Proposta recebida em Outubro de 2019 e aceite para publicação em Novembro de 2019.

Identificação automática de unidades de informação em testes de reconto de narrativas usando métodos de similaridade semântica

Automatic identification of information units in tests based on narrative retelling using semantic similarity methods

Leandro Borges dos Santos
Universidade de São Paulo
leandrobs@usp.br

Sandra Maria Aluísio
Universidade de São Paulo
sandra@icmc.usp.br

Resumo

Os diagnósticos da Doença de Alzheimer (DA) e do Comprometimento Cognitivo Leve (CCL) baseiam-se na análise das funções cognitivas do paciente pela administração de baterias de avaliação cognitiva e neuropsicológica. O emprego do reconto de narrativas é comum para auxiliar a identificação e quantificação do grau de demência: é atribuído um ponto para cada unidade recordada, e o escore final representa a quantidade de unidades recordadas. Avaliamos duas tarefas da área clínica: a identificação automática de quais elementos de uma narrativa recontada foram recordados; e a classificação binária da narrativa produzida por um paciente, tendo as unidades identificadas como atributos, visando uma triagem automática dos pacientes com comprometimentos cognitivos. Utilizamos dois conjuntos de dados de reconto transcritos que possuem as sentenças divididas e anotadas manualmente com as unidades de informação e os disponibilizamos publicamente. São eles: a Bateria Arizona para Desordens de Comunicação e Demência (*ABCD*) com narrativas de pacientes com CCL e Controles Saudáveis e a Bateria de Avaliação da Linguagem no Envelhecimento (BALE), com narrativas de pacientes com DA e CCLs, e Controles Saudáveis. Avaliamos dois métodos baseados em similaridade semântica, chamados de *STS* e *Chunking*, e transformamos o problema multirrótulo de identificação de elementos de uma narrativa recontada em problemas de classificação binária, encontrando um ponto de corte para o valor de similaridade de cada unidade de informação. Dessa forma, conseguimos superar dois *baselines* para os dois conjuntos de dados na métrica *SubsetAccuracy*, que é a mais punitiva para o cenário multirrótulo. Na classificação binária nem todos os seis métodos de aprendizado de máquina avaliados tiveram melhor desempenho do que os baselines de identificação de unidades de informação. Para a *ABCD*, os melhores métodos foram Árvores de Decisão e *KNN*, e para a BALE, o *SVM* com *kernel RBF*.

Palavras chave

testes neuropsicológicos, reconto de narrativas, métodos de similaridade semântica

Abstract

Diagnoses of Alzheimer's Disease (AD) and Mild Cognitive Impairment (CCL) are based on the analysis of the patient's cognitive functions by administering cognitive and neuropsychological assessment batteries. The use of retelling narratives is common to help identify and quantify the degree of dementia. In general, one point is awarded for each unit recalled, and the final score represents the number of units recalled. In this paper, we evaluated two clinical tasks: the automatic identification of which elements of a retold narrative were recalled; and the binary classification of the narrative produced by a patient, having the units identified as attributes, aiming at an automatic screening of patients with cognitive impairment. We used two transcribed retelling data sets in which sentences were divided and manually annotated with the information units. These data sets were then made publicly available. They are: the Arizona Battery for Communication and Dementia Disorders (*ABCD*) that contains narratives of patients with CCL and Healthy Controls and the *Avaliação da Linguagem no Envelhecimento* (BALE), which includes narratives of patients with AD and CCLs as well as Healthy Controls. We evaluated two methods based on semantic similarity, referred to here as STS and Chunking, and transformed the multi-label problem of identifying elements of a retold narrative into binary classification problems, finding a cutoff point for the similarity value of each information unit. In this way, we were able to overcome two baselines for the two datasets in the SubsetAccuracy metric, which is the most punitive for the multi-label scenario. In binary classification, however, not all six machine learning methods evaluated performed better than the baselines methods. For ABCD, the best methods were Decision Trees and KNN, and for BALE, SVM with RBF kernel stood out.

Keywords

neuropsychological tests, narrative retellings, semantic similarity methods

DOI: 10.21814/lm.11.2.304

1 Introdução

O envelhecimento da população é uma tendência social conhecida em países desenvolvidos e que tem se tornado cada vez mais pronunciada também nos países em desenvolvimento (Fichman et al., 2011). O Brasil, por exemplo, está mudando sua pirâmide etária, segundo os censos do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2000 e 2010[1].

A maior expectativa de vida é um bem desejável, porém o envelhecimento pode ser acompanhado de doenças neurodegenerativas, como as demências, dentre as quais a Doença de Alzheimer (DA) é a mais proeminente, correspondendo a $50 - 80\%$ dos casos (Abbott, 2011). Assim, as demências são consideradas pela Organização Mundial de Saúde como um desafio para as próximas décadas, devido aos seus custos sociais e econômicos (Wortmann, 2012). Outra enfermidade que tem recebido atenção nos últimos anos é o Comprometimento Cognitivo Leve (CCL), que ocasiona declínio em funções cognitivas, podendo progredir para um quadro demencial. Assim, o CCL tem sido descrito como uma condição pré-clínica da DA (Clemente & Ribeiro-Filho, 2008; Frota et al., 2011). Mas, em alguns casos, o quadro pode se reverter para um estado normal e compatível com indivíduos da mesma faixa etária e nível de escolaridade, sendo melhor definido como uma condição heterogênea (Frota et al., 2011).

O diagnóstico das demências e síndromes relacionadas, comumente, baseia-se na análise das funções cognitivas do paciente, pela administração de baterias de avaliação cognitiva e neuropsicológica (McKhann et al., 2011; Frota et al., 2011; Mapstone et al., 2014; Hübner et al., 2019). As baterias avaliam as funções que são mais afetadas como diferentes tipos de memória, orientação, linguagem e resolução de problemas. Estas baterias são usadas antes, durante e depois de tratamentos, como diagnóstico, acompanhamento e direcionamento de tratamento (de Abreu et al., 2005). Como exemplos de baterias e testes temos: o teste Memória Lógica da Wechsler Memory Scale (Wechsler, 1997), a Bateria Montreal de Avaliação da Comunicação (Nasreddine et al., 2005), a Bateria Arizona para Desordens da Comunicação e Demência (ABCD) (Bayles & Tomoeda, 1993), o teste de Boston para o Diagnóstico da Afasia (Goodglass & Kaplan, 1983), dentre outros.

O emprego do reconto de narrativas é comum para auxiliar a identificar e quantificar o grau de demência. Em geral, as tarefas de reconto de narrativas utilizam uma história curta que é contada ao paciente, a quem se solicita que reconte a história imediatamente após ouvi-la com o máximo de detalhes. Em alguns casos, é solicitado ao paciente recontar novamente após 30 minutos. O reconto é gravado para posterior transcrição e análise.

Como pacientes com quadros demenciais tendem a possuir um vocabulário e usar estruturas sintáticas mais simples, são analisados os aspectos lexicais e sintáticos dos recontos. Também é possível mensurar a capacidade de memória de um paciente, por isso a narrativa é dividida em unidades de informação, podendo ser palavras ou orações. Em geral, é atribuído um ponto para cada unidade recordada, e o escore final representa a quantidade de unidades recordadas. As principais desvantagens dessa análise são: (i) a demanda de tempo, por ser uma tarefa de avaliação manual; (ii) a subjetividade do avaliador na checagem da presença das unidades de informação da narrativa no reconto. Assim, torna-se bem-vinda e importante a aplicação de métodos computacionais tanto para a automatização dessa tarefa, o que viabiliza sua aplicação em larga escala, como para a manutenção da uniformidade na correção.

Entretanto, há desafios computacionais também para a automatização do cálculo do escore por um sistema computacional. O sistema deverá resolver vários fenômenos que são comuns para essa tarefa, por exemplo: mudança da ordem de palavras da história original, uso de palavras similares às da história original, comentários que não estão relacionados com a história, e disfluências que tornam a história recontada bastante diferente da original.

Na Figura 1 (a), apresentamos a história do teste do reconto da bateria *ABCD*, traduzida para o português. Ela possui 5 sentenças e 61 palavras. Em (b), a mesma história é dividida em 17 unidades de informação, com possíveis alternativas entre parênteses, sendo 17 a sua pontuação máxima. Em (c), apresentamos um reconto imediato de um paciente com CCL, com pontuação 12, pois a avaliação manual de seu reconto contabilizou 12 unidades relembradas. Neste reconto, há trechos com disfluências ((1) e (3)), duplicação de unidades de informação recontadas ((1) e (3); (5) e (6)) e comentários não relacionados com a história ((2), (10) a (13)).

