

Área Temática 17:

Linguística Computacional

Descrição de construções interrogativas QU- em Português Brasileiro para o desenvolvimento de um Chatterbot Institucional: o projeto LOURDES

Autores: Natália Duarte Marção ¹, Ely Edison da Silva Matos ¹, Tiago Timponi Torrent ¹

Instituição: ¹ UFJF - Universidade Federal de Juiz de Fora, ² UFJF - Universidade Federal de Juiz de Fora, ³ UFJF - Universidade Federal de Juiz de Fora

Resumo: Este trabalho tem por objetivo apresentar a descrição e a modelagem linguístico-computacional (DIAS-DA-SILVA, 1996) das construções interrogativas QU- em Português Brasileiro com vias a sustentar o desenvolvimento de um *chatterbot* institucional para a Universidade Federal de Juiz de Fora, no âmbito do projeto LOURDES – Localizador Unificado via ReDE Semântica –, desenvolvido no Laboratório FrameNet Brasil de Linguística Computacional. Uma vez que as interações planejadas para o *chatterbot* LOURDES se relacionam à extração de informações sobre os procedimentos da universidade, a descrição e modelagem das construções QU- se faz necessária para que se possa garantir o acesso à informação buscada pelo usuário através de uma interface de pesquisa baseada em compreensão de língua natural. No primeiro estágio da pesquisa, o qual será reportado neste trabalho, realizou-se, primeiramente, levantamento bibliográfico acerca das construções QU- em Português Brasileiro. Feito o levantamento, passou-se a uma proposta de descrição das construções QU- com base na Gramática das Construções de Berkeley (KAY & FILLMORE, 1999), através da elaboração de Matrizes de Atributo e Valor. Posteriormente, tais análises foram convertidas em modelos linguístico-computacionais armazenados na base de dados da FrameNet Brasil, na forma de construções. Para tanto, utilizou-se a ferramenta FrameNet Brasil WebTool 2.0. Até o presente momento, foram modeladas sete construções interrogativas QU-, quais sejam: Cxn. Interrogativa QU-, a qual é mais abstrata e é herdada pelas demais construções; Cxn. Interrogativa Que; Cxn. Interrogativa Qual; Cxn. Interrogativa Quem; Cxn. Interrogativa Quando; Cxn. Interrogativa Onde e Cxn. Interrogativa Prep-Que, a qual licencia subcasos como “por que”, “para que”, etc. **Referências** DIAS-DA-SILVA, Bento Carlos. A Face Tecnológica dos Estudos da Linguagem: o processamento automático das línguas naturais. 274 p. Tese de doutorado – USP. Araraquara 1996. KAY, Paul; FILLMORE, Charles J. Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. In: Language, p. 1-33, 1999.

Palavras-chave: Construções interrogativas QU-, Gramática de Construções, Chatterbot, FrameNet Brasil

I can tell you are stressed: modeling data from Brazilian Portuguese for lexical stress detection in ASR using KALDI

Autores: Simone Harmath-de Lemos ¹

Instituição: ¹ Cornell - Cornell University

Resumo: The present work trains the automatic speech recognition system (ASR) KALDI (Povey et al., 2011) to detect lexical stress in Brazilian Portuguese (BP). Lexical stress, or the relative prominence of portions of a word, merits the attention of Natural Language Processing, especially in the fields of Automatic Speech Recognition and Computer Assisted Language Learning (CALL), as numerous works have shown that, crosslinguistically, stressed syllables are associated with attracting nuclear intonational tones (Hayes, 1995), defining phonological rules, cuing speakers for syntactic ambiguity resolution (Gravina & Fernandes-Svartman, 2013), and pointing to the presence and nature of empty categories (Nunes & Santos, 2009). In second language acquisition (SLA), studies have shown that deviance in prosody by non-native speakers of English has a strong effect in intelligibility rating (Anderson-Hsieh, Johnson & Koehler, 1992). The work is innovative in which: KALDI is trained using data models labeled according to linguistic theory on lexical stress in BP, sparing the costly process of having a human labeler, and of building an additional module to process the labeled data; a relatively small lexicon is used in the training, from a read-speech corpus, the West Point Brazilian Portuguese Speech LDC2008S04 (Morgan, Ackerlind & Sterling, 2008). Moreover, a finely grained phonetic transcription is used in the phonetic dictionary, and monosyllables are included in the first pass of the training. KALDI is subsequently trained with more complex models of the data to rule out optimistic outcomes. Results show overall accuracy rate of detection comparable to previous works that use human labelers, ranging from 80.5% to 85.6% for the different models, 87.7% to 90.6% when

monosyllables are excluded. Using this model, the next step is underway to train KALDI to detect stress shift in contexts of stress clash, using data from an experimental study.

Palavras-chave: lexical stress detection, automatic speech recognition, training data

Caderno de resumos do X Congresso Internacional da ABRALIN – Pesquisa linguística e compromisso político. / Organizadores: Anabel Medeiros de Azerêdo; Beatriz dos Santos Feres; Patrícia Ferreira Neves Ribeiro; Roberta Viegas Noronha; Silmara Dela Silva. Niterói: UFF, 2017.
Disponível em: <<http://abralin.org/congresso2017/programacao-1?prog=simposios>>.