desenvolvida para os Jogos Olímpicos do Rio 2016,apresentava 52 frames e mais de 2000 unidades lexicais, modelando os domínios do turismo [Gamonal & Torrent 2015] e dos esportes [Costa 2017] em três línguas: Português, Espanhol e Inglês. Passados os jogos, a base de dados está sendo expandida para modelar de maneira mais detalhada os vários aspectos da experiência turística, incluindo-se aqueles que não se distinguem, ao menos do ponto de vista terminológico, de modelos lexicais genéricos.

Esses frames podem ser entendidos como modelos de sistemas conceituais que representam tanto eventos – tais como fazer uma reserva – como entidades – tais como atrações urbanas e naturais. Cada frame é composto por uma definição e um conjunto de elementos que definem a cena que funciona de pano de fundo para o significado da unidade lexical que evoca o frame [Fillmore 1982]. A Figura 2 apresenta o frame de `Serviço_turístico_reservar`, evocado pelas unidades lexicais *reserva.n* e *reservar.v*, conforme consta na base de dados da FrameNet Brasil.

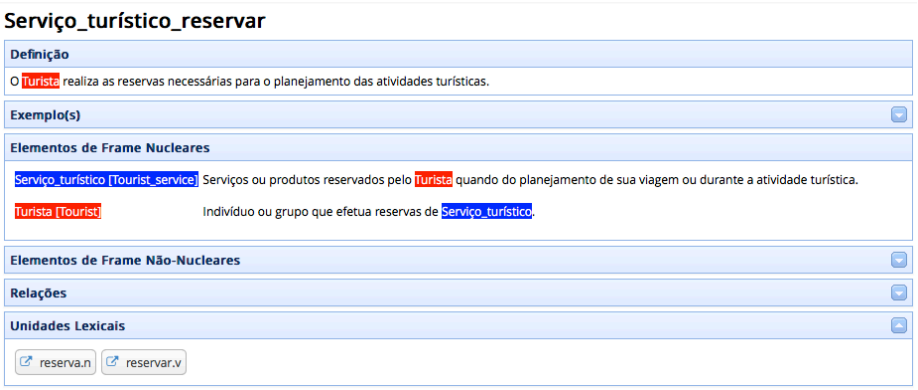

Figura 2: Frame Serviço_turístico_reservar

No m.knob, assim como nas demais FrameNets, os frames se relacionam entre si em uma rede, conforme Figura 3. As relações podem indicar se um frame é um sub-evento dentro de um evento complexo (setas azuis), se ele apresenta uma perspectiva específica sobre uma cena neutra (setas rosas indicando as perspectivas do turista e do prestador do serviço turístico), se ele ocorre necessariamente antes de outro frame (seta preta) ou se ele pressupõe algum outro frame (setas verdes).

Os frames do m.knob incluem tanto eventos e entidades específicos do turismo, quanto outros de domínio geral, uma vez que ambos são necessários para o entendimento dos comentários dos turistas sobre as atrações. Itens lexicais tais como *museu.n, visitar.v, restaurante.n* representam interesses primários do turistas. Além desses interesses, o turista pode necessitar de informações secundárias como se o lugar é acessível, se é bom para dias chuvosos, entre outros. Dessa forma, o banco de dados do

m.knob inclui frames relevantes para descrever tanto interesses primários, ou seja, os tipos de eventos e entidades tipicamente turísticos,quanto secundários, ou seja, características e propriedades que, em princípio, se aplicam a quaisquer eventos ou entidades.

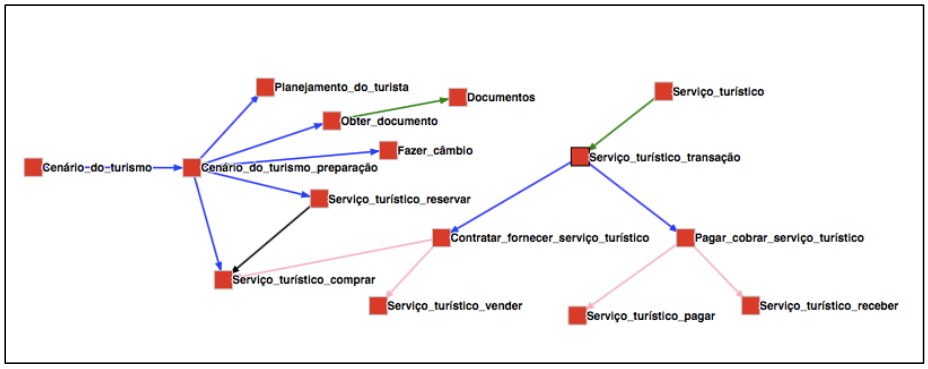

Figura 3: Extrato da Rede de frames do m.knob

Nesse sentido, a base atual expandida do m.knob conta com 331 frames, que incluem tanto frames modelados pela FrameNet Brasil, quanto frames que já constavam da base de dados de vocabulário genérico da Berkeley FrameNet [Fillmore et al. 2003].