[1] https://censo2010.ibge.gov.br/sinopse/webservice/frm_piramide.php

(a) Enquanto uma senhora fazia compras, sua carteira caiu da bolsa, mas ela não viu. Quando ela foi ao caixa, não tinha como pagar as compras. Então, ela colocou as compras de lado e foi para casa. Assim que ela abriu a porta da casa, o telefone tocou e uma menininha disse-lhe que tinha achado a carteira. A senhora ficou muito aliviada.
(b) **Senhora (mulher)** // **estava fazendo compras (na loja, foi às compras, foi ao mercado)** // Sua carteira (seu porta-notas, sua moedeira) // **carteira caiu (derrubou a carteira, perdeu a carteira, perdeu a bolsa)** // da sua bolsa (da sua mochila, de sua pasta) // **Ela não viu a carteira cair (ela não notou)** // No caixa (quando ela foi pagar, guichê) // **não tem como pagar (ela não tinha dinheiro, não tinha sua carteira)** // Coloca as mercadorias de lado (coloca as mercadorias de volta) // **foi para sua casa (voltou para sua casa)** // Quando ela abriu a porta (quando ela chegou em casa, assim que ela entrou) // **telefone tocou (fone tocou, ela recebeu uma ligação)** // Pequena (jovem) // menina (garota) // lhe disse (falou, contou) // **ela achou a carteira (achou sua moedeira, achou o porta-notas)** // **Senhora aliviada (senhora estava feliz, senhora estava radiante, senhora estava agradecida)**
(c) (1) ahm uma senhora foi fazer compras no me foi no mercado. (2) não lembrava o local. (3) no me fazer compras. (4) e quando ela foi pagar a conta no caixa percebeu que estava sem a carteira. (5) aí ela foi deixou a mercadoria. (6) não levou a mercadoria. (7) voltou para casa. (8) chegando em casa toca o telefone. (9) era uma garotinha avisando ela que que tinha achado a carteira. (10) é isso. (11) tem mais coisa. (12) não cortei. (13) eu resumi o que eu ouvi.

Figura 1: (a) Narrativa original da bateria ABCD; (b) Narrativa original separada em unidades de informação; as nove unidades marcadas em negrito são as principais, o resto são detalhes; (c) Transcrição do reconto imediato de um paciente com CCL, segmentada manualmente em sentenças.

Existem poucos trabalhos na literatura que tratam da automatização da identificação de unidades de informação em recontos de narrativas. Podemos dividi-los em: métodos de busca de palavras (Pakhomov et al., 2010; Fraser et al., 2016), métodos de alinhamento (Prud'hommeaux & Roark, 2015), e métodos de *clustering* (Yancheva & Rudzicz, 2016; Fraser et al., 2019).

Neste artigo, avaliamos automaticamente quais elementos de uma narrativa recontada foram recuperados, utilizando dois métodos baseados em similaridade semântica. Esses elementos são usados como atributos para métodos de classificação binária da narrativa produzida por um paciente realizando um teste neuropsicológico baseado em reconto. No melhor do nosso conhecimento, não há trabalhos na literatura sobre a identificação das unidades de informação em recontos modelada com métodos de similaridade semântica.

O restante deste artigo é organizado do seguinte modo: na Seção 2 são apresentadas as principais características do diagnóstico da Doença de Alzheimer e do Comprometimento Cognitivo Leve (Seção 2.1) e uma descrição dos trabalhos sobre identificação automática de unidades de informação em recontos (Seção 2.2). Na Seção 3, são descritos os córpus utilizados neste estudo, os métodos de similaridade semântica propostos e as baselines; já na Seção 4, mostramos os resultados dos experimentos para a classificação das unidades de informações nos dois datasets avaliados neste artigo. Na Seção 5, são apresentados os resultados dos métodos de classificação automática de narrativas que se basearam nos atributos recuperados automaticamente pelos métodos descritos na Seção 3. Por fim, na Seção 6, trazemos as conclusões e apresentamos sugestões de trabalhos futuros.

2 Trabalhos relacionados

2.1 Diagnóstico de Demências e Síndromes Relacionadas

O envelhecimento acarreta algumas perdas de funcionalidades, como a capacidade motora, a diminuição dos mecanismos de defesa natural do organismo e da adaptação ao ambiente. Acarreta, também, a diminuição de funcionalidades cognitivas, como a linguagem, que tem um papel fundamental na vida das pessoas, possibilitando a comunicação e as demais atividades sociais. Essas modificações não ocorrem de forma isolada e sim estão relacionadas com as alterações na memória operacional, na atenção e nas habilidades visuoespacial (Freitas, 2010). Nos idosos, são notadas alterações dos padrões discursivos conforme o estímulo. Para tarefas nas quais é exigido o reconto de narrativas ouvidas recentemente, são obtidos textos curtos e simples. Ao contrário das tarefas em que é necessária a produção de narrativas livres, elicitadas por meio de estímulo visual de um livro de figuras, nas quais os idosos tendem a elaborar textos mais longos, mas contendo um número maior de informações irrelevantes e com baixa coesão (Garcia & Mansur, 2006). Nesta seção, apresentamos as principais características do diagnóstico da Doença de Alzheimer e do Comprometimento Cognitivo Leve.

2.1.1 Doença de Alzheimer

No Brasil, as recomendações para o diagnóstico da DA foram elaboradas em 2011, pelos membros do Departamento de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia (Frota et al., 2011). Seguem abaixo os critérios clínicos principais para o diagnóstico de demência de qualquer tipo:

1. Demência é diagnosticada quando há sintomas cognitivos ou comportamentais (neuropsiquiátricos) que: (i) Interferem com a habilidade no trabalho ou em atividades usuais; (ii) Representam declínio em relação a níveis prévios de funcionamento e desempenho; (iii) Não são explicáveis por delirium (estado confusional agudo) ou doença psiquiátrica maior;
2. O comprometimento cognitivo é detectado e diagnosticado mediante combinação de (i) Anamnese com paciente e informante e (ii) Avaliação cognitiva objetiva, mediante exame breve do estado mental ou avaliação neuropsicológica;
3. Os comprometimentos cognitivos ou comportamentais afetam no mínimo dois dos seguintes domínios: memória, funções executivas, habilidades visuoespaciais, linguagem e personalidade/comportamento.

O declínio cognitivo progressivo é confirmado com exames sucessivos e a positividade de biomarcadores. Também são utilizados exames de imagens para exclusão de outros diagnósticos.

Mesmo que a perda de memória seja a característica mais frequente, alterações na linguagem também podem aparecer nos estágios iniciais da DA. Uma das formas de se avaliar a linguagem é a produção de narrativas, sendo observado que estas narrativas apresentam sentenças simples e curtas, maior número de proposições irrelevantes, vocabulário pobre, ruptura no desenvolvimento do tema, maior número de erros ortográficos e menor nível de complexidade sintática (Mansur et al., 2005).

2.1.2 Comprometimento Cognitivo Leve

Para a identificação do comprometimento cognitivo leve são utilizados testes neuropsicológicos, por serem mais sensíveis, embora não exista uma norma para o valor do ponto de corte. Essa dificuldade decorre pelo fato de ser uma situação entre o envelhecimento normal e a demência. Frota et al. (2011) sugerem as principais características utilizadas para identificação do CCL:

- Queixa de alteração cognitiva relatada pelo paciente ou informante próximo;
- Evidência de comprometimento cognitivo em um ou mais dos seguintes domínios: memória, função executiva, linguagem e habilidades visuoespaciais;
- Preservação da independência funcional;
- Não preenche critérios de demência.

Recentemente, há evidências de que indivíduos com CCL têm mais risco para desenvolver DA, devido a comprometimentos em múltiplos domínios, incluindo a linguagem. Por essa razão, é importante compreender a natureza do comprometimento de linguagem. Em Fleming & Harris (2008) foi realizado um estudo comparativo do discurso produzido por pacientes com CCL e idosos normais, observando que o discurso produzido por pacientes com CCL contém um número menor de palavras, e as suas características se comparam com os estágios iniciais de DA. Enquanto que em Chapman et al. (2002) foram comparadas as habilidades de compreensão, memória e expressão de texto discursivo extenso, identificando que a capacidade de fornecer informações detalhadas e realizar síntese de ideais a partir das narrativas estava comprometida quando comparadas com as habilidade dos pacientes normais, sendo muito similar à de pacientes acometidos por DA. Hodges et al. (1996) examinaram o desempenho de controles saudáveis e indivíduos com diversos graus de comprometimento de DA em tarefas de nomeação e geração de definições e reconheceram que a qualidade da definição produzia diferenças entre os grupos. Os estudos sobre descrição (oral e escrita) de figuras simples e complexas realizados por Forbes-McKay & Venneri (2005) também distinguem indivíduos com DA em grau leve e indivíduos saudáveis.

Em resumo, para avaliar a linguagem de indivíduos com CCL é importante dispor de instrumentos/testes sensíveis para detectar déficits sutis. Além disso, o monitoramento dessas dificuldades também carece de instrumentos acurados. A análise do discurso mostra-se interessante, pois abrange os diferentes componentes da linguagem, em uma perspectiva linguístico-cognitiva.

