UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

FELIPE S. F. PAULA

Identificação de Condições Clínicas em Testes de Fluência Verbal

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dra. Aline Villavicencio
Co-orientador: Dr. Rodrigo Wilkens

Porto Alegre
2018

“What might have been and what has been
Point to one end, which is always present.
Footfalls echo in the memory
Down the passage which we did not take
Towards the door we never opened
Into the rose-garden.”

— T. S. ELLIOT

AGRADECIMENTOS

A Nice e ao Jorge, meus pais, que com todo o amor do mundo me trouxeram a esse lugar. Os dois nunca mediram esforços e sempre me incentivaram a buscar os meus sonhos, por isso, a cada conquista minha, sempre lembrarei da dedicação dos dois. Também agradeço a Aline e ao Rodrigo, por serem modelos de pesquisadores que pretendo seguir. Ambos sempre me motivaram e, com absoluta certeza, o contato com eles me elevou a um patamar mais alto. Estendo esse agradecimento ao Marco e aos colegas do grupo de Processamento de Linguagem Natural da UFRGS: Diego, Matheus, Machado, Alex, Sílvio e Zilio. Agradeço, em especial, a Paola que há mais de 9 anos entende minhas ausências e meus problemas (muitas vezes exagerados). Esse trabalho com certeza seria pior sem o apoio e o afeto dela. Por último agradeço ao professor Roberto da Física da UFRGS e aos professores da Informática da UFRGS, com quem aprendi muito.

RESUMO

Abordagens computacionais têm sido cada vez mais presentes na análise de resultados de tarefas neuropsicológicas, pois podem ser menos custosos e podem identificar dinâmicas que são dificilmente detectadas por humanos. Um tipo de tarefa apropriada para esse tipo de análise é o teste de fluência verbal. Nessa tarefa, é pedido que uma pessoa fale uma sequência de palavras em um tempo limitado. É hipotetizado que a identificação de subsequências de palavras relacionadas, chamadas de cadeias semânticas, pode ajudar no diagnóstico de doenças. Nesse trabalho, investigamos abordagens computacionais para a detecção de cadeias semânticas em tarefas de fluência verbal, como a identificação de cadeias através de aprendizado de máquina e como a detecção através da média de similaridades entre palavras. Adicionalmente, avaliamos a performance dessas cadeias na identificação de doenças neuropsicológicas. A detecção de cadeias semânticas é validada através de um experimento no qual perguntamos para julgadores se existe uma quebra semântica em uma sequência de palavras. Também propomos descritores baseados nessas cadeias e os avaliamos contra o estado da arte na identificação de doença de Alzheimer e Comprometimento Cognitivo Leve. O melhor detector de cadeias semânticas proposto obtém uma medida-f de 0.81 no conjunto de dados obtido dos anotadores. Quando combinamos os descritores baseados em cadeias e os descritores estado da arte, não houve um aumento significativo de performance. Além disso, esses descritores não superam o baseline baseado no número de palavras faladas. Esse fato reforça a necessidade de que precisamos entender melhor as técnicas de detecção de cadeias semânticas em tarefas de fluência verbal e as suas relações com diversas condições neuropsicológicas.

Palavras-chave: Processamento de Linguagem Natural. Aprendizado de Máquina. Psicolinguística Computacional.

Identification of Clinical Conditions in Verbal Fluency Tests

ABSTRACT

Computational approaches have been increasingly present in the analysis of results of neuropsychological tasks, as they may be less costly and can identify dynamics that are difficult to detect by humans. One type of task appropriate for this type of analysis is the verbal fluency test. In this test, a person is asked to speak a sequence of words in a short time. It is hypothesized that the identification of subsequences of related words, called semantic chains, can help in the diagnosis of diseases. We investigate computational approaches for the detection of semantic chains in tasks of verbal fluency, and in addition, we evaluate the performance of these chains in the identification of neuropsychological diseases. The detection of semantic chains is validated through an experiment in which we ask judges if there is a semantic break in a sequence of words. We propose features based on these chains and evaluate them against the state of the art in the identification of Alzheimer's disease and Mild Cognitive Impairment. The best proposed semantic chain detector obtains a f-measure of 0.81 in the dataset obtained from the annotators. When we combined the chain-based features and the state of the art features, there was no significant increase in performance. In addition, these features do not exceed the baseline based on the number of spoken words. This fact reinforces the need to better understand the techniques of detecting semantic chains in tasks of verbal fluency and their relations with various neuropsychological conditions

Keywords: natural language processing, machine learning, computational psycholinguistics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

TFV	Teste de Fluência Verbal
ALZ	Doença de Alzheimer
aMCI	Comprometimento Leve Amnéstico (amnestic mild cognitive deficit)
a+mdMCI	Comprometimento Leve Amnéstico Multidomínio (multidomain amnestic mild cognitive deficit)
NB	Naive Bayes
RF	Random Forest
RL	Regressão Logística
SVM	Máquinas de Vetor de Suporte (support vector machines)

SUMÁRIO

1 INTRODUÇÃO

A linguagem é um fenômeno complexo produto da cognição humana. Para entendê-la, é necessária uma discussão interdisciplinar que envolve diversas áreas de conhecimento como a linguística, a psicologia, a neurociência, a ciência da computação, entre outras. Uma abordagem para a compreensão desse fenômeno, é observar as produções de diferentes grupos com em tarefas psicolinguísticas controladas e tentar explicar o comportamento computacionalmente.

Os Testes de Fluência Verbal (TFV) têm sido uma das fontes de entendimento da memória semântica e do léxico mental humano. Esse teste consiste em pedir para o indivíduo falar o máximo de palavras que pertençam a mesma categoria semântica, como animais, itens de supermercado ou vegetais. Desde os anos 40, o estudo da tarefa de fluência é chave para a compreensão dos processos cognitivos presentes nos processos recuperação e armazenamento das palavras na memória (BOUSFIELD; SEDGEWICK, 1944).

A aplicação do mesmo teste em pessoas com condições clínicas conhecidas é uma maneira de entender melhor a dinâmica dos mecanismos da linguagem. Pessoas com diferentes tipos de condições, como lesões em áreas específicas do cérebro, podem produzir testes com alterações interessantes. Dessa maneira, estudando computacionalmente essas alterações, podemos fazer uma "engenharia reversa" para entender o comportamento normal através de comportamentos anômalos.

Tarefas psicolinguísticas, como TFV, além de servir para desvendar processos cognitivos, também servem para o diagnóstico de doenças. Profissionais da saúde, como os de avaliação neuropsicológica, submetem os pacientes à baterias de testes que envolvem diferentes domínios da cognição, como memória, controle mental e linguagem.

Os TFV estão presentes em diversas dessas baterias e por si só, podem indicar a presença de diversas doenças (SHAO et al., 2014). Uma dinâmica conhecida nos TFV é a dinâmica de *cluster e switch* (BOUSFIELD; SEDGEWICK, 1944) e pode ser indicativo da presença de demências (TROYER et al., 1998). Essa dinâmica fala sobre características de subgrupos semânticos nas sequências de palavras dos TFV.

Os profissionais da saúde analisam os testes identificando os subgrupos semânticos através de taxonomias que ditam quais palavras pertencem a quais grupos. Essa abordagem pode apresentar problemas pois ela é essencialmente subjetiva. Diferentes palavras podem pertencer a diferentes categorias dependendo do entendimento do profissional.

1.1 Objetivo

Nesse trabalho, vamos investigar maneiras de aproximar os subgrupos semânticos nos TFV e usar essa informação para identificar as condições clínicas. Mais precisamente, estamos interessados em responder as perguntas:

Pergunta 1: Como identificar computacionalmente as quebras semânticas em sequências de palavras?

Pergunta 2: Qual o poder de predição das características provindas da identificação computacional na identificação de casos clínicos?

Para responder a primeira pergunta, propomos abordagens computacionais para o problema que serão validadas por dados provenientes de identificação humana dessas quebras semânticas. Esses dados provêm de um experimento em que estudamos como esses subgrupos são identificados por pessoas. Adicionalmente, fazemos um estudo quantitativo e qualitativo do experimento para melhor entender a tarefa. Para responder a segunda pergunta, estudamos a performance de descritores na classificação de sujeitos clínicos. Essa tarefa de classificação é avaliada com diferentes combinações de identificação de subgrupos semânticos, conjuntos de descritores e algoritmos de classificação. Reproduzimos o sistema estado da arte e estudamos se os melhores descritores adicionam poder de classificação.

O trabalho está organizado da seguinte maneira. O capítulo 2 faz uma breve revisão sobre trabalhos relevantes no estudo computacional da tarefa de fluência. O capítulo 3 fala sobre as principais técnicas e metodologia utilizada no trabalho. Também, nesse capítulo, detalharemos como procedemos para coletar os dados com anotadores. No capítulo 4, reportaremos o resultado do experimento juntamente de uma avaliação das respostas. Vamos estudar a performance de técnicas de identificação de subgrupos semânticos e também de identificação de sujeitos clínicos. O capítulo 5 descreve as contribuições do trabalho e direções de pesquisa futura.

2 TRABALHOS RELACIONADOS

Nesse capítulo trataremos das abordagens computacionais para a modelagem das tarefas de fluência. Na Seção 2.1, falamos sobre os casos clínicos tratados aqui. Na Seção 2.2, das cadeias semânticas e switches. Na Seção 3.3, falamos sobre abordagens para a modelagem do léxico mental e de como demências podem ser representadas computacionalmente. Na Seção 3.4, revisaremos alguns trabalhos que usam informação de cadeias semânticas para identificar grupos clínicos.