3. Proposta Metodológica

Embora a cultura colaborativa da internet tenha trazido avaliações subjetivas sobre atrações turísticas através de ferramentas diversas, isso ainda não é suficiente para que o usuário possa aproveitar essas informações, dada a impossibilidade de ler todas as avaliações postadas. Dessa forma, o projeto proposto nesse trabalho supera essas limitações através de uma base de conhecimento multilíngue que modela o domínio do turismo e de um categorizador algorítmico que usa essa base de conhecimento para gerar representações semânticas detalhadas de atrações turísticas.

Com base no categorizador algorítmico, o sistema irá verificar os comentários postados e extrairá o significado das palavras candidatas. Em um primeiro estágio, é reunido o conjunto de frames evocados nos comentários. Em seguida, os frames evocados serão pesados quanto à sua frequência nos dados. Em uma terceira etapa, os agrupamentos de frames que representam cada lugar serão derivados e armazenados no banco de dados do m.knob. Como objetivo final dessa pesquisa, uma interface de usuário será desenvolvida, onde o turista informará, através de uma interface conversacional, o que ele gostaria de fazer, usando língua natural. No estágio final, o sistema fornecerá ao turista recomendações de lugares classificados de acordo com os resultados de um processo de correspondência entre a representação semântica gerada para a sentença do usuário e aquelas geradas para as atrações a partir da análise dos comentários. Uma visão geral do sistema de categorização é apresentada na Figura 4.

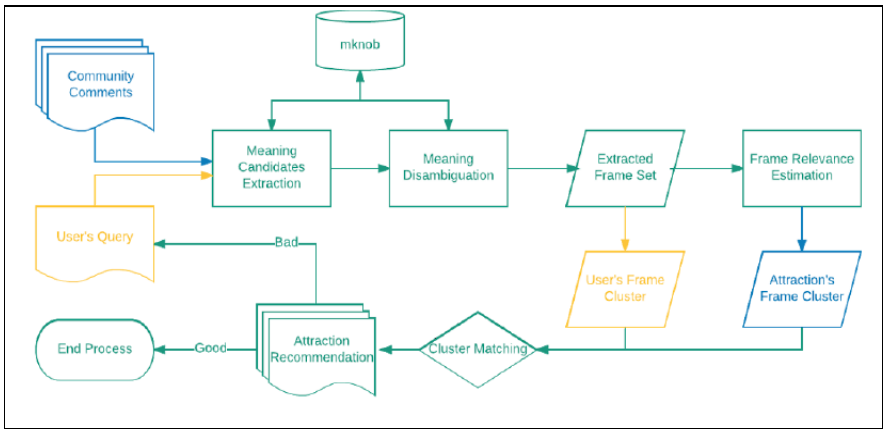

Figura 4. Visão geral do sistema de categorização. Os elementos em azul representam os processos atuantes sobre comentários disponíveis em plataformas online; aqueles em amarelo, os atuantes sobre a entrada do usuário, e aqueles em verde as partes do sistema que trabalham sobre ambos.

4. Análise de Exemplos

Para o estudo piloto, um corpus que conta com 3495 comentários sobre 939 locais em São Francisco (EUA), foi extraído do Google Places API. Uma versão alfa do analisador semântico identificou os frames potencialmente evocados por cada comentário. Dentre os 50 frames mais evocados, 24 se referiam a interesses secundários, 20 se referiam a um vocabulário mais geral, não sendo necessariamente interesses turísticos e 6 eram menos importantes.

Em relação à natureza e aos exemplos dos frames evocados a partir dos comentários dos usuários, entre os frames que se referiam a interesses primários, destacam-se `Locais_por_uso` (com ULs como *museu.n, igreja.n* e *praça.n*) e `Locais_naturais` (*praia.n* e *vale.n*). Já entre os secundários, destacam-se `Parentesco` (*filho.n, avô.n, irmão.n*), `Pessoas_por_idade` (*criança.n, idoso.a*), `Custo` (*caro.a, barato.a*), `Foco_no_estímulo` (*lindo.a, majestoso.a*), entre outros. Esses dados sugerem como os frames evocados nos comentários dos turistas, podem auxiliar na representação semântica detalhada de atrações turísticas, contribuindo para o fornecimento de dados para o aplicativo.