2.2 Métodos de Identificação Automática das Unidades de Informação em Recontos

Nesta seção, organizamos a descrição dos métodos de identificação automática das unidades de informação em recontos da literatura em três abordagens: métodos de busca de palavras (Pakhomov et al., 2010; Fraser et al., 2016), métodos de alinhamento (Prud'hommeaux & Roark, 2015), e métodos de *clustering* (Yancheva & Rudzicz, 2016; Fraser et al., 2019).

2.2.1 Métodos de Busca de Palavras

Pakhomov et al. (2010) compilaram uma lista com palavras e frases que representavam algum

conceito da cena do Roubo do Biscoito, que é uma subtarefa da Bateria de Boston (*Boston Diagnostic Aphasia Examination —BDAE*) (Goodglass & Kaplan, 1983). As narrativas foram divididas em *n-grams*, de 1 a 4, e para cada *n-gram* os autores realizaram uma busca na lista de palavras. Se o *n-gram* era encontrado, considerou-se que o paciente se relembrou dessa unidade de informação.

Os autores utilizaram 38 narrativas de idosos com Degeneração Lobar Frontotemporal, com o seguintes subtipos: Afasia Progressiva Primária, Demência Semântica, variante comportamental da Demência Frontotemporal, e Afasia Logopênica. Entretanto, não encontraram diferença estatisticamente significativa entre os grupos na contagem de unidades de informação recordadas.

Fraser et al. (2016) utilizaram uma lista de palavras para cada possível unidade de informação. Para as unidades de informação que representam uma ação, os autores utilizaram o *parser* de *Stanford* para identificar o verbo e o sujeito, e analisaram se essa combinação estava na lista de palavras. As unidades de informação foram utilizadas como atributos binários em conjunto com métricas de PoS, de complexidade sintática, psicolinguísticas, de diversidade lexical, de constituintes gramaticais, de repetividade de informações, e acústicas, totalizando 370 atributos. O objetivo dos autores foi distinguir narrativas de pacientes com Doença de Alzheimer e envelhecimento saudável no conjunto de dados *DementiaBank* (Becker et al., 1994), neste conjunto os pacientes são solicitados a descrever a cena do Roubo do Biscoito (Goodglass & Kaplan, 1983). Os autores utilizaram 233 narrativas de 97 participantes com envelhecimento saudável e 240 narrativas de 168 participantes com possível ou provável DA. Para a classificação final, usaram o algoritmo de Regressão Logística, *10-fold-cross-validation*, e a métrica acurácia para avaliação, dado que a classificação era binária. O melhor resultado foi 0,819 de acurácia, utilizando 35 atributos selecionados com o método de Correlação de Pearson.

2.2.2 Métodos de Alinhamento

Prud'hommeaux & Roark (2015) propuseram um método de alinhamento baseado em grafos, utilizando a técnica de passeios aleatórios (*Random Walks*) para automatizar o teste de reconto de narrativas do teste de Memória Lógica de Wechsler. Na abordagem proposta, cada palavra do reconto ou da narrativa original representa um nó do grafo e o alinhamento entre as palavras representa as arestas. São utilizadas narrativas de 235 pacientes, sendo 72 pacientes com CCL, 163 com envelhecimento saudável, e 48 narrativas de pacientes inelegíveis, i.e., que não se enquadraram em algum critério e não podem fazer parte dos grupos CCL ou Envelhecimento Saudável.

No método proposto, primeiramente, cada narrativa de reconto é alinhada com a narrativa original e as demais narrativas de reconto. Para obter os alinhamentos é utilizado o alinhador *Berkeley Aligner* (Liang et al., 2006). A partir desses alinhamentos é construído um grafo, em que é verificado se o alinhamento possui uma probabilidade maior que 0.5. Neste caso, é adicionado um vértice entre essas palavras. Desse modo, podem existir dois tipos de alinhamentos: o alinhamento com uma palavra da narrativa fonte, e o alinhamento com uma palavra da narrativa de reconto. Dada uma palavra da narrativa de reconto, esta é definida como o vértice inicial da caminhada aleatória. A cada passo da caminhada é gerado um valor aleatório, e caso este seja maior que um λ, é realizada uma transição para uma palavra da narrativa original; caso contrário, a transição é realizada para uma palavra da narrativa de reconto. Quando a caminhada aleatória atingir uma palavra da narrativa fonte, é proposto um novo alinhamento entre a palavra inicial e a palavra fonte, e a caminhada é encerrada.

Para cada palavra presente nas narrativas de reconto são realizados mil passeios aleatórios. O novo alinhamento, entre a palavra de reconto e a palavra fonte, é definido pelo alinhamento mais frequente dos passeios aleatórios. Após a obtenção dos alinhamentos, estes são utilizados como atributos para um classificador final; se alguma palavra da narrativa de reconto estiver alinhada com a narrativa original é considerado que o paciente se recordou desse trecho.

Na tarefa de classificação final (Envelhecimento Saudável *versus* CCL), Prud'hommeaux & Roark (2015) exploram duas representações para cada paciente: (i) o *Summary score* que é a quantidade de unidades de informações recordadas no reconto imediato e tardio; (ii) *Element scores* em que cada unidade de informação representa um atributo. É marcado se o paciente se recordou ou não dessa unidade de informação no reconto imediato e no tardio. O melhor resultado na classificação final, utilizando o método automático com o *Element scores* foi de $0,792$ de AUC, enquanto que utilizando os escores obtidos de forma manual o resultado foi de $0,813$.

2.2.3 Métodos de Clustering

Yancheva & Rudzicz (2016) e Fraser et al. (2019) automatizaram a análise de unidades de informação aplicando algoritmos de agrupamento, em que os *clusters* são considerados como um indicador (*proxy*) para as unidades de informação e são utilizados para extrair atributos. Os detalhes de cada método são explicados a seguir.

Yancheva & Rudzicz (2016) avaliaram o método no *DementiaBank*, utilizaram 241 narrativas de 98 participantes com envelhecimento saudável e 255 narrativas de 168 participantes com possível ou provável DA. Os verbos e os substantivos das transcrições são convertidos em uma representação densa com o método *GloVe* (Pennington et al., 2014); para cada grupo é aplicado o algoritmo *K-means* com a distância euclidiana e k igual a 10. A partir dos *clusters* são criados atributos baseados nas distâncias.

Na tarefa de classificação final, os autores optaram pelo classificador *Random Forest* via *10-fold-cross-validation*. A abordagem proposta foi comparada com o resultado da classificação utilizando uma lista de palavras para recuperar as unidades de informação. Os autores também adicionaram atributos de métricas linguísticas e acústicas. O classificador que utiliza os atributos do modelo de *cluster* do grupo de Envelhecimento Saudável e do grupo de DA obteve $0,74$ de F1, e quando combinado com as métricas linguísticas e acústicas obteve $0,80$. Observa-se que o *baseline* com lista de palavras de cada unidade de informação obteve $0,73$ de F1.

Fraser et al. (2019) substituíram o modelo *GloVe* pelo *FastText* (Bojanowski et al., 2017), possibilitando inferir palavras que não estão presentes no vocabulário do modelo de *embeddings*, e optarem pela distância do cosseno em vez da euclidiana. Foram utilizados três conjuntos de dados: o *DementiaBank*, com 97 participantes saudáveis e 19 participantes com CCL; o *Gothenburg* (Wallin et al., 2016), que é composto por transcrições da descrição da cena do Roubo do Biscoito de pacientes suecos, com 36 participantes saudáveis e 31 com CCL; e o *Karolinska*, que foi coletado por Cromnow & Landberg (2009), tendo sido solicitado a 96 indivíduos com envelhecimento saudável que produzissem uma descrição escrita da cena do Roubo do Biscoito em 5 minutos.

Os autores extraíram todos os verbos e os substantivos das transcrições. Em seguida, as palavras foram transformadas em vetores a partir do alinhamento das matrizes de *word embeddings* em Inglês e Sueco do *FastText*, em seguida aplicaram o algoritmo *k-means* com três variações do parâmetro k, sendo: 10, 23, e k_{sil}, onde $k_{sil} \in \{2, 3, \ldots, 30\}$. Para k_{sil} o valor é selecionado de forma automática pelo método da silhueta (Kaufman & Rousseeuw, 2009). Para cada configuração de k foram construídos modelos de agrupamento para o Inglês, Sueco, e uma versão multilíngue (Inglês e Sueco). Após a obtenção dos agrupamentos, foram extraídos atributos baseados nas distâncias em relação aos centroides.

Para a etapa de classificação, os autores utilizaram o *SVM* linear e o *leave-one-out*, e avaliaram acurácia no conjunto de dados do *DementiaBank* balanceado, e *Gothenburg*. O conjunto de dados *Karolinska* e 78 participantes saudáveis restantes do *DementiaBank* foram utilizados no treinamento dos modelos de agrupamentos. Para cada iteração do *leave-one-out*, os autores executaram um *inner-cross-validation* para selecionar os parâmetros do *SVM*, e selecionaram o modelo de agrupamento a partir de 10 execuções.