2.1 Doenças neurodegenerativas e condições relacionadas

A doença de Alzheimer é uma das principais causas de demência em adultos idosos (BRAMBATI et al., 2009). Embora não seja uma doença rara, é ainda mal compreendida. Suas principais características são a perda da memória e o declínio não reversível das habilidades cognitivas dos pacientes. Um aspecto terrível dessa doença é a perda do córtex cerebral. A doença começa com nenhum sintoma detectável, depois, progride para a perda de memória a curto prazo, para perda de linguagem, então as habilidades motoras e, em seguida, o declínio cognitivo geral, ou seja, o paciente perde totalmente sua independência.

Como o Alzheimer é uma doença altamente destrutiva, muitos esforços foram empregados em prevenção e detecção precoce. Um conceito útil que foi criado para ajudar a identificar os sinais iniciais de demência é a Comprometimento Cognitivo Leve Amnéstico (*amnestic mild cognitive deficit* - aMCI) (BRAMBATI et al., 2009) . Essa condição é principalmente caracterizada como uma queixa de memória por parte do paciente ou por parte de sua família. As pessoas afetadas por aMCI apresentam um déficit de memória enquanto mantêm habilidades de vida diária, ou seja, elas são capazes de viver (até certo ponto) de forma independente. Outra condição é o Comprometimento Cognitivo Leve Amnéstico Multidomínio (*multidomain amnestic mild cognitive deficit* - a+mdMCI) (BRAMBATI et al., 2009). Nessa condição, além da queixa da memória, o paciente também mostra dificuldades em outros domínios da cognição, como por exemplo, linguagem.

Um critério para diagnóstico de Alzheimer (MCKHANN et al., 1984) exige a presença de múltiplos déficits cognitivos, como deficiência de memória, deficiência de linguagem, comprometimento das habilidades motoras e assim por diante. Além disso, o sujeito também deve apresentar comprometimento funcional (diminuição dos índices

de atividades diárias) e progressão geral dos sintomas. Os critérios de diagnóstico para aMCI normalmente requerem que o paciente apresente uma queixa de memória, mas os pacientes devem manter o seu funcionamento intelectual geral e a capacidade normal de vida diária (WINBLAD et al., 2004). Para a+mdMCI, também deve apresentar declínio de performance em tarefas relacionadas a mais de um domínio cognitivo (WINBLAD et al., 2004). Os níveis cognitivos dos indivíduos são avaliados através de baterias de exames neuropsicológicos. Um tipo de teste chave nas avaliações neuropsicológicas são os testes de fluência verbal.

Nos testes de fluência verbal, é pedido aos pacientes falar tantas palavras quanto forem possíveis em um tempo delimitado, evitando repetições. Esse tipo de tarefa é muito usado na avaliação neuropsicológica porque está diretamente relacionado aos processos cognitivos de busca de palavras na memória semântica.

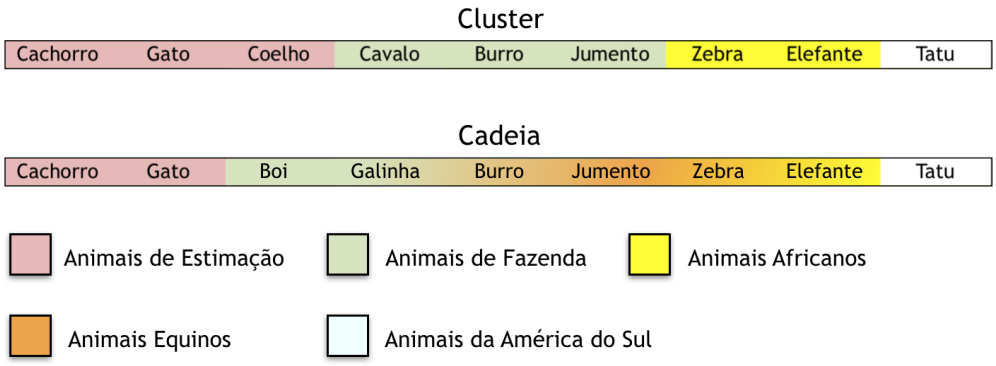
Esse teste geralmente é classificado em dois tipos: teste de fluência verbal fonêmica e teste de fluência verbal semântica. Na versão fonêmica, os participantes são convidados a falar tantas palavras quanto possível começando com a mesma letra. Na versão semântica, pede-se palavras pertencentes a mesma categoria, como itens de supermercado, vegetais ou animais. No restante desse trabalho focamos apenas na fluência verbal semântica.

2.2 Cadeias e switches

As pessoas tendem a produzir palavras em subgrupos naturalmente. Na categoria semântica de animais, por exemplo, uma sequência como “cão, gato, rato, cavalo, porco, vaca, ...” contém dois subgrupos semânticos (animais de estimação e animais de fazenda). Os subgrupos são chamados *clusters* e a alternância entre eles *switches*. A análise de clusters e switches pode revelar deficiências nas capacidades de linguagem de um falante e ser indicativo de uma condição clínica (TROYER et al., 1998).

Outra maneira de análise do teste é avaliar como o paciente navega através de suas representações de palavras. Isso é feito no estudo das *cadeias semânticas* (PAKHOMOV; HEMMY, 2014). Uma cadeia é uma seqüência de palavras relacionadas semanticamente. Por exemplo, a sequência “cão, gato, papagaio, galinha, vaca ”, possui dois clusters, animais de estimação e animais de fazenda, mas uma única cadeia, já que o gato e o papagaio são animais de estimação, o papagaio e a galinha são aves e, finalmente, frango e e vaca são animais de fazenda. A avaliação de *clusters* é mais estrita do que de cadeias, pois

Figura 2.1: Exemplo de sequências de palavras com identificação de clusters e cadeias. Na primeira sequência, temos switches entre coelho-cavalo, zebra-elefante e elefante-tatu. Na segunda sequência, temos switches entre gato-boi e elefante-tatu.



muitas vezes precisamos decidir qual cluster "fechar". Por exemplo, no exemplo anterior, poderíamos ter os clusters "cão e gato", "papagaio e galinha" e "vaca". Em uma avaliação de cadeias semânticas, olhamos para as palavras duas a duas, sequencialmente para decidir se existe quebra semântica ou não.Aqui nesse trabalho, usaremos a abordagem de cadeias semânticas, pois acreditamos que é uma boa aproximação do conceito de clusters. Na Figura 2.1, temos um ilustração de sequências com identificação de clusters e cadeias.

Na prática da avaliação neuropsicológica, existem listas de quais entidades pertencem a quais categorias. Essas taxonomias são subjetivas e podem apresentar problemas para o diagnóstico. Duas taxonomias diferentes podem influenciar o resultado do teste para um lado ou para o outro. Além disso, pode ser muito caro analisar manualmente esses testes. Nesse trabalho, vamos investigar métodos automáticos para a detecção dessas quebras semânticas nos TFV e como podemos identificar casos clínicos com essa informação.

2.3 Representações computacionais para fluência e associação de palavras

Em Borge-Holthoefer, Moreno e Arenas (2011), é proposta uma explicação computacional para o fato de que pessoas com Alzheimer em estágio inicial apresentar maior *priming* entre algumas palavras. Na tarefa de decisão lexical, é mostrada uma palavra, estímulo, e em seguida mais outra, alvo, que pode ser uma palavra ou pseudopalavra. Os indivíduos devem decidir se é a palavra é verdadeira ou falsa. Pessoas com Alzheimer inicial mostram tempos menores em alguns pares de palavras, o que indica um *priming*

maior. Para explicar isso, Borge-Holthoefer, Moreno e Arenas (2011) modela o léxico mental como um grafo ponderado contruído a partir de uma base de dados de associações entre palavras. As palavras são os nodos e as forças de associação são as arestas. Todas as arestas que saem de um nodo são normalizadas de maneira que somam 1. O *priming* entre palavras é modelado através dessa associação que dois nodos possuem. A doença é atua como um "ataque" ao grafo, em que a cada rodada da simulação, subtrai do peso de cada aresta um valor pequeno. Quando o peso de uma aresta chega a zero, ela é perdida. Por causa da degradação do grafo, arestas são perdidas, por causa da renormalização das arestas, conexões são reforçadas, assim explicando o fenômeno.

Em Hills, Jones e Todd (2012) e também em Abbott, Austerweil e Griffiths (2015), a busca de palavras na memória é explicada como um animal buscando comida em um ambiente. O léxico mental é modelado como uma rede de associações e o processo de busca é modelado como uma caminhada aleatória. É mostrado, que nas tarefas de fluência, os switches ocorrem de maneira análoga ao forrageamento ótimo. Isto é, um animal busca por comida em ambiente em diversos arbustos, quanto mais ele consome comida em um arbusto, mais difícil é encontrar comida nesse arbusto, então ele é confrontado com a decisão de ficar no mesmo lugar ou procurar outra fonte de comida. O problema do forrageamento é solucionado pelo teorema do valor marginal que fala que o "animal" permanece no mesmo lugar até que a sua taxa de ganhos seja maior que a taxa de ganhos média de todo o ambiente. Analogamente, as pessoas permanecem o mesmo cluster até que a sua taxa de ganhos, que no problema é o tempo entre palavras, atinja o valor médio de todas as palavras que ela fala.

Borge-Holthoefer, Moreno e Arenas (2011), Hills, Jones e Todd (2012) e Abbott, Austerweil e Griffiths (2015) nos motiva conceitualmente. Acreditamos que doenças como o Alzheimer degradam as estruturas do léxico mental, que por sua vez, interferem nos processos de busca que dependem dessas representações. Essas alterações devem gerar anomalias na produção de palavras nos testes. O forrageamento de palavras e o teorema do valor marginal nos inspira em uma abordagem heurística para identificação dos switches.

