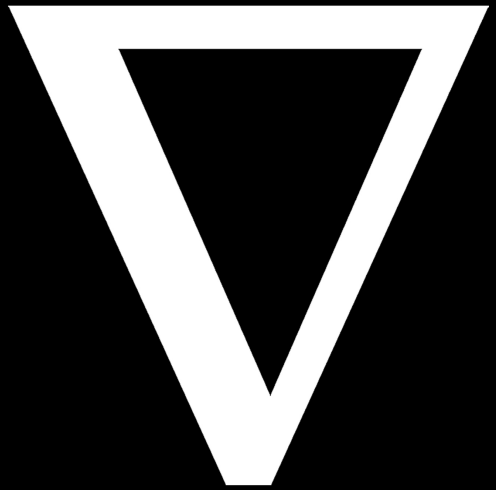


PLN na fala

Overview dos caps 1, 2 e 3 do livro Brasileiras em PLN

Por Artur Lima

Cap 1

Cap 1

O PLN se divide em duas grandes subáreas:

- Interpretação (ou Compreensão) de Linguagem Natural – NLU (do inglês, *Natural Language Understanding*)
- Geração de Linguagem Natural – NLG (do inglês, *Natural Language Generation*)

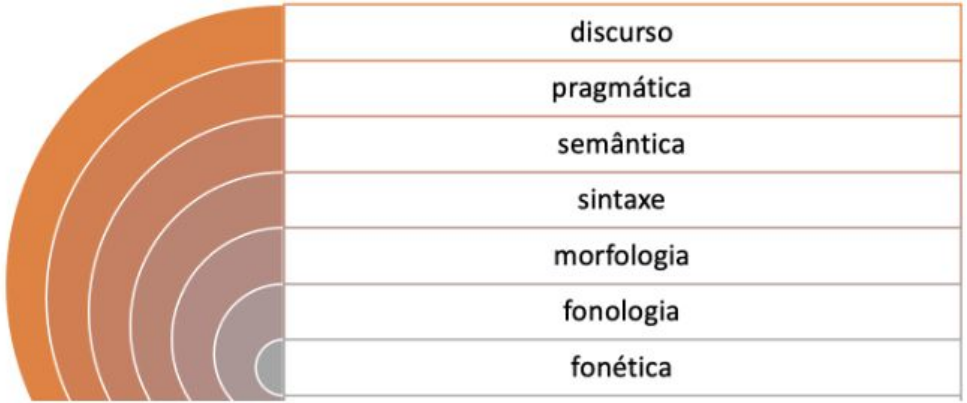
Assim NLU envolve a segmentação e classificação dos componentes linguísticos assim como interpretar o significado do texto, assim podemos citar como exemplos a compreensão de um texto por um chatbot, corretor do word ou até a extração de emoções.

Já o NLG, geração de conteúdos sejam eles textos ou áudios.

Fiquei em dúvida em qual categorias tarefas como resumo e tradução se encaixam.

Cap 1

- **Fonética** e **fonologia**: sons e sua organização
- **Morfologia** : morfemas e formação de palavras
- **Sintaxe**: organização das palavras na frase
- **Semântica** e pragmática: significado e finalidade do uso das palavras
- **Discurso** : Texto como um todo, relação entre frases distintas, parágrafos...



Cap 2 e 3

Termos importantes

- ***Part-of-speech* ou PoS**: categoria gramatical de uma palavra.
- **Frequência fundamental(f0):** altura de um som, diferença entre fino e grave
- **Forma espectral:** timbre

Termos importantes

- **RNN:** Rede neural recorrente, possui uma memória da sequência ao contrário das redes normais.
- **transformer**: honestamente ,adoraria que a gente discutisse um pouco que eu mesmo não peguei ainda, mas aqui está a definição do livro:
  “De forma resumida, diferentemente das RNN, nos Transformers, os vetores de entrada e de saída têm o mesmo tamanho e cada bloco de atenção tem acesso às entradas dos blocos anteriores”