No *DementiaBank*, o melhor resultado foi o modelo multilíngue com k igual a 10, que obteve uma acurácia de $0,63$; para o modelo de agrupamento monolíngue a melhor acurácia foi de $0,47$ com k igual a 10 e k_{sil}. No *Gothenburg*, o melhor resultado foi de $0,72$ com o modelo de multilíngue, e k igual a 23, enquanto que o melhor resultado do modelo monolíngue foi de $0,55$ com k igual a 10. Além disso, os autores avaliaram no mesmo cenário de Yancheva & Rudzicz (2016), ou seja, identificação de pacientes com DA *versus* idosos saudáveis, para k igual a 10, e obtiveram F1 score de $0,83$, enquanto que Yancheva & Rudzicz (2016) obtiveram $0,74$.

3 Desenvolvimento

Na Seção 3.1, são apresentados os dois córpus utilizados neste trabalho, bem como a metodologia proposta para compilá-los. Na Seção 3.2, são descritos dois métodos de identificação automática de unidades de informação, baseados em métodos de similaridade semântica. Na Seção 3.3, são descritas as *baselines* para comparação de desempenho dos métodos da Seção 3.2.

3.1 Conjuntos de dados utilizados

Utilizamos dois conjuntos de dados de reconto que possuem as sentenças anotadas manualmente com as unidades de informação, descritos em Santos et al. (2019). Os conjuntos estão disponíveis para download no GitHub[2].

[2] https://github.com/lbsantos/ANAA-Dementia/tree/master/conjutos_de_dados

A Tabela 1 apresenta as estatísticas dos dois conjuntos, que trazem uma média do tamanho de sentenças bem próxima entre CCLs e Controles na Bateria Arizona para Desordens de Comunicação e Demência (*ABCD*), mas uma diferença maior dos grupos DA e CCL com o grupo de Controle da Bateria de Avaliação da Linguagem no Envelhecimento (BALE) (Hübner et al., 2019).

O primeiro conjunto de dados é formado por transcrições da *ABCD* que é composta de 17 subtestes. Nos interessa neste trabalho o subteste de reconto no qual é contada uma história ao paciente, e este tem quer recontar a história imediatamente e depois de 30 minutos. O teste do reconto foi aplicado em 23 idosos com CCL e 12 adultos com envelhecimento saudável (Controles), na Faculdade de Medicina da USP. Este teste possui 17 unidades de informação, apresentadas na Figura 1 (b), com possíveis alternativas entre parênteses; a sua pontuação máxima é 17.

O segundo conjunto de dados é formado por transcrições da BALE que possui diversas tarefas, sendo uma delas o reconto de uma história apresentada oralmente (História da Lúcia). A História da Lúcia possui originalmente 24 unidades de informação que foram reagrupadas neste trabalho, resultando em 21 unidades (Figura 2). O teste do reconto foi aplicado em 11 idosos com DA, 5 idosos com CCL e 53 adultos com envelhecimento saudável (Hübner et al., 2019).

Lúcia // **mora** // **interior** // do Paraná // Numa manhã de 2a feira // **ela saiu de casa** // **para buscar emprego (foi para uma uma entrevista, foi buscar trabalho)** // na capital do estado (em Curitiba) // **Foi para rodoviária** // **foi de carona (pegou carona)** // **com amigo Pedro (com Pedro)** // Estava chovendo // naquela manhã // O carro // passou (caiu) // por um buraco // **o pneu furou** // Pensou que ia perder (achou que ia perder) // o ônibus // **Pegou um táxi** // **conseguiu chegar chegou a tempo (chegou a tempo)**

Figura 2: Narrativa utilizada na BALE, separada em unidades de informação; as onze unidades principais são marcadas em negrito.

Nas Figuras 1 e 2, anotamos as unidades da macroestrutura em negrito, seguindo o modelo de análise de Kintsch & van Dijk (1978) em que as unidades de informação do texto são organizadas de forma hierárquica, sendo a macroestrutura correspondente às ideias principais e a microestrututura às ideias acessórias e detalhes.

Para cada conjunto de dados, o áudio do participante foi transcrito manualmente, seguindo os princípios do NURC / SP No 338 EF e 331 D (Preti, 2005) e segmentado manualmente em sentenças por um anotador experiente, usando conhecimento prosódico (pausas), sintático e semântico. Optamos por utilizar uma segmentação manual para isolarmos os efeitos de erros de um sistema de segmentação automática. Embora Treviso & Aluísio (2018) tenham desenvolvido um sistema de segmentação sentencial para narrativas elicitadas com estímulo visual (livros de figuras), este ainda não consegue generalizar para narrativas elicitadas com estímulos orais (recontos).

Para criarmos os conjuntos de dados anotados com as unidades de informação sobre as unidades de interesse (sentenças anotadas no pré-processamento), utilizamos o sistema de anotação *brat* (*brat rapid annotation tool*) (Stenetorp et al., 2012), realizando a anotação em duas fases. Na primeira fase, cada sentença da transcrição foi classificada de acordo com a lista de unidades de informação de cada bateria por um único anotador; na segunda fase, outro anotador revisou a anotação e os casos discordantes foram discutidos, visando obter uma anotação concordante.

A narrativa da *ABCD* foi mantida com as 17 unidades de informação originais, mas para a narrativa da BALE, realizamos algumas modificações nas unidades de informação (ora separando, ora juntando) para termos uma anotação manual uniforme, sem discrepâncias e possibilitar a aplicação de métodos automáticos. A partir dessas modificações, finalizamos com 21 unidades de informação (Figura 2) em vez das 24 unidades originais, com 11 delas sendo unidades macroestruturais.

3.2 Modelando a identificação de unidades de informação via similaridade semântica

Métodos de similaridade semântica textual (*STS*, *Semantic Textual Similarity*) e inferência textual (*RTE*, *Recognizing Textual Entailment*) têm aplicações em diversas tarefas de Processamento de Línguas Naturais como: recuperação de informação, sistemas de perguntas-repostas, avaliação de sistemas de tradução, dentre outras (Agirre et al., 2012, 2015).

Na tarefa de *STS* o objetivo é indicar o grau de similaridade entre dois textos, ou seja, dado um par de textos $(\mathbf{S}_i^1, \mathbf{S}_i^2)$, estamos interessados em atribuir um valor y_i em alguma escala, geralmente de 0 a 5 ou 1 a 5 (Agirre et al., 2012, 2015; Marelli et al., 2014; Fonseca et al., 2016). Essa gradação naturalmente captura as diferenças sutis de similaridade, como sentenças que possuem o mesmo significado (pontuação 5), possuem pequenas diferenças semânticas (pontuação 4), compartilham apenas alguns detalhes

Bateria	Grupo	Sujeitos	Média Sentenças (Desvio Padrão)	Média de palavras por sentença (Desvio Padrão)
ABCD	CCL	23	8,17 (1,92)	60,76 (17,39)
	Controle	12	7,67 (2,06)	58,96 (14,73)
BALE	DA	11	6,09 (2,63)	36,18 (17,10)
	CCL	5	6,00 (1,00)	36,40 (5,68)
	Controle	53	7,68 (2,67)	52,06 (19,18)

Tabela 1: Estatísticas dos Conjuntos de Dados.

(pontuação 3), sentenças não relacionadas, mas versam sobre o mesmo assunto (pontuação 2), ou mesmo que não possuem nada em comum (pontuação 1).

Já o *RTE* pode ser definido como uma relação direcional entre dois textos, em que dado um texto **T** permite-se que se conclua que uma hipótese **H** é verdadeira (Dagan et al., 2006; Marelli et al., 2014; Fonseca et al., 2016). Com essa definição é assumido que pessoas lendo o par $(\mathbf{T}, \mathbf{H})$ compartilham: (i) o conhecimento da língua em que os textos são formulados, e (ii) o mesmo conhecimento prévio sobre o tema (Dagan et al., 2006).

Uma das questões investigadas nesse artigo foi a possibilidade de utilizar métodos de *STS* para identificar as unidades de informação recordadas; abordagem inédita, até onde sabemos.

Para obter a similaridade semântica de duas sentenças neste artigo, utilizamos dois métodos:

1. O método de Hartmann (2016)[3], o qual obteve o melhor resultado de similaridade semântica na Avaliação de Similaridade Semântica e INferência textual (ASSIN) (Fonseca et al., 2016) — chamamos esse método de *STS*;
2. O método chamado de *Chunking*, proposto neste trabalho, explora a similaridade de representações vetoriais obtidas por *embeddings*.

Hartmann (2016) utilizou uma abordagem baseada no valor da similaridade do cosseno de duas representações vetoriais de cada sentença.

Na primeira representação, o autor soma os vetores de cada palavra obtidos pelo word2vec (Mikolov et al., 2013). Na segunda representação, é realizada uma expansão do vocabulário: para cada palavra de conteúdo são buscados os sinônimos no TEP (Thesaurus para o português do Brasil) (Maziero et al., 2008). Essa expansão é restrita apenas a palavras que possuam até um sinônimo, o que corresponde a 28% das entradas do TEP. Em seguida, os pares são transformados em uma representação esparsa, utilizando o *TF-IDF* (*frequency–inverse document frequency*), e é obtida a similaridade do cosseno.