A partir dessa proposta metodológica, nesta seção, será apresentada uma análise exemplar para o Português. Usaremos como exemplo a sentença em (1), um comentário sobre a Universidade Federal de Juiz de Fora, extraído da plataforma Google Local Guides:

(1) A UFJF é uma boa instituição de ensino e um ótimo lugar para passar as tardes de sábado e domingo, praticando esportes e atividades ao ar livre, como vôlei e peteca. Além disso, é frequente a ocorrência de eventos noturnos, como shows.

Aplicando-se a esta sentença o procedimento de indicação dos frames evocados por cada UL constante da base da FrameNet Brasil, temos a anotação proposta em (2).

(2) A UFJF é uma [boa^{Avaliar}] [instituição de ensino^{Locais_por_uso}] e um [ótimo^{Ser_desejável}] [lugar^{Local}] [para^{Finalidade}] [passar^{Estada_temporária}] as [tardes^{Unidades_calêndricas}] de [sábado^{Unidades_calêndricas}] e [domingo^{Unidades_calêndricas}], [praticando esportes^{Atividades_de_lazer}] e [atividades ao ar livre^{Atividades_de_lazer}], como [vôlei^{Esporte}] e [peteca^{Esporte}]. Além disso, é [frequente^{Frequência}] a [ocorrência^{Evento}] de [eventos noturnos^{Evento}], como [shows^{Artes_performáticas}].

Como se pode notar, a grande maioria das palavras e expressões que constituem o comentário encontram correspondência em algum frame. Considere-se ainda que, apenas para a Universidade Federal de Juiz de Fora, a plataforma do Google Local Guides conta com 224 comentários de usuários. Isso posto, fica demonstrada a riqueza dos dados que podem ser extraídos e compilados na plataforma m.knob a partir da web.

5. Considerações Finais

O presente artigo teve como objetivo apresentar uma metodologia de categorização automatizada para atrações turísticas baseada na informação semântica extraída de comentários de turistas em plataformas online. Além disso, buscou-se demonstrar como a Semântica de Frames pode auxiliar na interpretação semântica de dados, através dos frames evocados em cada sentença. Outro ponto de destaque foi a apresentação do repositório de frames multilíngues de domínio específico chamado Multilingual Knowledge Base (m.knob) que a FrameNet Brasil vem desenvolvendo. A partir da apresentação do m.knob, buscou-se apontar como essa ferramenta pode auxiliar na extração de informações semânticas de comentários de usuários, através de uma proposta metodológica de categorização automatizada.

Esse estudo inicial sugere que o banco de dados do m.knob tenha cobertura suficiente de interesses turísticos primários e secundários para apoiar a extração das informações semânticas necessárias para o sistema de categorização proposto.

Referências

Costa, A. D. (2017). *A Tradução por Máquina Enriquecida Semanticamente com Frames e Estruturas Qualia*. Qualificação. (Progressão ao Doutorado) – Universidade Federal de Juiz de Fora. Juiz de Fora, p.139.

Fillmore, C. J. (1982). Frame Semantics. In: Linguistic Society of Korea (Eds.), *Linguistics in the morning calm* (pp. 111 – 137). Seoul: Hanshin.

Fillmore, C. J. Johnson, R. C., Petruck, M. R. L. (2003a) Background to FrameNet. *International Journal of Lexicography,* 16 (3), p. 235-250.

Fillmore, C. J., Petruck, M. R., Ruppenhofer, J., & Wright, A. (2003). FrameNet in action: The case of attaching. *International Journal of Lexicography*, *16*(3), 297-332.

Fillmore. C. J.; Baker, C. (2010) A frames approach to semantic analysis. In: HEINE, B. & HEIKO, N. (Eds.). *The Oxford Handbook of Linguistic Analysis*. New York: The Oxford University Press, p. 313-339.

Gamonal, M.; Torrent, T. T. (2015). Diretrizes para a criação de um recurso lexical multilíngue a partir da semântica de frames: a experiência turística em foco. Domínios de Lingu@gem 9, p. 56-75.

Torrent, T. T. & Ellsworth, M. (2013) Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. In: *Veredas*: Frame Semantics and its technological applications. Juiz de Fora: UFJF, v. 17, p.44-65.