2.4 Abordagens computacionais para detecção de casos clínicos

Os trabalhos que identificam computacionalmente os grupos clínicos utilizando TFV se dividem em dois tipos:

- Detecção através de grafos (BERTOLA et al., 2014b)
- Detecção através de modelagem de clusters e cadeias semânticas (PRUD'HOMMEAUX; SANTEN; GLINER, 2017) (LINZ et al., 2017)

Na detecção baseada em grafos, introduzida em Bertola et al. (2014b), cada palavra no TFV é convertida em um nodo e se duas palavras são consecutivas, então uma aresta direcionada é criada entre elas. Em um teste que não houve repetições de palavras, o grafo seria uma "linha". Quando ocorrem repetições, o grafo se condensa, pois um nodo vai passar a ter mais uma aresta incidente. Nesse tipo de abordagem, utilizamos essa informação estrutural que obtemos dos testes para classificar os indivíduos nos grupos clínicos. A segunda abordagem utiliza os clusters ou cadeias para detectar as doenças. O desafio nesse tipo de técnica é a detecção das quebras semânticas. Para isso, são empregados anotadores para indentificação ou são empregados modelos semânticos para a detecção automática.

Em Prud'hommeaux, Santen e Gliner (2017), são comparadas as populações de crianças com desenvolvimento padrão e crianças no espectro do autismo. Para identificar cadeias semânticas, foram utilizadas as similaridades provenientes dos modelos Word2Vec (MIKOLOV et al., 2013) e WordNet. As cadeias foram identificadas através de aprendizado supervisionado. Uma porção dos TFV possuía informação das cadeias, logo, essa informação foi utilizada para treinamento. A classificação foi feita a partir de máquinas de vetor de suporte. É reportada uma performance nessa detecção de AUC entre 0.6 e 0.85, porém os autores fazem uma ressalva que estavam não interessados em aprender totalmente a anotação humana, pois acreditam em uma detecção computacional sem a tendenciosidade do anotador. As palavras foram analisadas duas a duas e foi extraída a similaridade Word2Vec e a similaridade de caminho da WordNet como descritores.

Esse trabalho é pertinente, pois a identificação por humanos não encontra diferenças entre os grupos (desenvolvimento padrão e autismo), enquanto a abordagem computacional encontra, $t = 2.41$, $p < 0.05$. É sugerido que os modelos computacionais podem conter mais informações que as taxonomias e, assim, ser um método alternativo na análise dos testes.

Em Linz et al. (2017), os autores estão preocupados em predizer MCI utilizando semântica distribucional. Nesse trabalho, são propostas uma abordagem para detecção de clusters semânticos e uma abordagem para cadeias semânticas. Essas detecções são baseadas na média de similaridade do TFV, ou seja, é a média das permutações da palavras tomadas duas a duas. No caso do cluster, são somadas as representações vetoriais das

palavras da subsequência e essa é comparada (cosseno) com a representação da primeira palavra fora do cluster, se a similaridade for maior que a média das permutações, então temos um switch. As cadeias são identificadas comparando a similaridade de palavras adjacentes no teste com o limiar proposto.

As similaridades são obtidas através do Word2Vec treinado na Wikipedia em Francês e o FrWac. Para classificar os indivíduos entre controle e MCI, foram usados os descritores *número de switches* e *tamanho cadeia/cluster médio*. O melhor modelo automático obteve performance de revocação de 79, precisão de 75 e medida-f de 77. O modelo que usava as taxonomias baseadas em Troyer et al. (1998) obteve revocação de 74, precisão de 71 e medida-f de 72.

Em Bertola et al. (2014b) são aplicados testes de fluência verbal em indivíduos idosos sauindicaveis e idosos com algum tipo de déficit cognitivo. São no total quatro grupos: controle, MCI amnéstico de um único domínio, MCI amnéstico de múltiplos domínios e doença de Alzheimer. Nesse trabalho é apresentado a abordagem de separação dos grupos clínicos através de descritores baseadas em grafos. Como dito na sessão anterior, nessa abordagem, as palavras se transformam em um conjunto de vértices (isto é, não existe repetições) e se a palavra w_1 e a palavra w_2 são consecutivas no TFV, então, no grafo, existe uma aresta direcionada entre os vértices que representam a palavra w_1 e a palavra w_2. Depois que o grafo é formado, são calculadas medidas topológicas:

- **N**: Número de nodos
- **E**: Número de arestas
- **Diametro**: O maior caminho entre os caminhos mais curtos
- **ASP**: (average shortest path) Média dos caminhos mais curtos
- **Densidade**: Proporção de nodos e arestas $D = \frac{|E|}{|V|(|V|-1)}$

Com essas medidas e um classificador naive bayes, foram classificados os grupos clínicos e obtiveram os resultados descritos na Tabela 2.1. Como nesse trabalho usamos os mesmos dados, decidimos reproduzir os resultados para melhor avaliar os nossos modelos.

Tabela 2.1: Resultados reportados em Bertola et al. (2014b) para separação do grupo de controle dos grupos clínicos

	Controle × Alzheimer	Controle × aMCI	Controle × a+mdMCI
AUC	0.875	0.619	0.710

Prud'hommeaux, Santen e Gliner (2017) foi o primeiro trabalho a usar um modelo

distribucional em conjunção com a WordNet para encontrar as quebras semânticas. Aqui, também usaremos essa abordagem. Linz et al. (2017) é um trabalho que, assim como o nosso, identifica os switches usando a média de similaridade do TFV de cada indivíduo para localizar as quebras semânticas. Além disso, Linz et al. (2017) identifica MCI usando características de "número de switches" e "tamanho de cadeia média". Bertola et al. (2014b) será o trabalho que iremos comparar a performance de identificação de casos clínicos. Esse fato é justificado por usarmos o mesmo conjunto de dados e também por esse trabalho ser o estado da arte na identificação de Alzheimer, aMCI e a+mdMCI através de TFV.

3 MATERIAIS E MÉTODOS

Nesse capítulo apresentamos os dados utilizados para avaliar as abordagens estudadas (Seção 3.1), então apresentamos maneiras de identificação de switches (Seção 3.2) e os algoritmos utilizados para a classificação (Seção 3.3). Na Seção 3.4, apresentamos as medidas para representar computacionalmente os dados dos TFV de cada sujeito e na Seção 3.5 os principais algoritmos de classificação usados. Por fim, na Secao 3.6 apresentamos nossa metodologia de avaliação.

3.1 Testes de Fluência Verbal

Utilizamos os mesmos dados reportados em Bertola et al. (2014a), que consistem em 100 indivíduos de escolaridade uniforme e idosos. Essas pessoas passaram por uma bateria de avaliações no Centro de Referência à Saúde do Idoso Jenny de Andrade Faria, Hospital Clínico, Universidade Federal de Minas Gerais e foram classificados entre sem patologias (Ctrl), com Comprometimento Cognitivo Leve Amnéstico (a+MCI), com Comprometimento Cognitivo Leve Amnéstico Multidomínio (a+mdMCI) e com doença de Alzheimer (Alz). Cada grupo possui $n = 25$ indivíduos. Os clusters foram identificados por um especialista na área que seguiu o esquema similar ao de Troyer, Moscovitch e Winocur (1997). Aqui, vamos chamar essa indentificação de **taxonomia**.

3.2 Detecção heurística dos switches

Um teste de fluência verbal gera uma lista $\boldsymbol{L}$ com $N + 1$ palavras. O *perfil de similaridade* de $\boldsymbol{L}$ são as palavras tomadas duas as duas, como uma janela deslizante, e posteriormente a função de similaridade $s(w_i, w_j)$ é aplicada. Por exemplo, se temos um $\boldsymbol{L} = \{w_1, w_2, w_3, w_4, \ldots, w_N, w_{N+1}\}$, o perfil de similaridade correspondente é $\boldsymbol{x} = \{s(w_1, w_2), s(w_3, w_4), \ldots, s(w_N, w_{N+1})\}$. Com isso, temos um vetor $\boldsymbol{x}$ de dimensão N em que cada componente x_i representa a similaridade de um par de palavras consecutivas. A partir desse perfil, podemos localizar os pares em que a troca de contexto na representação semântica aconteceu. Para isso, temos os *mapas de switch* $\boldsymbol{m}$ do perfil, que é um vetor de dimensão N cujas componentes m_i possuem valor de 1 caso o par de palavras seja um switch e 0 caso contrário. Um algoritmo de detecção pode ser

considerado uma função $\Psi(\boldsymbol{x}) = \{\psi(x_1), \ldots, \psi(x_N)\}$ que avalia cada par de palavras.

3.2.1 Detecção baseada na média de similaridade

Uma abordagem heurística é extrair informação desse perfil de similaridade para detectar os switches. Em Hills, Jones e Todd (2012) e Abbott, Austerweil e Griffiths (2015) é indicado que existe uma relação muito forte entre a média de similaridade entre os pares de palavras de um indivíduo e os switches. Isso pode ser explicado intuitivamente. Espera-se que a similaridade de um item com outro item dentro de seu grupo semântico seja alta e que a similaridade desse item com outro item fora seja menor. Durante os TFV, na maior parte das vezes, as pessoas falam as palavras em grupos, com isso a média de similaridade estará próximo da média do que é considerado grupo.