Por fim, os valores dos cossenos entre as duas representações (*TF-IDF* e *word2vec*) de cada par são dados como entrada para um regressor linear que determina a similaridade do par.

Na Tabela 2, são apresentados alguns exemplos de sentenças das narrativas de reconto e as sentenças da narrativa original com os valores de similaridade. Com esses exemplos, é possível perceber a viabilidade da exploração de *STS* para a tarefa avaliada neste artigo. As duas primeiras linhas da Tabela 2 apresentam valores altos de similaridade semântica, sendo que, pela definição da anotação do ASSIN, os valores indicam que as sentenças são muito semelhantes, mas apresentam algumas informações exclusivas. Enquanto que a terceira e quarta sentenças apresentam valores de similaridade próximos de 3, sendo que esse valor indica que as sentenças possuem similaridade e podem se referir ao mesmo fato.

Dado que o sistema de *STS* recebe como entrada dois textos curtos, para possibilitar a aplicação desse sistema as narrativas originais de cada bateria foram sentenciadas e para cada sentença foram atribuídos seus respectivos rótulos. Nas Tabelas 3 e 4 são apresentados os resultados dessa etapa. Assim como nas sentenças dos pacientes, algumas sentenças possuem mais que um rótulo.

Como temos um problema multirrótulo, adotamos a abordagem de transformação de problema *Binary Relevance* (Tsoumakas et al., 2009), para reduzirmos o problema multirrótulo para vários problemas binários. Em cada problema queremos identificar se a sentença do reconto é uma respectiva unidade de informação ou não. Dessa forma, para cada sentença é criado um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, sendo que UI_j é a unidade de informação j, e para cada par é obtido o valor de similaridade. Na Tabela 5, são apresentados

[3] O autor gentilmente nos forneceu o código fonte e os modelos utilizados no sistema.

Similaridade	Sentença do reconto	Sentença da narrativa
4,17	e ela ficou aliviada.	a senhora ficou muito aliviada.
4,55	uma senhora fazia as compras no mercado.	uma senhora fazia compras.
3,09	e ai foi pegou um táxi pra chegar com tempo.	então ela pegou um táxi até a rodoviária.
2,93	ela pegou carona.	ela foi para a rodoviária de carona com seu amigo.

Tabela 2: Exemplos para os valores de similaridade semântica.

Sentenças	Rótulos
Lúcia mora no interior do Paraná	LUCIA MORA INTERIOR PARANA
numa manhã de segunda-feira	NUMA_MANHA_SEGUNDA
ela saiu de casa para mais uma entrevista de trabalho na capital do estado	SAIU_DE_CASA BUSCAR_EMPREGO NA_CAPITAL
ela foi para a rodoviária de carona com seu amigo Pedro	FOI_RODOVIARIA
estava chovendo naquela manhã	ESTAVA_CHOVENDO NAQUELA_MANHA
de repente o carro passou por um buraco	CARRO PASSOU_CAIU BURACO
e o pneu furou	PNEU_FUROU
Lúcia pensou que iria perder o ônibus	PENSOU_ACHOU_PERDER ONIBUS
então ela pegou um táxi até a rodoviária	PEGOU_TAXI
e conseguiu chegar a tempo	CONSEGUIU_CHEGAR_TEMPO

Tabela 3: Sentenças da narrativa original da BALE rotuladas com as unidades de informação.

os valores de similaridade para a sentença "*uma senhora fazia as compras no mercado*" contrastada com cada rótulo das sentenças da narrativa original, apresentada na Tabela 4.

Na Figura 3, são dispostos os histogramas e a estimação de densidade por *kernel* para cada unidade de informação da *ABCD*. É possível perceber a separação para algumas unidades de informação, como: SENHORA, ESTAVA_FAZENDO_COMPRAS, QUANDO_ELA_ABRIU_A_PORTA. Entretanto, outras não apresentam uma separação clara, como ELA_NAO_VIU_A_CARTEIRA_CAIR.

Dado o valor de similaridade semântica do par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, é necessário definir um ponto de corte para transformar esse valor em uma resposta binária. Para encontrar o ponto de corte que maximizasse a medida $F1$ para classe UI_j, aplicamos um otimizador Bayesiano com a técnica *TPE* (*Tree-structured Parzen Estimator*) (Bergstra et al., 2013).

Os passos do segundo método avaliado e chamado de *Chunking* são elencados abaixo:

1. Um *tagger* probabilístico (López & Pardo, 2015) que atribui a classe gramatical mais frequente do conjunto de dados foi utilizado para filtrar as palavras de conteúdo das sentenças dos recontos, de forma semelhante aos trabalhos de Yancheva & Rudzicz (2016) e Fraser et al. (2019). Escolhemos esse *tagger*, pois as narrativas de reconto possuem ruídos que podem afetar o desempenho de *PoS taggers* treinados em córpus.
2. Em seguida, as palavras são convertidas para uma representação densa com o *FastText*, e calculamos a média dos vetores em $\mathbf{S}_{UI_j}^2$.
3. Lembrando que queremos identificar se a sentença é uma respectiva unidade de informação ou não, então para cada sentença é criado um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, sendo que UI_j é a unidade de informação j, e para cada par é obtido o valor de similaridade.
4. Dada uma sentença da narrativa de reconto, $\mathbf{S}_i^1$, esta é dividida em *n-grams*, variando de 1 a 3.
5. Para cada n-gram, é calculada a média dos vetores que compõem esse *n-gram*.
6. Por fim, calculamos a similaridade do cosseno desses vetores e retornamos o valor mais próximo de $\mathbf{S}_{UI_j}^2$.
7. Utilizamos um otimizador Bayesiano com a técnica *Tree-structured Parzen Estimator — TPE* para encontrar o ponto de corte.

Na Figura 4, apresentamos um exemplo da aplicação do método de *Chunking*. Dada uma sentença $\mathbf{S}_i^1$ contendo 4 palavras de conteúdo

Sentenças	**Rótulos**
uma senhora fazia compras	SENHORA ESTAVA_FAZENDO_COMPRAS
sua carteira caiu da bolsa	SUA_CARTEIRA CARTEIRA_CAIU DA_SUA_BOLSA
mas ela não viu	ELA_NAO_VIU_A_CARTEIRA_CAIR
quando ela foi ao caixa	NO_CAIXA
não tinha como pagar as compras	NAO_TEM_COMO_PAGAR
então ela colocou as compras de lado	COLOCA_AS_MERCADORIAS_DE_LADO
foi para casa	FOI_PARA_SUA_CASA
assim que ela abriu a porta da casa	QUANDO_ELA_ABRIU_A_PORTA
o telefone tocou	TELEFONE_TOCOU
uma menininha disse-lhe que tinha achado a carteira	PEQUENA MENINA LHE_DISSE ELA_ACHOU_A_CARTEIRA
a senhora ficou muito aliviada	SENHORA_ALIVIADA

Tabela 4: Sentenças da narrativa original da *ABCD* rotuladas com as unidades de informação.

Rótulos	**Similaridade**
SENHORA	4,55397
ESTAVA_FAZENDO_COMPRAS	4,55397
SUA_CARTEIRA	1,2814
CARTEIRA_CAIU	1,2814
DA_SUA_BOLSA	1,2814
ELA_NAO_VIU_A_CARTEIRA_CAIR	1,24164
NO_CAIXA	1,20222
NAO_TEM_COMO_PAGAR	2,71347
COLOCA_AS_MERCADORIAS_DE_LADO	2,48818
FOI_PARA_SUA_CASA	1,31341
QUANDO_ELA_ABRIU_A_PORTA	1,46262
TELEFONE_TOCOU	1,08743
PEQUENA	1,08353
MENINA	1,08353
LHE_DISSE	1,08353
ELA_ACHOU_A_CARTEIRA	1,08353
SENHORA_ALIVIADA	2,52263

Tabela 5: Valores de similaridade da sentença "uma senhora fazia as compras no mercado".

(P_1, P_2, P_3, P_4), esta é dividida em 9 *n-grams* (linhas da tabela à esquerda), e para cada *n-gram* calculamos a distância do cosseno para cada sentença que representa uma unidade de informação. Em seguida, utilizamos o valor máximo para cada classe (C_1, C_2, C_3) e aplicamos os pontos de corte para definir quais unidades de informação a sentença contém.

3.3 Baselines

Para comparar os métodos apresentados na Seção 3.2 na tarefa de identificação de unidades de informação, foram usadas duas estratégias.

A primeira, chamada de Casamento Exato, utiliza uma lista de palavras para identificar as unidades de informação, via casamento exato. Essa abordagem também foi utilizada em outros trabalhos (Prud'hommeaux & Roark, 2015; Pakhomov et al., 2010; Fraser et al., 2016). A segunda utiliza a saída do sistema *baseline* de inferência textual do ASSIN. Nessa abordagem, chamada aqui de *Inferência*, consideramos que a sentença contém a unidade de informação se o sistema de inferência retornar os rótulos *"Inferência"* e *"Paráfrase"*.