Resumo breve

- **Reconhecimento de fala(ASR):** consiste em conseguir transcrever uma fala em texto
- **Síntese de fala:** processo inverso, gerar um áudio correspondente a certo texto
    - **Segmentação prosódica:** Trabalha com a tonicidade, força das sílabas das palavras
- **Reconhecimento de emoções:** com base na duração e intensidade da fala caracterizá-la quanto às suas emoções

Menções honrosas

- **reconhecimento de locutor e verificação de locutor**
- **restauração (ou aprimoramento) da fala**
- **diarização da fala**
- **modelagem de tópicos a partir das transcrições dos áudios**
- **clonagem de voz**


ASR

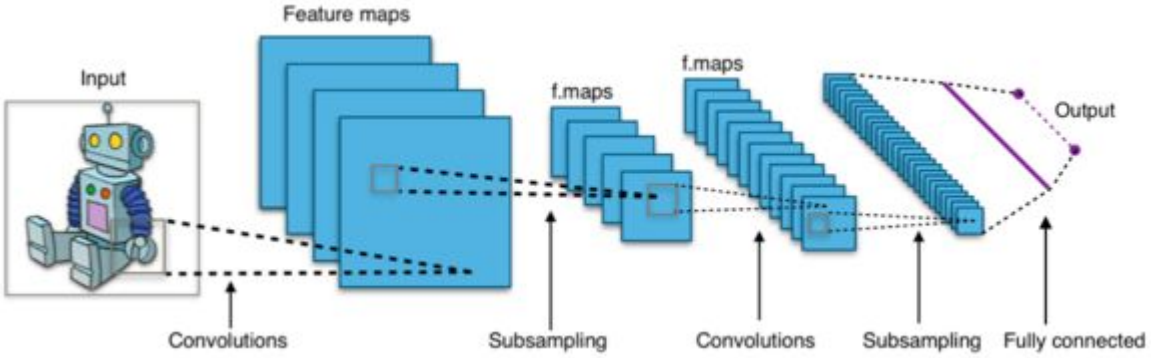
Reconhecimento de fala(ASR)

Consiste na transformação do sinal acústico de um trecho de fala em um trecho de texto. Como isso é feito, analisa amplitude das vibrações do ar, com uma FFT(fast fourier transform) obtém a frequência e a pressão correspondente, que serão as entradas para os modelos.

Antigamente(paradigma estatístico híbrido) usava-se o HMM(Hidden Markov Model), que consistia na junção de três modelos, um para extrair fonemas dos sons, um para transformar os fones em palavras e por fim um lexical com dicionário de pronúncias para corrigir erros gramaticais. Bons no geral, mas sofriam para estrangeirismos por trabalharem com n-gramas.

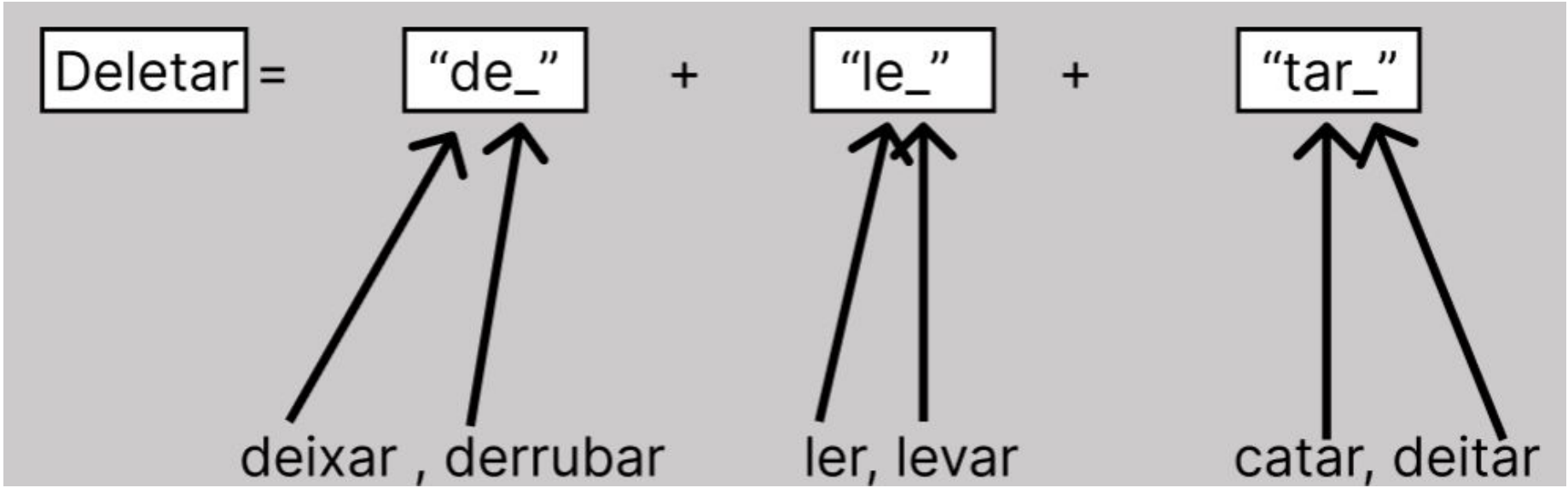
Reconhecimento de fala(ASR)