Esse algoritmo considera que o par de palavras x_i é um switch quando o seu valor de similaridade é menor que a média de similaridades desse vetor. É uma transcrição do teorema do valor marginal para a busca na memória semântica.

$$\psi(x_i) = \begin{cases} 1 & \text{se } x_i < \sum_{j=1}^{N} \frac{x_i}{N} \\ 0 & \text{cc} \end{cases}$$

3.2.2 Detecção baseada em aprendizado de máquina

Assim como em Prud'hommeaux, Santen e Gliner (2017), também vamos usar aprendizado de máquina para detectar os switches. Em uma sequência de n palavras (com n par), queremos identificar se entre as palavras do meio existe uma quebra semântica. Por exemplo, para cada conjunto de quatro palavras $\{w_1 w_2 w_3 w_4\}$, que chamamos de *quadrigramas*, queremos responder se existe um switch entre $w_2 w_3$. Aqui, nesse trabalho, escolhemos $n = 4$ palavras, pois a decisão do switch pode ser influenciada pelas palavras do entorno e essa é a configuração mais simples que modela isso. Esse fato vai ser estudado experimentalmente no próximo capítulo.

Como descritores, utilizamos as similaridades GloVe e WordNet dos pares de palavra consecutivos. Vamos investigar a performance dos algoritmos random forest, regressão logística e SVM (máquinas de vetor de suporte) para a predição dos switches. Esses algoritmos foram escolhidos, pois o random forest apresenta uma performance boa

em uma gama grande de problemas, regressão logística é um modelo clássico e SVM foi a escolha de Prud'hommeaux, Santen e Gliner (2017).

3.3 Descritores para a identificação de casos clínicos

A partir da hipótese de que as características dos grupos semânticos presentes nos TFV podem dar pistas sobre a presença de doenças neurológicas (TROYER et al., 1998), usaremos os seguintes descritores:

swiches: Número de switches (quebras semânticas) em um teste.

cadeia média: Tamanho médio das cadeias semânticas nos testes.

cadeia max: O tamanho da maior cadeia em um teste.

cadeia min: Quantas cadeias mínimas existem no teste. Calculamos o tamanho da cadeia mínima, em seguida, vemos quantas cadeias de tamanho mínimo existem no teste. Dividimos esse número pelo total de cadeias no teste. Dessa maneira, temos a proporção de cadeias mínimas.

Além disso, nesse trabalho nós estudamos a performance de descritores léxicos para a classificação dos grupos, como:

número de palavras: Número total de palavras ditas no TFV.

número de repetições: Total de palavras repetidas.

tamanho da palavras: Descritores estatísticos do tamanho das palavras. Usamos a média, desvio padrão, curtose e obliquidade.

3.4 Similaridade entre palavras

Para identificar os switches, como mencionado na sessão anterior, usamos modelos computacionais para descobrir o nível de similaridade de duas palavras. A primeira abordagem é a de semântica distribucional, codifica distribuição de palavras em vetores, e a segunda, é baseada em ontologia, que são bases de conhecimento criadas por especialistas.

3.4.1 GloVe

Junto do Word2Vec (MIKOLOV et al., 2013), o GloVe (PENNINGTON; SOCHER; MANNING, 2014) é um dos modelos distribucionais mais usados na área de processamento de linguagem natural. Resumidamente, o algoritmo "codifica" as informações de coocorrências de palavras em vetor de dimensão dada pelo usuário. O GloVe faz isso minimizando uma função de custos que faz a similaridade dos vetores serem proporcionais a log-coocorrência deles no corpus de treinamento.

Inicialmente, as coocorrências das palavras são contadas em janelas de tamanho t, que é um parâmetro do modelo. Essas informações são guardadas em uma matriz X, na qual X_{ij} corresponde a coocorrência da palavra i com o contexto j. Pela hipótese distribucional, palavras relacionadas devem ter probabilidade de coocorrência alta. A probabilidade $P(\text{gelo}|\text{sólido})$ deve ser mais alta que $P(\text{gelo}|\text{moda})$. Se temos $P(\text{gelo}|k)$ e $P(\text{vapor}|k)$, se $k =$ água, então ambas probabilidades vão ser altas, porém se k for uma propriedade relacionada a apenas uma palavra do par, como $k =$ sólido, apenas gelo terá probabilidade alta. O GloVe explora razões de probabilidades de palavras $\frac{P(i|k)}{P(j|k)}$. Quando a razão for pequena, elas são relacionadas, quando for grande, elas não são relacionadas.

Pennington, Socher e Manning (2014) através de várias manipulações algébricas, partindo dessa premissa da razão da probabilidade das palavras e relacionando com as representações de palavras que queremos ter, w_i e w_j, chega na equação

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) \tag{3.1}$$

na qual k é a palavra de "inspeção".

O modelo minimiza uma função $J(\theta)$, sendo θ os parâmetros do modelo.

$$J(\theta) = \frac{1}{2} \sum_{ij}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ij}))^2 \tag{3.2}$$

O vocabulário possui tamanho V. A função $f(\cdot)$, é uma função que vai ponderar as frequências de coocorrências. Ela "corta" pares de palavras com frequência muito alta. No trabalho original é definida uma possibilidade de $f(\cdot)$.

No nosso trabalho, treinamos o GloVe com janela de tamanho 7 e todos os outros

parâmetros padrão do pacote disponibilizado pelos autores. O corpus de treinamento foi a Wikipedia em português (*dump* junho de 2016) lematizado com o parser Palavras (BICK, 2000). O corpus possui 44.345 tipos e $118.095.367$ tokens. Stopwords foram removidas.

3.4.2 WordNet

A WordNet (FELLBAUM, 1998) é uma base de dados que possui informação sobre relações entre as palavras. É um modelo que se propõem a representar a organização léxica da linguagem. Sua construção foi manual e é fruto do trabalho de vários linguistas através das décadas. Ela é organizada através de *synsets*, que são as entendidades básicas do modelo. Por exemplo, nas sentenças "um banco no meio do rio" e "o banco roubou meu dinheiro", a palavra "banco" está relacionada a dois synsets. O primeiro synset é "banco de areia" e o segundo é o "banco" como "instituição financeira". Os synsets estão organizados em uma rede semântica na qual temos relações como hiperonímia, hiponímia, sinonímia, entre outras.

No nosso trabalho, estamos interessados em relações de hiperonímia, pois elas induzem uma taxonomia. Esse tipo de relação pode ser traduzida como "é-um". Por exemplo, um cachorro *é-um* canídeo, um gato *é-um* felino, um canídeo *é-um* mamífero, um felino *é-um* mamífero. Navegando na hierarquia da rede semântica, podemos descobrir a relação entre dois animais. Para termos uma noção numérica da similaridade, utilizamos a *path-similarity* que é baseada no menor caminho entre dois synsets.

Como a WordNet é uma base de dados em inglês, precisamos utilizar versões traduzidas dos nomes dos animais. Contudo, como animais são facilmente traduzíveis, não tivemos grandes problemas em achar correspondentes em inglês para as palavras em português.

3.5 Algoritmos de aprendizado supervisionado

Nessa sessão, falaremos brevemente dos algoritmos de aprendizado de máquina que usamos. Informações mais completas podem ser encontradas em Hastie, Tibshirani e Friedman (2009).

3.5.1 Naive Bayes

Esse método é um modelo probabilístico de classificação. Seu nome vem do fato de usar o teorema de Bayes para estimar as probabilidades posteriores assumindo, ingenuamente (naive), que as probabilidades dos descritores são idependentes.

O método Naive Bayes estima a probabilidade condicional $P(C = k|X = \boldsymbol{x})$ sendo $k \in C$ uma classe do conjunto de classes C e $\boldsymbol{x} = (x_1, \ldots, x_N) \in X$ o conjunto de valores que os atributos podem tomar. Vamos simplificar a notação removendo C e X, que ficam implícitos. Para encontrar essa distribuição é utilizada a regra de bayes:

$$P(k|\boldsymbol{x}) = \frac{P(k)P(\boldsymbol{x}|k)}{P(\boldsymbol{x})} \tag{3.3}$$

Dessa maneira, o esforço computacional é direcionado para encontrar a distribuição $P(\boldsymbol{x}|k)$. O problema é facilitado pela suposição de independência entre os atributos, isto é, $P(\boldsymbol{x}|k) = \prod_{i=1}^{N} P(x_i|k)$. Com isso, temos que estimar apenas os $P(x_i|k)$, que por sua vez, dependem da ocorrência de x_i em k. Como $P(X = \boldsymbol{x})$ é apenas um termo de normalização e não influencia no resultado de comparações, ele pode ser descartado. Logo, o resultado da classificação é dada por:

$$\arg\max_{k} \; P(k) \prod_{i=1}^{N} P(x_i|k) \tag{3.4}$$

Isso quer dizer que escolhemos o k (classe), que tem a maior probabilidade.

Aqui, nesse trabalho, temos atributos contínuos, como similaridade e número de switches, logo, utilizamos o algoritmo Naive Bayes com estimativa de densidade. É possível estimar $P(x_i|k)$ apenas discretizando os valores contínuos, porém um abordagem mais interessante é assumir os dados como normalmente distribuídos. Para isso nós calculamos os valores de x_i associados a classe k, isto é, tomamos a média μ_k e a variância σ_k dos valores de x_i quando eles "acontecem com k". Esses valores serão parâmetros de uma distribuição normal. Logo, a probabilidade de x_i assumir um valor contínuo v é dada por:

$$P(x_i = v|k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(v-\mu_k)^2}{2\sigma_k^2}) \tag{3.5}$$

3.5.2 Regressão Logística

Apesar do nome, a regressão logística é uma abordagem de classificação que usa a função logística para estimar a probabilidade das classes. Essa função é interessante pois tem a propriedade de ser limitada entre 0 e 1, além de ser contínua e derivável.