4 Avaliação de Métodos de Identificação de Unidades de Informação

Nesta seção, são apresentados os experimentos para identificação de unidades de informação, para os métodos descritos na Seção 3.2 e as *baselines* da Seção 3.3.

Os conjuntos foram separados em treino e teste, utilizando 70% para treino e 30% para teste de forma estratificada para cada grupo.

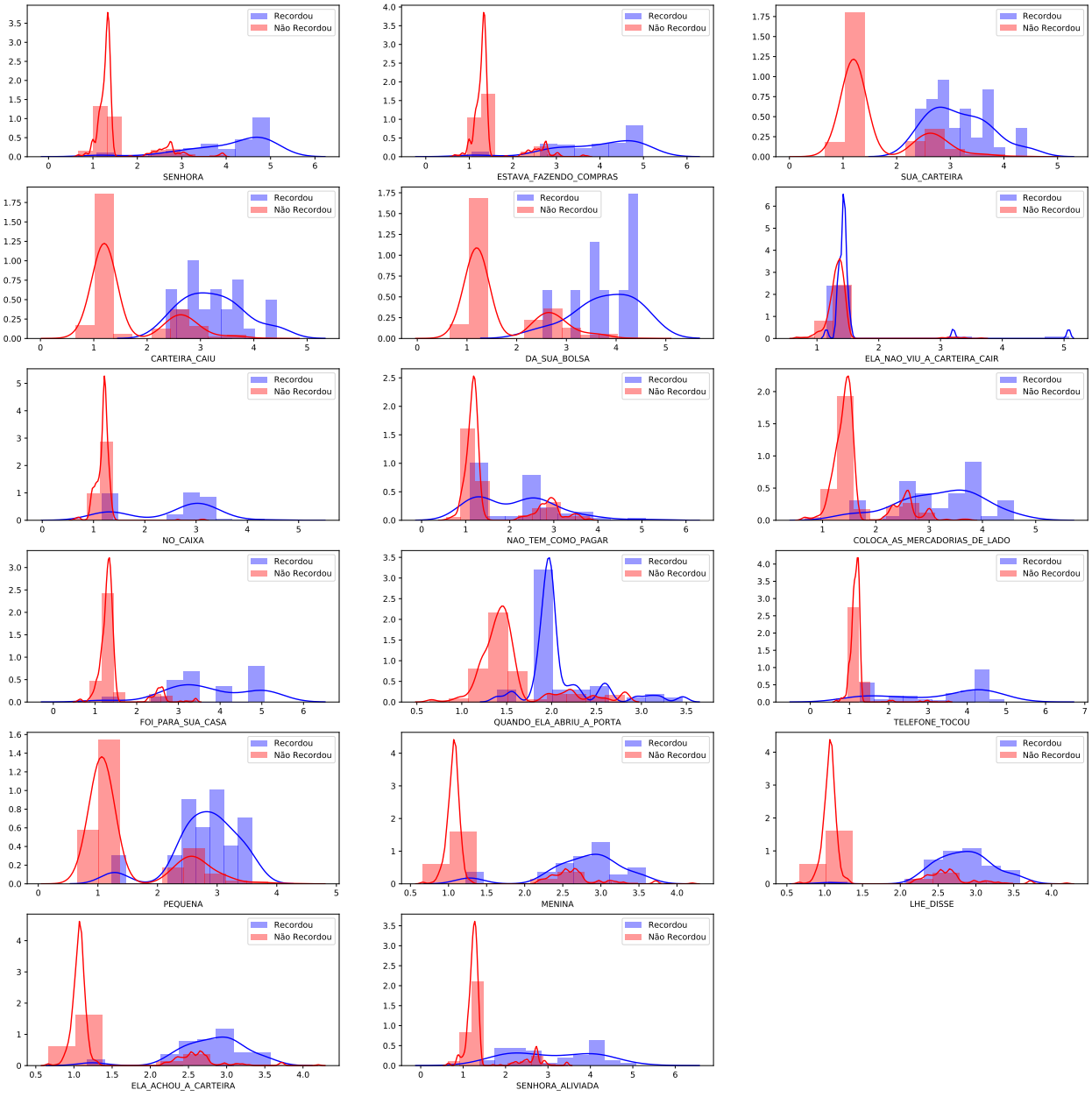

Figura 3: Histograma e distribuição acumulada para cada rótulo da *ABCD*.

S^1_i	C_1	C_2	C_3
P_1	0.55	0.10	0.01
P_1, P_2	0.60	0.15	0.30
P_1, P_2, P_3	0.30	0.13	0.40
P_2	0.40	0.18	0.35
P_2, P_3	0.30	0.30	0.55
P_2, P_3, P_4	0.15	0.20	0.70
P_3	0.20	0.15	0.48
P_3, P_4	0.30	0.12	0.60
P_4	0.15	0.08	0.05

Valor máximo para cada classe →

	C_1	C_2	C_3
S^1_i	0.60	0.30	0.70

Aplicação do ponto de corte para definir as classes finais →

$S^1_i = C_1, C_2$

Figura 4: Aplicação do método de *Chunking* para uma sentença com quatro palavras de conteúdo.

Na *ABCD*, cada participante produz duas narrativas, uma imediatamente após ouvir a história e a outra após 30 minutos. Para não enviesar a avaliação, os pares de narrativas produzidas foram alocados no conjunto de treino ou no conjunto de teste. Para a *BALE*, agrupamos as narrativas dos idosos com CCL e DA, dado o baixo número de idosos com CCL.

A Tabela 6 apresenta os resultados obtidos no conjunto de dados da *ABCD*. Por se tratar de um problema multirrótulo, reportamos a Precisão micro (Pr_{micro}), Precisão macro (Pr_{macro}), $F1$ micro ($F1_{micro}$), $F1$ macro ($F1_{macro}$), *SubsetAccuracy*, e o *HammingLoss*.

O método *baseline* Casamento Exato obteve os valores mais baixos, já o método *baseline* Inferência obteve os melhores resultados para precisão e o *HammingLoss*, mas foi superado pelo método de similaridade semântica *STS* nas outras medidas.

A Tabela 7 mostra os resultados obtidos no conjunto de dados da BALE. Os métodos *baselines* Casamento Exato e Inferência obtiveram os melhores resultados para precisão, enquanto o método *STS* obteve os melhores resultados em $F1$ e *SubetAccuracy*.

O objetivo final da pesquisa é criar um classificador para narrativas de testes neuropsicológicos de idosos saudáveis e idosos com comprometimento cognitivo (CCL e DA), para poder identificar os primeiros sinais de problemas cognitivos. Neste artigo, avaliamos o desempenho de classificadores utilizando unidades de informação como atributos. Na próxima seção, avaliamos se os métodos com medidas altas de $F1$ e *SubsetAccuracy* conseguem obter resultados de classificação próximos da anotação manual.

5 Classificação Automática de Narrativas visando uma Triagem Automática de Pacientes

Realizamos duas tarefas de classificação automática de narrativas, uma para cada conjunto de dados avaliado neste trabalho.

O conjunto *ABCD* possui as classes CCL e Controles Saudáveis e o conjunto BALE as classes CCL e DA, que foram agrupadas, e contrastadas com os Controles Saudáveis. Nessas duas tarefas, utilizamos as unidades de informação como vetores de atributos binários. Desta forma, cada narrativa da *ABCD* e da BALE possui 17 e 21 atributos, respectivamente. Avaliamos seis algoritmos de aprendizado de máquina: *SVM* (com kernel liner e RBF), *Naïve Bayes*, Árvores de decisão, *Gradient Boosting*, e *KNN*, implementados no *scikit-learn* (Pedregosa et al., 2011) versão 0.21.2, com os hiperparâmetros *default*.

Algumas particularidades destes métodos são destacadas abaixo:

Naïve Bayes é um dos algoritmos de aprendizado de máquina mais simples, pois assume que os atributos são independentes;

SVM é um algoritmo de classificação linear; sua função de otimização busca encontrar o hiperplano com margem máxima. Para esse algoritmo, utilizamos o *kernel* linear e o Radial Basis Function.

Árvores de decisão é um algoritmo que recursivamente particiona o espaço de entrada, geralmente de forma binária, definindo um modelo local em cada região resultante do espaço de entrada. É possível visualizar o modelo final na forma de uma árvore, em que cada partição representa um nó.

Gradient Boosting utiliza diversas árvores de decisão; cada árvore é treinada de forma sequencial para corrigir os erros da anterior.

KNN pertence à categoria *lazy*, pois não necessita de uma fase de treinamento para predizer um novo exemplo; busca no conjunto de treinamento os k exemplos mais similares e retorna o rótulo mais frequente.

