VII SIMAC
SIMPÓSIO ACADÊMICO
FACULDADE ENG. SALVADOR ARENA

AUTOMAÇÃO DE TRANSCRIÇÃO E ANÁLISE DE REUNIÕES UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Matheus Barbosa Meloni, matheus.b.meloni@gmail.com, Universidade Presbiteriana Mackenzie

Victor Inacio de Oliveira, victor.inacio@mackenzie.br, Universidade Presbiteriana Mackenzie

Resumo

Este artigo apresenta uma solução de automação para a transcrição e análise de reuniões utilizando modelos de Inteligência Artificial (IA). O objetivo é otimizar o registro e a organização de conteúdos discutidos em ambientes acadêmicos e corporativos, reduzindo a dependência de transcrições manuais, que são demoradas e suscetíveis a erros. A proposta envolve a captação de áudio, transcrição automática com o modelo Whisper, segmentação semântica do conteúdo por meio de embeddings e geração de relatórios estruturados. Espera-se que a solução proporcione maior eficiência, precisão e acessibilidade no acompanhamento e consulta de reuniões, beneficiando equipes de pesquisa, gestores e organizações que dependem de documentação confiável.

Palavras-chave: Reconhecimento de Fala; Processamento de Linguagem Natural; Transcrição Automática; Inteligência Artificial.

Introdução

A documentação de reuniões é uma prática essencial em contextos corporativos e acadêmicos, garantindo que informações estratégicas, decisões e compromissos sejam registrados de forma confiável. No entanto, a execução manual dessa atividade é trabalhosa, consome tempo e pode apresentar falhas de interpretação. O avanço das técnicas de Reconhecimento Automático de Fala (ASR) e Processamento de Linguagem Natural (PLN) possibilitou o surgimento de soluções capazes de transcrever e organizar grandes volumes de dados de maneira automatizada.

Modelos recentes, como o Whisper, desenvolvido pela OpenAI, demonstraram robustez na transcrição de áudios em cenários adversos, incluindo ambientes com ruído de fundo e sotaques variados, alcançando resultados próximos ao desempenho humano (RADFORD et al., 2022). Essa característica torna a tecnologia especialmente adequada para aplicações em reuniões corporativas e acadêmicas, nas quais fatores acústicos frequentemente dificultam a precisão da documentação.

No campo da representação semântica, Mikolov et al. (2013) introduziram o modelo Word2Vec, que permite representar palavras em um espaço vetorial contínuo, capturando regularidades sintáticas e semânticas. Essas representações tornam possível, por exemplo, calcular relações linguísticas complexas, como “rei – homem + mulher ≈ rainha”, demonstrando a capacidade dos embeddings de preservar relações semânticas latentes.

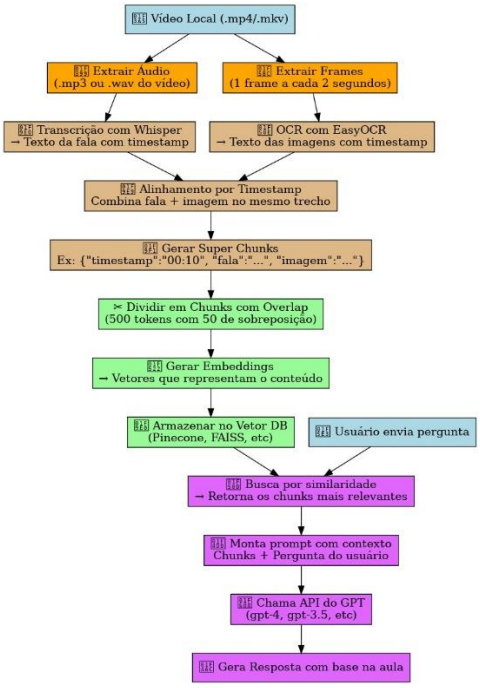
Além disso, estudos como o de Alemi e Ginsparg (2015) evidenciam que algoritmos de segmentação textual baseados em embeddings superam métodos tradicionais, permitindo cortes mais naturais e coerentes no

texto. Essa abordagem, aplicada a transcrições de reuniões, possibilita a organização automática de tópicos, decisões e pendências, tornando os registros mais úteis para consulta posterior.
Portanto, ao integrar modelos robustos de transcrição automática com técnicas de embeddings e segmentação semântica, este trabalho propõe uma solução inovadora para otimizar o registro e a análise de reuniões. Como defendem Jurafsky e Martin (2023), a combinação de reconhecimento de fala e processamento semântico constitui um caminho essencial para a evolução de sistemas práticos e escaláveis em Processamento de Linguagem Natural.