Atualmente, uma junção de transformers com CNNs é muito utilizada, com os transformers trabalhando o contexto mais global e as CNNs o local, eles conseguem até representar neologismos com facilidade por meio do uso de wordpieces, construindo palavras novas com partes de outras.



Fonte wikipedia

Reconhecimento de fala(ASR)



Métricas de avaliação

- ***Word Error Rate* (WER):** Percentual de palavras que foram transcritas erradas
- ***Sentence Error Rate* (SER):** Percentual de frases que continham erro de transcrição.
- Sugestão de um **terceiro erro**, que atribui **peso** aos erros de acordo com a **importância** da palavra.

Problemas enfrentados no ASR

Falta de dados, e aqui entra

- Ausência de transcrição da fala
- Diferentes condições de gravação(com ou sem ruído)
- Ausência de locutores diferentes
- Ausência de dialetos diferentes

Os modelos de ponta precisam de muitos dados, o que não há em português brasileiro

Problemas enfrentados no ASR

Coarticulação

- Quando fonemas interagem entre si, pode haver o suprimento uns dos outros, formando por exemplo em dialeto mineirês “você sabe se esse ônibus passa na Savassi”, passível como “cêsasessonspasansavas”
- Isso ocorre também em casos envolvendo encadeamento de sons a e o, como em “Mande um beijo par**a a A**manda” que pode facilmente virar “Mande um beijo par**a A**manda”

Problemas enfrentados no ASR

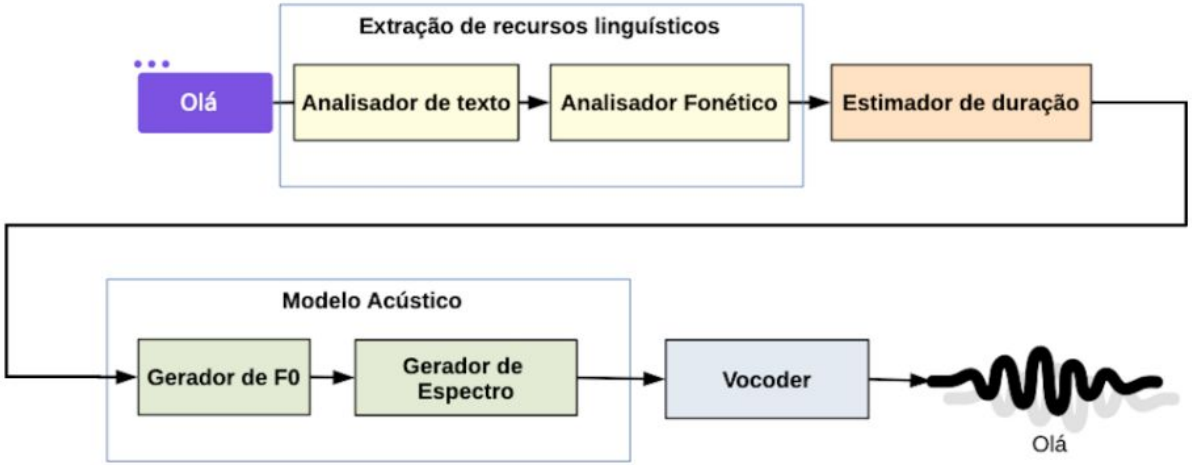
Pontuação

- Embora pareça simples, a pontuação é importante para facilitar a leitura humana dos textos transcritos
- Dificultada pois diferenças sutis como tom de voz, podem indicar pontuações diferentes, falta de dados e pouco modelos treinados com isso