Para as tarefas de classificação binária, os conjuntos de dados foram balanceados. Para a *ABCD*, utilizamos 12 idosos por grupo (Controle e CCL), sendo que cada idoso produziu 2 narrativas. O conjunto de dados final para a *ABCD* possui 48 narrativas. No caso da BALE, agrupamos os pacientes com CCL e DA (o grupo contém 16 narrativas), e selecionamos de forma randômica 16 narrativas do grupo de Controle. O conjunto de dados final para a BALE possui 32 narrativas.

Para a avaliação, utilizamos *10-fold-cross-validation* e a métrica de acurácia. Comparamos os métodos desenvolvidos para a identificação de unidades de informação e os *baselines* com a anotação manual, para analisar o impacto dos métodos na classificação final.

Os métodos *STS* e *Chunking* necessitam de uma busca de hiperparâmetros. Foi utilizada a metodologia *10-fold-cross-validation* para construir o novo conjunto de dados e este foi utilizado na avaliação dos classificadores.

Os resultados obtidos por todos os modelos na *ABCD*, em termos de acurácia, são apresentados na Tabela 8. Na segunda coluna da tabela,

Método	$\mathbf{Pr_{macro}}$		$\mathbf{Pr_{micro}}$		$\mathbf{F1_{macro}}$		$\mathbf{F1_{micro}}$		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,570	0,469	0,903	0,758	0,337	0,283	0,246	0,233	0,246	0,169	0,078	0,081
Inferência	0,858	**0,793**	0,891	**0,873**	0,552	0,531	0,500	0,478	0,348	0,384	0,062	**0,062**
Chunking	0,705	0,699	0,587	0,577	0,640	**0,624**	0,668	**0,656**	0,668	**0,395**	0,076	0,076
STS	0,651	0,569	0,595	0,552	0,672	0,598	0,670	0,552	0,670	0,273	0,072	0,081

Tabela 6: Resultados da identificação de unidades de informação na *ABCD*

Método	$\mathbf{Pr_{macro}}$		$\mathbf{Pr_{micro}}$		$\mathbf{F1_{macro}}$		$\mathbf{F1_{micro}}$		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,680	**0,740**	0,830	**0,930**	0,510	0,540	0,440	0,460	0,430	0,300	0,040	**0,040**
Inferência	0,625	0,690	0,808	0,781	0,485	0,519	0,359	0,404	0,447	0,324	0,042	0,050
Chunking	0,620	0,580	0,510	0,480	0,600	0,570	0,630	0,560	0,460	0,340	0,060	0,070
STS	0,680	0,640	0,670	0,650	0,740	**0,700**	0,720	**0,650**	0,520	**0,430**	0,030	**0,040**

Tabela 7: Resultados da identificação de unidades de informação na BALE

apresentamos os resultados para a anotação manual, que trouxe valores próximos para dois dos seis algoritmos de aprendizado (Árvores de Decisão e *Naïve Bayes*). Em geral, o classificador de Árvores de Decisão apresenta diferenças negativas maiores entre os valores da anotação manual e dos quatro modelos automáticos.

Dentre os dois métodos propostos para a identificação de unidades de informação (*STS* e *Chunking*), os melhores desempenhos para a *ABCD* foram do método de *Chunking*. Já o método de Inferência apresentou, em geral, resultados melhores do que o *STS*.

Na Tabela 9, são dispostos os resultados dos modelos na BALE. Para a anotação manual, tivemos quatro empates no desempenho de classificadores. O método *Chunking* apresenta diferenças negativas maiores entre os valores da anotação manual para todos os classificadores. Já o método *STS* apresenta as menores diferenças.

Em geral, o método *baseline* Casamento Exato superou os métodos automáticos propostos de identificação de unidades de informação.

Resumindo, os métodos com desempenhos adequados para a identificação de unidades de informação trazem resultados para classificação próximos da anotação humana, e as unidades de informação auxiliam mais na classificação final de pacientes com DA *versus* Controles Saudáveis (caso da BALE). Já para a *ABCD*, em que temos dois grupos balanceados de idosos saudáveis e com CCL, foi mais difícil separar as classes, corroborando com resultados da literatura (Santos et al., 2017; Fraser et al., 2019). Portanto, seguindo os trabalhos da literatura, há necessidade de trazer mais atributos para a classificação de pacientes saudáveis e com CCL, para melhorar o desempenho da classificação final dos pacientes.

6 Conclusões e Trabalhos Futuros

Este trabalho tratou de duas avaliações no cenário clínico: (i) avaliou métodos de similaridade semântica textual para a tarefa de identificação de unidades de informação em narrativas de recontos, e (ii) usou as unidades recuperadas como atributos para a classificação binária dos grupos idosos saudáveis *versus* idosos com comprometimento cognitivo, avaliando vários algoritmos de aprendizado de máquina.

Como observado na revisão dos trabalhos da literatura para identificação automática de unidades de informação em recontos, a grande dificuldade de utilizar listas de palavras para cada unidade de informação é a necessidade de um trabalho humano e subjetivo, pois nem sempre a lista possui todas as paráfrases/sinônimos possíveis. Métodos de *clustering* são úteis, pois conseguem criar automaticamente as unidades de informação, entretanto, podem gerar unidades pouco representativas ou não relacionadas às unidades que um dado teste neuropsicológico avalia.

Em estudos envolvendo análise de narrativas clínicas, geralmente a quantidade de dados é limitada, dado o alto custo de aquisição dos dados. Por se tratar de uma tarefa multirrótulo (identificação de unidades de informação), o cenário tratado neste artigo é ainda mais desafiador. Neste artigo, contornamos essas limitações aproximando a tarefa de identificação automática de unidades de informação em narrativas com similaridade semântica.

Avaliamos um método de similaridade semântica que explora a similaridade de representações vetoriais obtidas por *embeddings* e outro que se destacou na avaliação ASSIN, e transformamos o problema multirrótulo em

Método	Manual	Casamento Exato	Inferência	*Chunking*	*STS*
Árvore de Decisão	**0,638**	0,475	0,475	0,525	**0,538**
Gradient Boosting	0,538	0,463	**0,625**	0,550	0,500
KNN	0,575	**0,513**	0,525	**0,663**	0,413
SVM-Linear	0,500	0,475	0,588	0,525	0,488
SVM-RBF	0,563	0,363	0,463	0,463	0,488
Naïve Bayes	0,625	0,425	0,588	0,638	0,525

Tabela 8: Resultados da classificação utilizando as unidades de informações na *ABCD*

Método	Manual	Casamento Exato	Inferência	*Chunking*	*STS*
Árvore de Decisão	0,600	0,700	0,525	0,400	0,600
Gradient Boosting	**0,675**	0,725	0,450	0,400	0,575
KNN	0,650	0,575	0,475	0,525	0,625
SVM-Linear	**0,675**	0,625	0,600	0,550	0,625
SVM-RBF	**0,675**	0,525	**0,675**	**0,625**	**0,725**
NaiveBayes	**0,675**	**0,775**	0,550	0,550	0,700

Tabela 9: Resultados da classificação utilizando as unidades de informações na BALE

problemas de classificação binária, encontrando um ponto de corte para o valor de similaridade de cada unidade de informação. Desse forma, conseguimos superar ambos os *baselines* para os dois conjuntos de dados avaliados.

O uso de uma representação densa para sentenças é comum na literatura. Aqui, adotamos uma combinação com a média dos *embeddings* das palavras como proposto por Hartmann (2016), mas nos últimos anos essa abordagem vem sendo superada por métodos mais complexos como *ELMo* (*Embeddings from Language Models*) (Peters et al., 2018) ou *BERT* (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). Como trabalhos futuros, pretendemos explorar esses modelos mais atuais, pois acreditamos que com sistemas melhores de similaridade semântica podemos obter métodos de identificação de unidades de informação também melhores.

Outro ponto para investigações futuras é a utilização de mais atributos para a classificação final, como os propostos em Santos et al. (2017). Observamos na avaliação dos classificadores finais deste trabalho que separar conjuntos com características próximas como os da *ABCD* (CCLs *versus* Controles Saudáveis) é mais difícil do que a classificação final de pacientes com DA *versus* Controles Saudáveis (caso da BALE). Novos atributos linguísticos e da representação de narrativas via redes complexas podem contribuir com essa tarefa.

Por fim, cabe também avaliar o desempenho dos métodos de identificação de unidades de informação em narrativas de recontos, usando métodos de segmentação automática das narrativas, como os explorados em Treviso et al. (2017a,b) e Treviso & Aluísio (2018), mas retreinados com os datasets de Testes Neuropsicológicos em Português do Brasil[4], disponibilizados publicamente recentemente.

Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, processos números 130100/2015-3, 155137/2015-8, e 153047/2016-0, e também contou com o apoio da Google via programa *Google Research Awards for Latin America*.