Metodologia

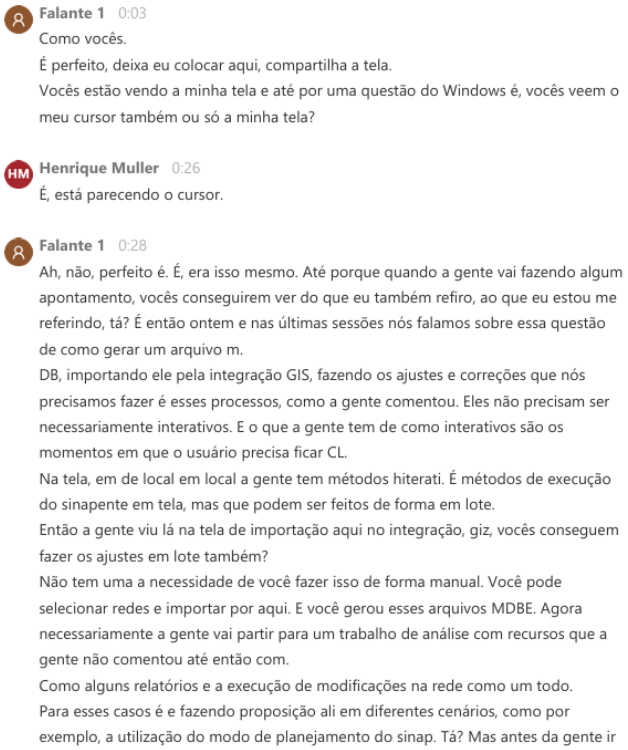
O sistema de automação de transcrição e análise de reuniões vem sendo desenvolvido de forma modular, o que tem garantido maior escalabilidade e facilidade de manutenção. A primeira etapa consistiu na captação de aulas e videos gravados de reuniões utilizando microfones e dispositivos convencionais, de modo a coletar dados em diferentes condições de ambiente. Esses áudios foram processados com o modelo **Whisper**, da OpenAI, que demonstrou robustez no reconhecimento de fala mesmo em cenários com ruído de fundo e sotaques diversos, alcançando resultados próximos ao desempenho humano (RADFORD et al., 2022). Essa etapa viabilizou a conversão automática dos sinais de voz em texto, substituindo com eficiência as transcrições manuais que demandavam tempo e estavam sujeitas a erros.

Figura 1: Fluxograma de teste com uma aula gravada com inputs de um usuário para futuros questionamentos



Fonte: Próprio Autor

Figura 2: Transcrição exportada em pdf da aula gravada



Falante 1 0:03
Como vocês.
É perfeito, deixa eu colocar aqui, compartilha a tela.
Vocês estão vendo a minha tela e até por uma questão do Windows é, vocês veem o meu cursor também ou só a minha tela?

Henrique Muller 0:26
É, está parecendo o cursor.

Falante 1 0:28
Ah, não, perfeito é. É, era isso mesmo. Até porque quando a gente vai fazendo algum apontamento, vocês conseguirem ver do que eu também refiro, ao que eu estou me referindo, tá? É então ontem e nas últimas sessões nós falamos sobre essa questão de como gerar um arquivo m.
DB, importando ele pela integração GIS, fazendo os ajustes e correções que nós precisamos fazer é esses processos, como a gente comentou. Eles não precisam ser necessariamente interativos. E o que a gente tem de como interativos são os momentos em que o usuário precisa ficar CL.
Na tela, em de local em local a gente tem métodos hiterati. É métodos de execução do sinapente em tela, mas que podem ser feitos de forma em lote.
Então a gente viu lá na tela de importação aqui no integração, giz, vocês conseguem fazer os ajustes em lote também?
Não tem uma a necessidade de você fazer isso de forma manual. Você pode selecionar redes e importar por aqui. E você gerou esses arquivos MDBE. Agora necessariamente a gente vai partir para um trabalho de análise com recursos que a gente não comentou até então com.
Como alguns relatórios e a execução de modificações na rede como um todo.
Para esses casos é e fazendo proposição ali em diferentes cenários, como por exemplo, a utilização do modo de planejamento do sinap. Tá? Mas antes da gente ir

Fonte: Próprio Autor

Na etapa seguinte, os textos resultantes passaram por processos de análise e segmentação temática. Para isso, têm sido empregadas técnicas de embeddings semânticos, capazes de representar sentenças em vetores numéricos de alta dimensionalidade (MIKOLOV et al., 2013). Essas representações permitiram organizar o conteúdo transcrito de acordo com a proximidade semântica entre trechos. Seguindo os métodos propostos por Alemi e Ginsparg (2015), a segmentação baseada em embeddings mostrou-se eficaz para estruturar automaticamente os tópicos discutidos nas reuniões, superando abordagens mais tradicionais baseadas apenas em palavras-chave.

Os resultados desse processo foram organizados em relatórios estruturados, compostos por seções temáticas e resumos automáticos. Para sustentar essa arquitetura, adotou-se uma divisão em camadas: a camada de usuário, responsável pela captação do áudio; a camada de interface, encarregada de exibir as transcrições e relatórios em formato navegável; a camada de controle, que atua como intermediária entre interface e algoritmos; e a camada de inteligência artificial, que reúne os módulos de transcrição e segmentação semântica (JURAFSKY; MARTIN, 2023).