Síntese de fala

Síntese de fala

Consiste na transformação de uma fala em um som. Esse processo é feito em duas etapas, uma primeira para normalizar o texto e transformá-lo em fonemas, e outra para produzir esses sons com a devida tonicidade(prosódia)



Normalização

Processo em que transformamos símbolos e abreviações em sua forma falada

Texto de entrada	Texto normalizado
1990	mil novecentos e noventa
68,3%	sessenta e oito vírgula três por cento
Av.	avenida
km^2	quilômetros quadrados
2:45PM	Duas e quarenta e cinco da tarde

Normalização

No português, surgem problemas como questão de gênero e número e se siglas são faladas ou soletradas, como IFSC e ICMC

Isso pode ser feito definindo regras, que com base nos tokens ao redor chega-se em uma resposta, ou via redes neurais estilo codificador-decodificador, mas que precisam de um grande volume de dados

Conversão grafema-fonema

Historicamente, a conversão das letras para sons se deu pelo uso de regras descritas de modo a mapear as letras do alfabeto para o som correspondente a ela, de acordo com o contexto em que tal letra aparece, pelo que eu entendi parecido com o HMM(Hidden Markov Model) do ARS.

Atualmente, redes neurais como Tacotron 2 se mostraram mais eficientes

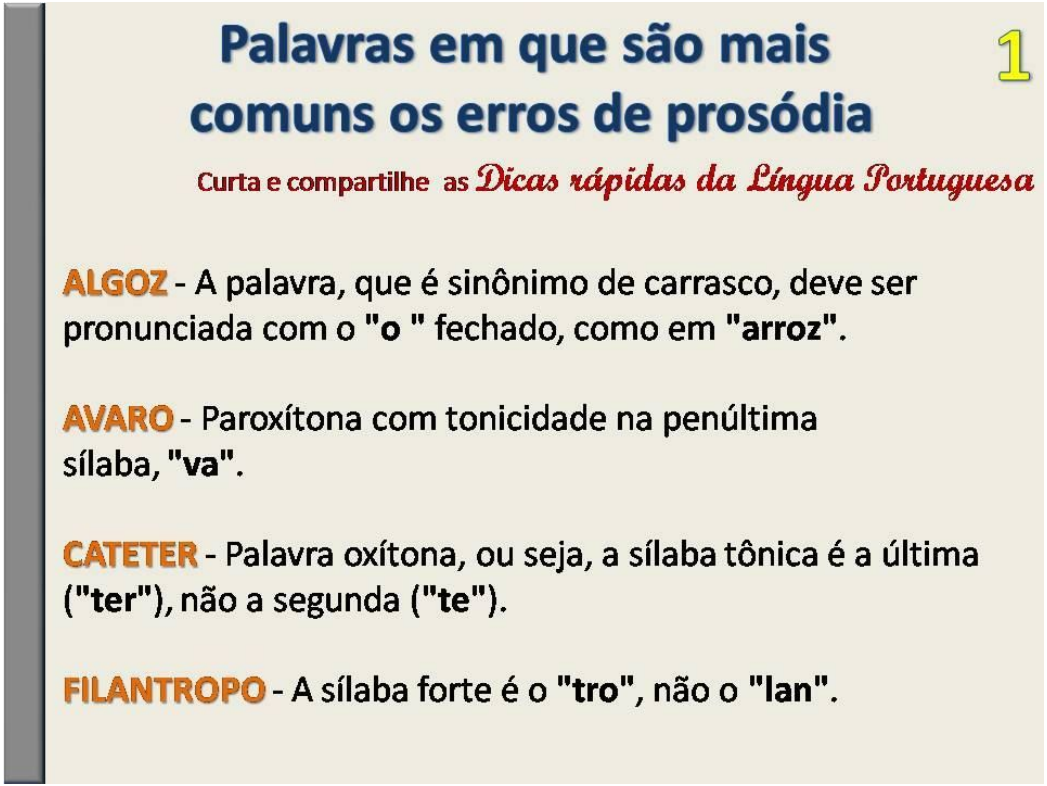
Métricas de avaliação