Na modelagem assumimos apenas duas classes, k e $\neg k$, adicionalmente, dizemos que $k = 1$ e $\neg k = 0$. O objetivo é descobrir a distribuição $P(k|\boldsymbol{x}, \theta)$, sendo $\boldsymbol{x}$ os descritores e θ os parâmetros do modelo. Definimos que:

$$h_\theta(\boldsymbol{x}) = \frac{1}{1 + e^{-\theta^T \cdot \boldsymbol{x}}} \tag{3.6}$$

Isto é, temos um modelo h para $\boldsymbol{x}$ parametrizado por θ. A probabilidade de k é dada por:

$$P(k|\boldsymbol{x}, \theta) = h_\theta(\boldsymbol{x})^k (1 - h_\theta(\boldsymbol{x}))^{(1-k)} \tag{3.7}$$

O próximo passo é estimar os valores de θ, que podemos pensar como pesos de $\boldsymbol{x}$. Isso é feito através de uma função de verossimilhança $\mathcal{L}(\theta|D)$. Essa função indica o quanto o modelo, que depende de θ, está alinhado com o conjunto de dados de treinamento D. O conjunto de dados D, pode ser visto como um conjunto de tuplas de label e atributo $\{(k_1, \boldsymbol{x}_1), (k_2, \boldsymbol{x}_2), \ldots, (k_n, \boldsymbol{x}_n)\}$. Queremos achar os valores de $\theta \in \Theta$ que mais explicam os dados, isto é, queremos o $\max_{\theta \in \Theta}\ \mathcal{L}(\theta|D)$. Na função de verossimilhança, assumimos que as probabilidades dos modelos são independentes, logo,

$$\mathcal{L}(\theta|D) = \prod_{(k_i, x_i) \in D} P(k_i|x_i, \theta) \tag{3.8}$$

$$\log \mathcal{L}(\theta|D) = \sum_{(k_i, x_i) \in D} P(k_i|x_i, \theta) \tag{3.9}$$

Por causa de problemas de estabilidade numérica, normalmente é procurado o $\min_{\theta \in \Theta} - \log \mathcal{L}(\theta|D)$. Para isso são empregadas técnicas de otimização como a descida do gradiente.

3.5.3 Random Forest

A técnica de random forest (BREIMAN, 2001), consiste na criação de várias árvores de decisão treinadas em diversas amostras do conjunto de dados. Essas amostragens são feitas com reposição, numa técnica chamada boostraping. As decisões do modelo são tomadas a partir do conjunto das árvores, com o objetivo de ter predições com menor variância. O algoritmo de random forest, no seu processo de treinamento, nos indica seleção de atributos naturalmente, o que o torna um tipo de abordagem muito fácil de se usar.

Na tarefa de classificação, essas árvores predizem as classes baseando-se em avaliações dos descritores. As árvores são sempre binárias e cada nova sub-árvore (bifurcação do caminho) corresponde a um teste $x_i < \lambda$. As folhas são as decisões da árvore. Para decidir se um item $\boldsymbol{x} = (x_1, \ldots, x_N)$ pertence à classe $k \in C$, caminhamos na árvore testando os atributos, logo, as folhas são estimativas de $P(k|\boldsymbol{x})$

O treinamento dessas árvores consiste em achar a topologia de árvore com testes que melhor separa os dados. Cada bifurcação tem como objetivo particionar os dados de maneira que tenhamos partições o mais "puras" possíveis. Um particionamento com pureza máxima separa os dados perfeitamente.

As árvores de decisão possuem alguns problemas bem conhecidos. O principal deles é o *overfitting*, isto é, é muito fácil chegar a modelos que não possuem poder de generalização. Uma estratégia para escapar desse problema é treinar várias árvores de decisão aleatórias e tomar a média de seus resultados. O random forest se baseia nessa ideia.

O algoritmo random forest se propõe a treinar B árvores de decisão aleatórias. Para construir várias árvores aleatoriamente, cada árvore vai ser treinada com um subconjunto dos dados. Esses dados são selecionados com distribuição uniforme e com reposição (*boostrap*). Cada árvore vai ser construída até uma altura máxima, que é um parâmetro. Selecionamos t atributos aleatórios do conjunto de descritores. Depois, encontramos as melhores bifurcações para esse conjunto de t atributos. Cada árvore b terá uma estimativa de probabilidade $P_b(k|\boldsymbol{x})$. A resposta do modelo será $P_b(k|\boldsymbol{x}) = \frac{1}{B} \sum_b^B P_b(k|\boldsymbol{x})$

3.5.4 Métricas para avaliação

Para classificadores que têm saída a probabilidade da decisão, podemos plotar a curva ROC (*receiver operator characteristic*). Essa curva tem como eixos a taxa de falsos positivos e a taxa de falsos negativos. Quando temos um problema de classificação binária, podemos definir o limiar de probabilidade para separar as classes como 0.5, porém não precisa ser assim. Por exemplo, quando temos um problema cujos os falsos negativos são mais graves que os falsos positivos, podemos aumentar o limiar de probabilidade de maneira que seja mais difícil a classe negativa. Um ponto na curva ROC corresponde ao limiar de probabilidade para determinada taxa de falsos positivos e taxa de falsos negativos. Com isso, temos informação bastante geral sobre o poder de separação do classificador.

Para resumir a informação da curva ROC, nós calculamos a área sob a curva, chamada de AUC (*Area under ROC curve*). Uma AUC de 0.5 corresponde ao classificador aleatório e uma AUC de 1 a classificação perfeita.

Outra medida que usamos é a medida-f, também chamada de F_1. Essa medida é uma média harmônica entre a precisão e a sensibilidade.

3.6 Avaliação da Identificação de Switches

Como os dados foram identificados usando uma taxonomia e nossa abordagem de identificação encontra cadeias semânticas, precisamos de recursos extras para avaliarmos nossos algoritmos de detecção. Além disso, outro fato que influencia a coleta de novos dados é que taxonomias e identificação dos switches não estão imunes a erros. A taxonomia mais conhecida (TROYER; MOSCOVITCH; WINOCUR, 1997), possui algumas peculiaridades, como tigre pertencer a animais da África. Outro motivo, é que essa lista de animais teve que ser adaptada para português e para o Brasil. Isso pode trazer alguns problemas, como, por exemplo, o animal mais próximo de onça na taxonomia, é *panther* (pantera) que é um animal da África, logo, no conjunto de dados, onça compartilha a categoria de zebra, enquanto poderia estar em um grupo mais interessante, como “Animais da Amazônia"ou “Animais do Brasil” (que não existem na taxonomia original).

Validaremos as nossas abordagens computacionais de detecção de switches através de experimentos com anotadores. No experimento que propomos, coletamos um conjunto de 594 julgamentos de quebra semântica que os modelos distribucionais têm dificuldade em prever. Com isso, teremos um conjunto de dados com uma avaliação da taxonomia

e avaliação dos anotadores. Calculamos a concordância entre os julgamentos para entender melhor a qualidade da detecção. As abordagens computacionais, têm seu poder de predição avaliado contra o conjunto de dados que tanto a taxonomia e os anotadores concordam. Além disso, também fazemos uma avaliação qualitativa dos resultados.

3.6.1 Coleta de dados

Perguntamos para um grupo de pessoas se em uma sequência de quatro palavras existe uma quebra semântica ou não. Esse experimento foi conduzido na forma de um questionário online no qual cada pessoa recebia um conjunto de 90 quadrigramas. Dessa maneira, conseguimos uma grande cobertura de avaliação de dados. O questionário foi aplicado em um conjunto de alunos da disciplina de Linguagens Formais e Autômatos de 2017 do Instituto de Informática, UFRGS.

A ideia do experimento é ter uma avaliação de switches alternativa à taxonomia. Para isso, extraímos conjuntos de 4 tokens dos testes de fluência verbal, de maneira que podemos comparar com os switches da avaliação original dos dados. Essas sequências foram extraídas dos TFV em uma janela deslizante de quatro posições. Tokens foram inseridos nas duas extremidades dos testes para conseguirmos extrair os switches das bordas (padding). Dessa maneira, obtemos um total de 1162 quadrigramas, dos quais, avaliamos 594. A partir do conjunto de quadrigramas, criamos os formulários automaticamente. Usamos Apps Script, a linguagem de scripts/macros de produtos Google, para transformar uma tabela em 60 formulários. Cada formulário contém 90 avalições de quadrigramas e eles foram escolhidos em uma política de subgrupos com sobreposição.

Inicialmente apresentamos um texto em que indicamos o objetivo do questionário e pedimos o consentimento do participante. Em seguida, o questionário pergunta dados demográficos, como idade, gênero e escolaridade. Na última parte, os avaliadores julgam as sequências de palavras.

3.7 Avaliação de identificação de sujeitos clínicos

Queremos entender o poder de classificação dos descritores baseados em cadeias semânticas. Para avaliar isso, entendemos que os descritores devem apresentar uma performance consistente em diferentes algoritmos de classificação. Estamos interessados

também em saber se esses descritores nos ajudam a superar o estado da arte, que é baseado em medidas de grafo. Como o estado da arte não usa informação semântica, os descritores aqui avaliados devem adicionar poder de classificação.

Comparamos a performance de classificadores baseados em diversas abordagens de detecção, inclusive aleatória, para ter uma ideia concreta da performance dos descritores.