A validação da metodologia tem ocorrido por meio da comparação entre transcrições automáticas e transcrições manuais realizadas pelos integrantes do estudo. Essa avaliação contínua tem permitido medir

a eficiência do modelo, considerando indicadores como tempo de execução, precisão das transcrições e consistência na segmentação temática

Resultados preliminares

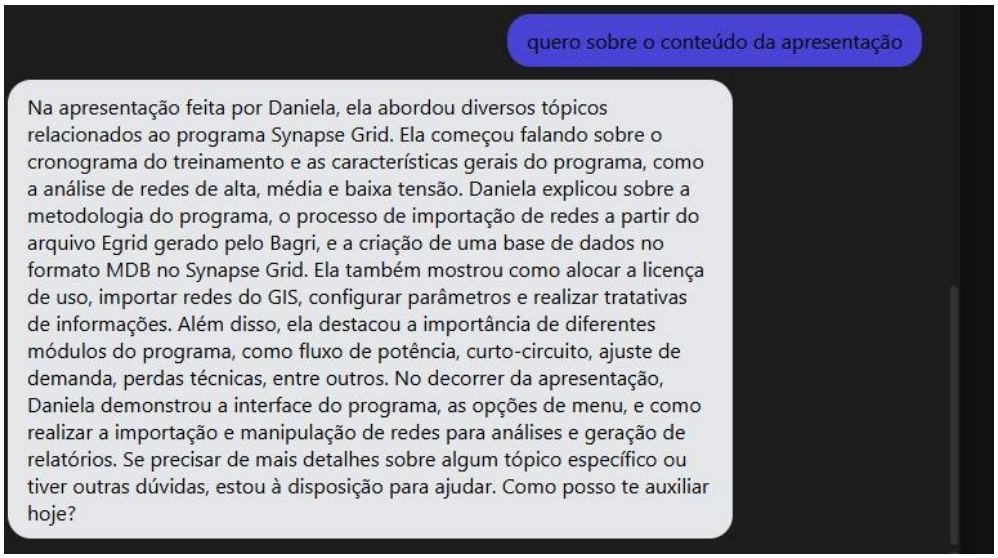
Os primeiros experimentos realizados com o modelo Whisper demonstraram resultados promissores em termos de acurácia e velocidade de transcrição. Em reuniões com áudio de boa qualidade, a taxa de erro manteve-se próxima a 10%, enquanto em ambientes com ruído moderado os valores se mantiveram aceitáveis, reforçando a robustez do modelo já apontada por Radford et al. (2022). Esse desempenho permitiu reduzir significativamente o tempo necessário para a documentação, já que uma reunião de uma hora passou a ser transcrita em poucos minutos, em contraste com as várias horas exigidas em processos manuais.

Na etapa de segmentação, a aplicação de embeddings semânticos possibilitou organizar automaticamente os conteúdos transcritos em blocos temáticos coerentes. Assim como apontado por Mikolov et al. (2013), os vetores distribuídos preservaram relações semânticas entre palavras e frases, o que foi fundamental para distinguir assuntos diferentes dentro de uma mesma reunião. Esse processo foi validado por testes práticos em que trechos de decisão, pendências e discussões foram corretamente agrupados, evidenciando ganhos de clareza e acessibilidade da informação.

Os métodos de segmentação baseados em embeddings mostraram resultados superiores às técnicas tradicionais de agrupamento apenas por palavras-chave, corroborando as observações de Alemi e Ginsparg (2015). Essa abordagem aumentou a precisão na separação dos tópicos, tornando os relatórios finais mais estruturados e de fácil consulta.

Os relatórios gerados têm se mostrado úteis tanto para pesquisadores quanto para profissionais corporativos, oferecendo resumos automáticos que destacam os principais pontos tratados nas reuniões. A comparação com transcrições manuais evidenciou que, embora ainda existam ajustes a serem feitos, a solução proposta já apresenta consistência e aplicabilidade prática. Como reforçam Jurafsky e Martin (2023), a integração entre reconhecimento de fala e processamento semântico constitui um avanço essencial para transformar dados não estruturados em conhecimento acessível e acionável.

Figura 3: resposta da IA respondendo o usuário sobre o treinamento gravado em video



Fonte: Próprio Autor

Referências

RADFORD, A. et al. **Robust speech recognition via large-scale weak supervision.** *arXiv preprint*, arXiv:2212.04356, 2022. Disponível em: https://cdn.openai.com/papers/whisper.pdf. Acesso em: ago. 2025.

ALEMI, A. A.; GINSPARG, P. **Text segmentation based on semantic word embeddings.** *arXiv preprint*, arXiv:1503.05543, 2015. Disponível em: https://arxiv.org/pdf/1503.05543. Acesso em: ago. 2025.

JURAFSKY, D.; MARTIN, J. H. ***Speech and language processing*.** 3. ed. Stanford: Stanford University, 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. Acesso em: ago. 2025.

MIKOLOV, T. et al. **Efficient estimation of word representations in vector space.** *arXiv preprint*, arXiv:1301.3781, 2013. Disponível em: https://arxiv.org/pdf/1301.3781. Acesso em: ago. 2025.