Referências

Abbott, Alison. 2011. A problem for our age. *Nature* 475(7355). S2–S4. doi 10.1038/475S2a.

de Abreu, Izabella Dutra, Orestes V. Forlenza & Hélio Lauar de Barros. 2005. Demência de alzheimer: correlação entre memóriaria e autonomia. *Revista de Psiquiatria Clınica* 32. 131–136. doi 10.1590/S0101-60832005000300005.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea et al. 2015. Semeval-2015 task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Em *9th International Workshop*

[4] https://github.com/nilc-nlp/DNLT-BP

on Semantic Evaluation (SemEval), 252–263. 10.18653/v1/S15-2045.

Agirre, Eneko, Mona Diab, Daniel Cer & Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. Em *1st Joint Conference on Lexical and Computational Semantics-Volume 1:*, 385–393.

Bayles, Kathryn & C. K. Tomoeda. 1993. *ABCD: Arizona battery for communication disorders of dementia*. Tucson, AZ: Canyonlands Publishing.

Becker, James T., François Boiler, Oscar L Lopez, Judith Saxton & Karen L. McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6). 585–594. 10.1001/archneur.1994.00540180063015.

Bergstra, James, Dan Yamins & David D. Cox. 2013. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Em *12th Python in Science Conference (SCIPY)*, 13–20.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146. 10.1162/tacl_a_00051.

Chapman, Sandra Bond, Jennifer Zientz, Myron Weiner, Roger Rosenberg, William Frawley & Mary Hope Burns. 2002. Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer disease & Associated Disorders* 16(3). 177–186. 10.1097/00002093-200207000-00008.

Clemente, Rená & Sergio Ribeiro-Filho. 2008. Comprometimento Cognitivo Leve: aspectos conceituais, abordagem clínica e diagnóstica. *Revista do Hospital Universitário Pedro Ernesto* 7(1). 68–77.

Cromnow, Karolina & Tove Landberg. 2009. *Skriftliga beskrivningar av bilden Kakstölden. Insamling av referensvärden från friska försökspersoner*: Division of Speech and Language Pathology, Karolinska institute. Tese de Mestrado.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. Em Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini & Florence d'Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, 177–190. 10.1007/11736790_9.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

Fichman, Helenice Charchat, Rosinda Martins Oliveira & Conceição Santos Fernandes. 2011. Neuropsychological and neurobiological markers of the preclinical stage of alzheimer's disease. *Psychology & Neuroscience* 4(2). 245–253. 10.3922/j.psns.2011.2.010.

Fleming, Valarie B. & Joyce L. Harris. 2008. Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology* 22(7-8). 729–740. 10.1080/02687030701803762.

Fonseca, Erick Rocha, Leandro Borges Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliaçao de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Forbes-McKay, K.E. & A. Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological Sciences* 26(4). 243–254. 10.1007/s10072-005-0467-9.

Fraser, Kathleen C., Kristina Lundholm Fors & Dimitrios Kokkinakis. 2019. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language* 53. 121–139. 10.1016/j.csl.2018.07.005.

Fraser, Kathleen C., Jed A. Meltzer & Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49(2). 407–422. 10.3233/JAD-150520.

Freitas, Maria Isabel D'Ávila. 2010. *Habilidades linguísticas de pacientes com demência vascular: estudo comparativo com a doença de alzheimer*: Universidade de São Paulo. Tese de Doutoramento.

Frota, Norberto Anízio Ferreira, Ricardo Nitrini, Benito Pereira Damasceno, Orestes Forlenza, Elza Dias-Tosta, Amauri B da Silva, Emílio Herrera Junior & Regina Miksian Magaldi. 2011. Critérios para o diagnóstico de doença de Alzheimer. *Dementia & Neuropsychologia* 5(supl 1). 5–10. 10.1590/S1980-57642011DN05030002.

Garcia, Flavia Helena Alves & Letícia Lessa Mansur. 2006. Habilidades funcionais de comunicação: idoso saudável. *Acta fisiátrica* 13(2). 87–89.

Goodglass, Harold & Edith Kaplan. 1983. *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger 2nd edn.

Hartmann, Nathan Siegle. 2016. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. *Linguamática* 8(2). 59–64.

Hodges, John R., Karalyn Patterson, Naida Graham & Kate Dawson. 1996. Naming and Knowing in Dementia of Alzheimer's Type. *Brain and Language* 54(2). 302–325. 10.1006/brln.1996.0077.

Hübner, Lilian Cristine, Fernanda Loureiro, Bruna Tessaro, Ellen Siqueira, Gislaine Jerônimo & Anderson Smidarle. 2019. BALE: bateria de avaliação da linguagem no envelhecimento. Em Nicolle Zimmermann, François Delaere & Rochele Paz Fonseca (eds.), *Tarefas de avaliação neuropsicológica para adultos: memória e linguagem*, vol. 3, Memnon.

Kaufman, Leonard & Peter J. Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

Kintsch, Walter & Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5). 363–394. 10.1037/0033-295X.85.5.363.

Liang, Percy, Ben Taskar & Dan Klein. 2006. Alignment by agreement. Em *Human Language Technology Conference of the NAACL*, 104–111. 10.3115/1220835.1220849.

López, Roque & Thiago Pardo. 2015. Experiments on sentence boundary detection in user-generated web content. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 227–237. 10.1007/978-3-319-18111-0_18.

Mansur, Letícia Lessa, Maria Teresa Carthery, Paulo Caramelli & Ricardo Nitrini. 2005. Linguagem e cognição na doença de Alzheimer. *Psicologia: reflexão e crítica* 18(3). 300–307. 10.1590/S0102-79722005000300002.

Mapstone, Mark, Amrita K. Cheema, Massimo S. Fiandaca, Xiaogang Zhong, Timothy R. Mhyre, Linda H. MacArthur, William J. Hall, Susan G. Fisher, Derick R. Peterson, James M. Haley et al. 2014. Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine* 20(4). 415–418. 10.1038/nm.3466.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi & Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 216–223.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo & Bento C. Dias da Silva. 2008. A base de dados lexical e a interface web do tep 2.0-thesaurus eletrônico para o português do brasil. Em *VI Workshop em Tecnologia da informação e da linguagem humana (TIL)*, 390–392.

McKhann, Guy M, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux et al. 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7(3). 263–269. 10.1016/j.jalz.2011.03.005.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Em *International Conference on Learning Representations (ICLR)*, s/p.

Nasreddine, Ziad S, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings & Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53(4). 695–699. 10.1111/j.1532-5415.2005.53221.x.

Pakhomov, Serguei V. S., Glenn E Smith, Dustin Chacon, Yara Feliciano, Neill Graff-Radford, Richard Caselli & David S. Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology* 23(3). 165–177. 10.1097/WNN.0b013e3181c5dde3.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 2227–2237. 10.18653/v1/N18-1202.

Preti, Dino (ed.). 2005. *O discurso oral culto*. São Paulo: Associação Editorial Humanitas 3rd edn. Projetos Paralelos. V.2.

Prud'hommeaux, Emily & Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics* 41(4). 549–578.

Santos, Leandro, Edilson Anselmo Corrêa Júnior, Osvaldo Oliveira Jr, Diego Amancio, Letícia Mansur & Sandra Aluísio. 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. Em *55th Annual Meeting of the Association for Computational Linguistics*, 1284–1296. 10.18653/v1/P17-1118.

Santos, Leandro, Lilian Cristiane Hübner, Anderson Dick Smidarle, Letícia Mansur & Sandra Aluísio. 2019. Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português. Em *XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 253–261.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou & Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. Em *Demonstrations Session at European Association for Computational Linguistics (EACL)*, 102–107.

Treviso, Marcos, Christopher Shulby & Sandra Aluísio. 2017a. Evaluating word embeddings for sentence boundary detection in speech transcripts. Em *XI Brazilian Symposium in Information and Human Language Technology (STIL)*, 151–160.

Treviso, Marcos, Christopher Shulby & Sandra Aluísio. 2017b. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 315–325. Association for Computational Linguistics.

Treviso, Marcos Vinícius & Sandra Maria Aluísio. 2018. Sentence segmentation and disfluency detection in narrative transcripts from neuropsychological tests. Em *Computational Processing of the Portuguese Language (PROPOR)*, 409–418. 10.1007/978-3-319-99722-3_41.

Tsoumakas, Grigorios, Ioannis Katakis & Ioannis Vlahavas. 2009. Mining multi-label data. Em Maimon O. & Rokach L. (eds.), *Data Mining and Knowledge Discovery Handbook*, 667–685. Springer, Boston, MA.

Wallin, Anders, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson et al. 2016. The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism* 36(1). 114–131. 10.1038/jcbfm.2015.147.

Wechsler, David. 1997. *Wechsler memory scale - third edition.* San Antonio, TX: The Psychological Corporation.

Wortmann, Marc. 2012. Dementia: a global health priority-highlights from an ADI and World Health Organization report. *Alzheimer's Research & Therapy* 4(5). 40. 10.1186/alzrt143.

Yancheva, Maria & Frank Rudzicz. 2016. Vector-space topic models for detecting Alzheimer's disease. Em *54th Annual Meeting of the Association for Computational Linguistics*, 2337–2346. 10.18653/v1/P16-1221.