4 EXPERIMENTOS

Nessa capítulo falaremos sobre os resultados dos experimentos que executamos. Na Seção 4.1 reportamos o resultado da coleta de dados com anotadores. Falaremos, na Seção 4.2, dos resultados da detecção computacional das quebras semânticas nos testes. Na Seção 4.3, mostramos a performance das combinações de modelos e descritores na tarefa de identificação de casos clínicos.

4.1 Resposta aos questionários

Recebemos respostas de 58 avaliadores. A idade média é de 21 ($\sigma = 2.6$) anos. Temos 10 mulheres, 48 homens e todos participantes possuem ensino superior incompleto. Todos os quadrigramas foram avaliados por pelo menos 3 sujeitos. Em média, cada quadrigrama foi julgado por 8.1 ($\sigma = 2.28$) sujeitos.

Na figura 4.1, podemos ver as distribuições das respostas. O primeiro painel, Figura a, mostra a distribuição de número de avaliadores por quadrigrama. No segundo, Figura b, vemos a distribuição proporção de *sim*, isto é, para um dado quadrigrama, o número de *sim*s dividido pelo total de *sim*s e *não*s. Podemos ver na figura que a maior parte das proporções ficam nas extremidades, isso indica um considerável nível de certeza por parte dos avaliadores.

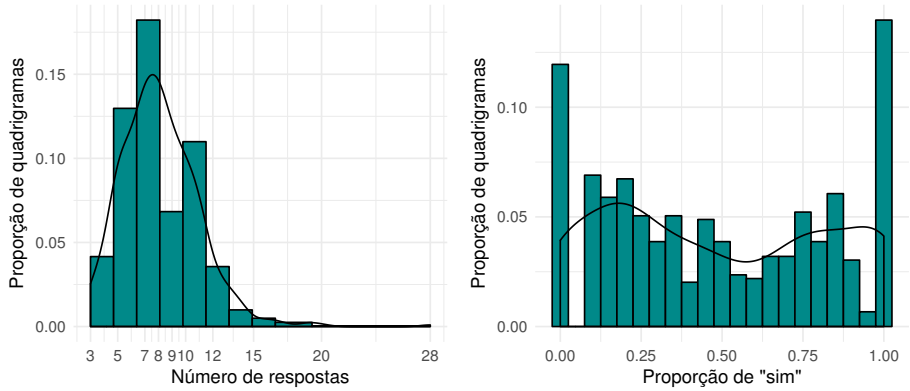
Com o resultado desse questionário, obtemos um conjunto de dados de referência para identificação de switches. Para os 594, fizemos a decisão baseada na maioria.

Apesar de não ser o objetivo desse experimento, podemos ver o quanto o contexto influencia a decisão da existência de switch no conjunto de palavras. Para isso, identificamos todos os pares de palavras que aparecem "no meio" dos quadrigramas e que repetem em contextos diferentes. Não consideramos a ordem, isto é, $w_1 w_2 = w_2 w_1$.

Existem 423 bigramas únicos, desses, 75 ocorrem na ordem reversa. Temos 116 casos de bigramas que aparecem em mais de um contexto. Desses 116, em apenas 11 existe pelo menos uma avaliação que não concorda com as outras. Como são poucos os casos, os listaremos na Tabela 4.1.

Podemos ver na tabela que muitos casos são realmente bastante ambíguos. Cachorro e porco, por exemplo, parecem serem relacionados se o contexto for de "animais de fazenda". No caso do quadrigrama "gato cachorro cavalo porco", o qual é considerado um switch, temos dois animais bastante característicos dos seus grupos, cachorro,

Figura 4.1: Distribuições das respostas. A figura (a) apresenta a distribuição do número de respostas por quadrigramas. Na figura (b) podemos ver a proporção de respostas *sim* em relação ao total de respostas para cada quadrigrama. O gráfico do *não* é uma versão espelhada da do *sim*.



(a) (b)

Tabela 4.1: Casos de pares de palavras em que houve avaliações conflitantes baseadas no contexto.

par	contexto	avaliaçao
Cachorro Porco	cavalo cachorro porco burro	s
	cavalo cachorro porco cavalo	s
	cavalo cachorro porco aranha	n
	elefante cachorro porco gato	n
	burro cachorro porco zebra	n
	cavalo porco cachorro galinha	s
	boi porco cachorro cavalo	s
	rinoceronte porco cachorro gato	n
Cachorro Cavalo	porco cachorro cavalo burro	s
	galinha cachorro cavalo cachorro	s
	cabrito cachorro cavalo burro	s
	# cachorro cavalo porco	s
	gato cachorro cavalo porco	n
	# cavalo cachorro porco	s
	cachorro cavalo cachorro jumento	n
Galinha Coelho	cachorro galinha coelho pato	s
	papagaio galinha coelho porquinho_da_índia	n
	elefante coelho galinha cobra	n
Macaco Veado	leão macaco veado tigre	s
	avestruz macaco veado cutia	n
Camelo Elefante	carneiro camelo elefante vaca	s
	cavalo camelo elefante zebra	n
Cobra Macaco	calango cobra macaco peixe	s
	galinha cobra macaco #	n
	porco macaco cobra leão	n
	# macaco cobra cachorro	n
Macaco Tigre	camelo macaco tigre onça	s
	tamanduá macaco tigre carneiro	n
	gato tigre macaco urso	s
	onça tigre macaco tartaruga	s
Onça Cobra	tigre onça cobra jacaré	s
	tigre onça cobra leão	n
	pássaro cobra onça leão	s
	leão cobra onça cachorro	s
	sapo cobra onça lagarta	s
Zebra Macaco	onça zebra macaco galinha	s
	girafa zebra macaco porco	n
Carneiro Veado	camelo carneiro veado ovelha	s
	gato carneiro veado #	n
	galinha veado carneiro cabrito	n
Gato Rato	cachorro gato rato cavalo	s
	cachorro gato rato onça	n

do grupo dos pets, e cavalo, do grupo dos animais de fazenda. A ligação de gato com cachorro e de cavalo com porco parece forçar o aparecimento de uma quebra semântica. As pessoas também parecem ter dúvidas sobre qual categoria colocar “macaco”, “cobra” e “veado”.

4.2 Avaliação de detecção computacional de switches

Vamos primeiro validar a detecção de switches usando os três pares de palavras do quadrigrama. Para isso, utilizamos a técnica baseada na média (Seção 3.2.1) e abordagens

Tabela 4.2: Performance dos modelos. **RF** corresponde ao modelo random forest, **NB** ao naive bayes, **SVM** ao modelo de máquinas de vetor de suporte e **Logística** ao modelo de regressão logística. A coluna **Média** representa a heurística de detecção baseada na média das palavras faladas no teste e **Random** o modelo aleatório.

Modelo		RF	NB	SVM	Logística	Media	Random
GloVe	F_1	**.63 (.07)**	.59 (.06)	.54 (.06)	.53 (.06)	.49 (.07)	.52 (.09)
	Precisão	**.63 (.12)**	.61 (.09)	.53 (.09)	.53 (.08)	.41 (.07)	.53 (.11)
	revocação	**.64 (.05)**	.57 (.07)	.56 (.06)	.53 (.05)	.42 (.08)	.51 (.08)
WordNet	F_1	.62 (.05)	**.65 (.10)**	.63 (.06)	.58 (.10)	.56 (.06)	.50 (.04)
	Precisão	.61 (.08)	**.72 (.14)**	.68 (.14)	.58 (.13)	.58 (.07)	.52 (.07)
	revocação	**.62 (.05)**	.60 (.08)	.60 (.07)	.60 (.09)	.55 (.05)	.49 (.03)
Glove + WordNet	F_1	**.74 (05)**	.67 (.05)	.64 (.09)	.61 (.05)	-	.47 (.05)
	Precisão	**.74 (.09)**	.70 (.08)	.64 (.12)	.65 (.07)	-	.47 (.06)
	revocação	**.75 (.04)**	.65 (.05)	.65 (.06)	.59 (.04)	-	.47 (.05)
Glove + WordNet - sem contexto	F_1	**.86 (.04)**	.69 (.04)	.70 (.06)	.60 (.05)	-	.48 (.05)
	Precisão	**.83 (.07)**	**.83 (.09)**	.82 (.09)	.60 (.06)	-	.50 (.07)
	revocação	**.90 (.03)**	.60 (.03)	.61 (.04)	.59 (.05)	-	.47 (.04)

usando aprendizado de máquina. Os algoritmos utilizados foram naive bayes (com estimação gaussiana de densidade), random forest, regressão logística e máquinas de vetor de suporte com kernel radial. Todos os modelos foram avaliados usando 10-fold-cross-validation (validação cruzada). Para melhor entender como os algoritmos se comportam, treinamos os modelos nos quadrigramas em que a taxonomia e os anotadores concordam. Os resultados podem ser vistos na Tabela 4.2

Podemos ver que o algoritmo random forest apresenta uma boa performance e que a informação do contexto confunde os algoritmos. O algoritmo baseado na média vai mal, porque esse conjunto de dados foi selecionado com base na pouca performance de modelos distribucionais. O melhor modelo é random forest, sem contexto e com 2 descritores: similaridade do par de palavras dado pelo GloVe e pela WordNet. Chamaremos, a partir daqui, esse melhor modelo de $\text{RF}_{\text{GloVe/WN}}$.