Diferente do ASR que quantiza o número de erros e acertos, a avaliação para geração de vozes consiste em humanos falantes nativos do idioma dando notas de 1 a 5 para as transcrições, o Mean Opinion Score (MOS), o que gera uma avaliação subjetiva dos modelos.

Problemas enfrentados na síntese de fala

- Falta de dados com variedade de locutores para possibilitar a geração de fala com vários locutores ou falta de dados mesmo para redes neurais mais robustas
- Prosódia, ou seja, descobrir qual a sílaba tônica de uma palavra para pronunciá-la corretamente. Muitos conjuntos de dados não possuem anotações prosódicas dificultando o processo

Segmentação prosódica



Fonte facebook

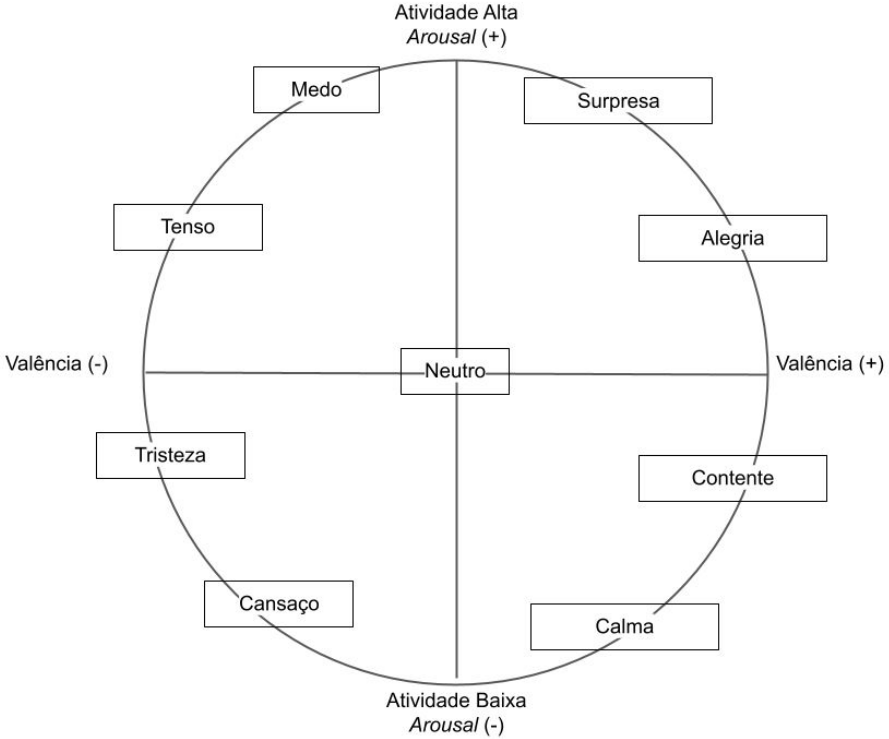
Reconhecimento de emoções

Não é muito trabalhado nos capítulos, mas a ideia geral é retirar de aspectos do áudio como a duração e a intensidade da fala para enquadrá-la em emoções, que são obtidas aplicando a transformada de *Fourier* em janelas de áudio.

Com essas informações do áudio conseguimos atribuir certos valores de valência e ativação para colocar a frase em um espaço bidimensional com os valores típicos de cada emoção já presentes.

Exemplo retirado do livro nota-se maior intensidade vocal na emoção “alegria”, enquanto a “tristeza” costuma ter intensidade vocal reduzida

Reconhecimento de emoções



Ideias de projetos

Ideias de projetos reconhecimento de voz, clonagem de voz, transcrição de voz para anotar, com correções gramaticais e tudo mais ou talvez até resumo