Nos 594 quadrigramas do experimento, $\text{RF}_{\text{GloVe/WN}}$ obteve $\kappa = 0.848$ com os anotadores e $\kappa = 0.891$ com as avaliações da taxonomia. A concordância entre a taxonomia e os anotadores é de $\kappa = 0.74$. No conjunto de dados dos 1126 quadrigramas, que consiste nos 594 que o algoritmo conhece e mais 532 avaliados apenas pela taxonomia, o modelo $\text{RF}_{\text{GloVe/WN}}$ obteve $\kappa = 0.6$ e $F_1 = 0.8$. A a abordagem baseada na média apresentou $\kappa = 0.3$ e $F_1 = 0.65$.

Desse experimento podemos dizer que a tarefa de indentificar quebras semânticas em sequências de palavras é bastante difícil, pois até mesmo anotadores apresentam uma concordância moderada com a taxonomia. Apesar desse fato, obtemos uma performance considerável e uma concordância razoável no conjunto de dados como um todo.

4.3 Classificação dos grupos clínicos

Outra maneira que temos como avaliar a detecção de switches é descobrindo se ela ajuda no diagnóstico de doenças. Para isso, fazemos uma análise do poder de classificação de cada descritor e de cada abordagem.

O nosso baseline é o número de palavras e o atual estado da arte é o conjunto de descritores baseados grafos (descritos na Seção 2.4). A Tabela 4.3 apresenta a performance desses classificadores, junto de outras combinações de descritores lexicais.

Tabela 4.3: Resultado da classificação dos baselines. *ALZ* corresponde a classificação de grupo controle versos grupo Alzheimer e assim por diante. As células contém o valor de AUC médio entre as 10 rodadas de 10-fold-cross-validation. O conjunto de descritores "lexicais" corresponde ao número de palavras, as estatísticas do tamanho das palavras e ao número de repetições.

	ALZ			aMCI			a+mdMCI		
	NB	RF	RL	NB	RF	RL	NB	RF	RL
num palavras	.88 (.17)	.76 (.23)	.87 (.17)	.59 (.28)	.53 (.27)	.63 (.25)	.67 (.26)	.58 (.24)	.69 (.26)
tamanho da palavra	.76 (.22)	.87 (.17)	.73 (.26)	.65 (.25)	.81 (.19)	.46 (.24)	.60 (.28)	.65 (.28)	.54 (.27)
num palavras + repetições	.86 (.19)	.81 (.23)	.88 (.19)	.71 (.23)	.63 (.24)	.71 (.24)	**.76 (.22)**	.62 (.26)	.75 (.23)
num palavras + tam palavra	.84 (.21)	.87 (.16)	.82 (.19)	.64 (.27)	.75 (.23)	.51 (.25)	.57 (.27)	.70 (.26)	.59 (.28)
lexicais	.85 (.16)	**.91 (.16)**	.73 (.26)	.68 (.24)	**.82 (.20)***	.62 (.28)	.62 (.24)	.71 (.24)	.67 (.26)
grafo	.88 (.16)	.83 (.19)	**.89 (.16)**	.70 (.26)	.63 (.23)	**.72 (.24)**	**.75 (.26)**	.64 (.26)	.65 (.27)

Fazemos os testes-t entre os grupos, isto é, melhor modelo para o grupo clínico x contra o segundo melhor modelo do grupo clínico x. Para o Alzheimer, os descritores lexicais junto do random forest não bateram as features de grafo com regressão logística ($t(197.69) = 1.25$, $p = 0.10$), porém essa configuração é superior para identificação do grupo aMCI ($t(192.04) = 3.33$, $p = 0.0005$). No grupo de a+mdMCI, o naive bayes com número de palavras junto do número de repetições não vence as features de grafo com naive bayes ($t(193.12) = 0.46$, $p = 0.32$).

Podemos ver na Tabela 4.3 que número de palavras para separar o grupo controle de Alzheimer, é muito próximo do estado da arte. Por outro lado, para separar os grupos controle e aMCI ele não possui o mesmo poder. No caso do aMCI, as estatísticas de tamanho da palavra parecem ajudar mais os modelos e, no caso do a+mdMCI, é o número de repetições. Outro fato interessante é que, nos três grupos, o conjunto de descritores "número de palavras + repetições" possui uma performance que chega perto do estado da arte.

Com esses resultados, podemos partir para a análise da performance dos descritores baseados em cadeias semânticas. Para cada detector de switches, temos 12 conjuntos de descritores que são referentes as combinações de número de switches, cadeia média, proporção de cadeia mínima e cadeia máxima. Todos os modelos foram treinados com

naive bayes, random forest e regressão logística. No total, temos 108 modelos para separar cada grupo. Na Tabela 4.4, cada linha representa o melhor modelo entre 108 para cada grupo. Apesar de que conseguimos $p < 0.05$ para o grupo aMCI, visto o número de comparações que fazemos para chegar na tabela, podemos concluir que os descritores de cadeias semânticas por si só não possuem mais poder de classificação que o número de palavras.

Tabela 4.4: Melhor modelo que usa apenas descritores baseados em informações de cadeia semântica. A coluna "Número de palavras" corresponde ao melhor modelo, para cada grupo, da primeira linha da Tabela 4.3. Na coluna "Configuração", média-WordNet e média-GloVe corresponde a detecção heurística baseada na média de similaridade. Switches corresponde ao número de switches e cadeia aos descritores de cadeia.

	Configuração	AUC	Número de palavras	t	p
Alzheimer	Média-WordNet switches + cadeia min + random forest	.89	.88	0.36	.35
aMCI	Média-GloVe switches + cadeia max + naive bayes	.64	.63	0.52	0.3
a+mdMCI	Média-WordNet switches + cadeia min + regressão logística	.76	.69	1.84	0.03

Na Tabela 4.5, combinamos os descritores da Tabela 4.3 com os descritores baseados em cadeia semântica. Comparamos, para cada grupo, com o melhor modelo da Tabela 4.3. O objetivo dessa análise é descobrir se esses descritores adicionam alguma informação nos melhores modelos. Olhando os resultados, não podemos dizer que esses descritores aumentam a performance do estado da arte.

Tabela 4.5: Melhor combinação de descritores de cadeia com os melhores modelos da Tabela 4.3. Na coluna "Estado da Arte", o modelo de referência para Alzheimer corresponde aos descritores lexicais com random forest, para aMCI, descritores lexicais com random forest e, para a+mdMCI, número de palavras juntamente de número de repetições e regressão logística. Na coluna "configuração", Lex corresponde a descritores lexicais, média-GloVe e média-Wordnet a detecção heurística pela média usando esses modelos e cadeia os descritores de cadeia.

	Configuração	AUC	Estado da Arte	t	p
Alzheimer	Lex + média-GloVe cadeia max + random forest	.92	.91	0.63	0.26
aMCI	Lex + média-GloVe cadeia média + random forest	.83	.82	-0.40	0.65
a+mdMCI	N plvras + rep + média-WordNet + cadeia média + min + regr logística	.74	.76	1.84	0.74

Os resultados apresentados nessa seção, nos faz questionar a validade da hipótese de que as informações de cadeias semânticas podem servir para a ajudar o diagnóstico de doenças. Descritores como número de palavras e número de repetições, são de fácil entendimento e não dependem de um aparato computacional sofisticado, como transformar os TFV em grafos ou usar técnicas de processamento de linguagem natural.

5 CONCLUSÃO

Doenças neuropsicológicas podem ser devastadoras e representam um risco crescente na nossa sociedade. Muitos pesquisadores têm utilizado técnicas de processamento de linguagem natural para criar novos métodos que auxiliem a detecção. O estudo computacional de testes de fluência verbal ganhou atenção no últimos anos. Em especial, a comunidade de processamento de linguagem natural procura métodos computacionais para analisar semanticamente os testes de fluência verbal, contudo, muitos desses trabalhos não são suficientemente validados. Essa pesquisa, como um todo, sugere maneiras de validar o estudo computacional de cadeias semânticas nos testes de fluência verbal. Além disso, o experimento de coleta com anotadores, indica ideia da dificuldade inerente do problema de decidir se palavras são relacionadas ou não.

Mostramos que a identificação de switches usando a similaridade do GloVe e WordNet, juntamente do random forest, obtém boa performance nos dados que coletamos. Avaliamos, comparativamente, o poder de classificação de descritores lexicais, descritores baseados em grafo e descritores baseados em cadeias semânticas. Adicionalmente, também mostramos que para o nosso conjunto de dados, os descritores baseados em cadeias semânticas não são melhores que os outros conjuntos de descritores.

Para a identificação de switches, a combinação de GloVe com WordNet se mostra promissora. Podemos argumentar que, como a WordNet possui conhecimento biológico dos animais, por exemplo, que gato e leão são felinos, e o GloVe conhecimento baseado em contextos, como gato e rato são inimigos, essa combinação é complementar.

5.1 Trabalhos futuros

Como continuidade do trabalho observamos a necessidade de replicar esse estudo em diferentes populações clínicas, principalmente, com mais dados. Além disso, temos que investigar a relação desses descritores semânticos com outros descritores de tarefas neuropsicológicas relacionadas. Um exemplo de tarefa interessante é o *Roubo do Pote de Biscoito* (GILES; PATTERSON; HODGES, 1996). Nessa tarefa, é mostrada uma ilustração na qual duas crianças roubam um pote de biscoito e pessoa que está sendo avaliada deve descrever a figura. Um estudo computacional pode ser feito procurando características da fala de casos clínicos e comparar com a performance dos descritores semânticos.

Outra variável que não temos no momento é o aspecto temporal do teste, isto é, o

instante de tempo em que cada palavra é emitida pelo sujeito. Os switches são marcados por uma pausa maior que o comum na sequência de palavras faladas. A identificação de cadeias semânticas e os tempos poderiam ser cruzados.

Com esses dados, poderemos partir para um entendimento mais profundo de como se indica a perda categorias semânticas na memória. Compreender melhor essas técnicas pode nos ajudar a criar ferramentas que ajudam o diagnóstico de doenças e, consequentemente, diminuir o sofrimento de muitas pessoas.

REFERÊNCIAS

ABBOTT, J. T.; AUSTERWEIL, J. L.; GRIFFITHS, T. L. Random walks on semantic networks can resemble optimal foraging. **Psychological Review**, American Psychological Association (APA), v. 122, n. 3, p. 558–569, 2015. Disponível em: <https://doi.org/10.1037%2Fa0038693>.

BERTOLA, L. et al. Impaired generation of new subcategories and switching in a semantic verbal fluency test in older adults with mild cognitive impairment. **Frontiers in Aging Neuroscience**, Frontiers Media SA, v. 6, jul 2014. Disponível em: <https://doi.org/10.3389/fnagi.2014.00141>.

BERTOLA, L. et al. Graph analysis of verbal fluency test discriminate between patients with alzheimers disease, mild cognitive impairment and normal elderly controls. **Frontiers in Aging Neuroscience**, Frontiers Media SA, v. 6, jul 2014. Disponível em: <https://doi.org/10.3389%2Ffnagi.2014.00185>.

BICK, E. **The Parsing System 'Palavras': Automatic Grammatical Analysis**. [S.l.]: Aarhus University Press, 2000. ISBN 8772889101.

BORGE-HOLTHOEFER, J.; MORENO, Y.; ARENAS, A. Modeling abnormal priming in alzheimer's patients with a free association network. **PLoS ONE**, Public Library of Science (PLoS), v. 6, n. 8, p. e22651, aug 2011. Disponível em: <https://doi.org/10.1371/journal.pone.0022651>.

BOUSFIELD, W. A.; SEDGEWICK, C. H. W. An analysis of sequences of restricted associative responses. **The Journal of General Psychology**, v. 30, n. 2, p. 149–165, 1944.

BRAMBATI, S. M. et al. Single- and multiple-domain amnestic mild cognitive impairment: Two sides of the same coin? **Dementia and Geriatric Cognitive Disorders**, S. Karger AG, v. 28, n. 6, p. 541–549, 2009. Disponível em: <https://doi.org/10.1159/000255240>.

BREIMAN, L. Random forests. **Machine Learning**, Springer Nature, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/a:1010933404324>.

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [S.l.]: A Bradford Book, 1998. ISBN 026206197X.

GILES, E.; PATTERSON, K.; HODGES, J. R. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimers type: Missing information. **Aphasiology**, Informa UK Limited, v. 10, n. 4, p. 395–408, may 1996. Disponível em: <https://doi.org/10.1080/02687039608248419>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)**. [S.l.]: Springer, 2009. ISBN 9780387848587.

HILLS, T. T.; JONES, M. N.; TODD, P. M. Optimal foraging in semantic memory. **Psychological Review**, American Psychological Association (APA), v. 119, n. 2, p. 431–440, 2012. Disponível em: <https://doi.org/10.1037%2Fa0027373>.

LINZ, N. et al. **Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task**. 2017. Disponível em: <http://www.aclweb.org/anthology/W17-6926>.

MCKHANN, G. et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. **Neurology**, v. 34, n. 7, p. 939–944, Jul 1984.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems**. USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Disponível em: <http://dl.acm.org/citation.cfm?id=2999792.2999959>.

PAKHOMOV, S. V.; HEMMY, L. S. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. **Cortex**, Elsevier BV, v. 55, p. 97–106, jun 2014. Disponível em: <https://doi.org/10.1016/j.cortex.2013.05.009>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics (ACL), 2014. Disponível em: <https://doi.org/10.3115%2Fv1%2Fd14-1162>.

PRUD'HOMMEAUX, E.; SANTEN, J. van; GLINER, D. Vector space models for evaluating semantic fluency in autism. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 32–37. Disponível em: <http://aclweb.org/anthology/P17-2006>.

SHAO, Z. et al. What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. **Frontiers in Psychology**, Frontiers Media SA, v. 5, jul 2014. Disponível em: <https://doi.org/10.3389%2Ffpsyg.2014.00772>.

TROYER, A. K.; MOSCOVITCH, M.; WINOCUR, G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. **Neuropsychology**, American Psychological Association (APA), v. 11, n. 1, p. 138–146, 1997. Disponível em: <https://doi.org/10.1037/0894-4105.11.1.138>.

TROYER, A. K. et al. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. **Neuropsychologia**, Elsevier BV, v. 36, n. 6, p. 499–504, jun 1998. Disponível em: <https://doi.org/10.1016%2Fs0028-3932%2897%2900152-8>.

WINBLAD, B. et al. Mild cognitive impairment - beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. **Journal of Internal Medicine**, Wiley-Blackwell, v. 256, n. 3, p. 240–246, sep 2004. Disponível em: <https://doi.org/10.1111/j.1365-2796.2004.01380.x>.

APÊNDICE A — TABELAS ADICIONAIS

A.1 Concordância entre julgamentos

Fazemos um estudo de concordância na interssecção entre quadrigramas que coletamos no experimento e os que foram avaliados pela taxonomia. Obtivemos um $\kappa = 0.74$ (Cohen). Isso quer dizer que tivemos uma boa concordância entre as duas avaliações. De 594 casos, apenas 74 apresentam conflito, isto é, a taxonomia e os anotadores discordaram. A Tabela A.1 lista todos os casos. Adicionalmente, colocamos o julgamento do nosso melhor modelos computacional.

A.2 Resultado detalhado dos descritores de cadeia

A Tabela A.2 mostra todos os resultados de uma sequência de 10 vezes 10-fold-cross-valitation de treinamento dos modelos. Essa rodada de treinamento é diferente da reportada no trabalho. Isso ilustra o quanto os resultados variam. A variância de todos os modelos está entre 0.17 e 0.32.

A.3 Performance dos descritores baseados em detecção aleatória

Outra maneira de entender o quanto a detecção de switches ajuda no problema de identificação de casos clínicos, é estudar a performance dos descritores baseados na detecção aleatória. O detector aleatório de switches diz que um par de palavras consecutivas é switch com probabilidade de 0.5. Treinamos os modelos com 10-fold-cross-validation repetida 30 vezes. A cada repetição, aplicamos novamente a detecção de switches. A Tabela A.3 mostra o resultado.

Podemos ver que o resultado da A.3 e A.2 muitas vezes é parecido. Isso pode vir do fato de que as medidas de switch são correlacionadas com o número de palavras. Para o detector aleatório, a correlação do número de palavras com o número de switches é 0.79, com o tamanho médio da cadeia 0.33, com o tamanho máximo da cadeia 0.49 e com a proporção da cadeia mínima -0.12 (todos os valores com $p < 0.001$).

Tabela A.1: Todos os casos em que o taxonomia e os anotadores não concordam. A coluna $RF_{GloVe/WN}$ corresponde aos julgamentos do melhor modelo de detecção computacional. Nos julgamentos **s** indica que os animais são relacionados e **n** indica que não são relacionados.

Quadrigrama				Taxonomia	Anotadores	$RF_{GloVe/WN}$
#	**camelo**	**zebra**	cachorro	s	n	n
cachorro	**gato**	**coelho**	cavalo	s	n	n
cachorro	**gato**	**galinha**	porco	s	n	n
cachorro	**gato**	**maritaca**	boi	s	n	s
cachorro	**onça**	**urso**	#	s	n	s
capivara	**paca**	**quati**	macaco	s	n	n
cavalo	**girafa**	**veado**	cobra	s	n	s
cobra	**aranha**	**mosquito**	pernilongo	s	n	s
cobra	**peixe**	**jacaré**	#	s	n	n
coelha	**jacaré**	**anta**	lontra	s	n	n
elefante	**zebra**	**camelo**	girafa	s	n	n
galo	**camelo**	**macaco**	tigre	s	n	n
gato	**camelo**	**zebra**	#	s	n	n
gato	**jacaré**	**peixe**	vaca	s	n	n
gato	**papagaio**	**coelho**	cobra	s	n	n
girafa	**zebra**	**macaco**	porco	s	n	s
jacaré	**cobra**	**jabuti**	tartaruga	s	n	n
jegue	**camelo**	**macaco**	#	s	n	n
leão	**onça**	**camelo**	andorinha	s	n	n
leão	**tigre**	**camelo**	elefante	s	n	n
leão	**veado**	**tigre**	#	s	n	n
macaco	**urso**	**leão**	veado	s	n	s
macaco	**veado**	**tigre**	rinoceronte	s	n	n
onça	**égua**	**galinha**	porco	s	n	s
onça	**veado**	**leão**	elefante	s	n	s
tamanduá	**macaco**	**tigre**	carneiro	s	n	s
tatu	**calango**	**cobra**	macaco	s	n	s
tigre	**macaco**	**urso**	leão	s	n	s
#	**cachorro**	**cavalo**	porco	n	s	n
boi	**cavalo**	**camelo**	elefante	n	s	n
boi	**galinha**	**pato**	pássaro	n	s	s
boi	**porco**	**cachorro**	cavalo	n	s	n
burro	**galinha**	**pássaro**	gato	n	s	s
burro	**jumento**	**zebra**	elefante	n	s	s
cabra	**carneiro**	**peru**	galinha	n	s	n
cabrito	**cachorro**	**cavalo**	burro	n	s	n
cachorro	**cabra**	**galinha**	pato	n	s	n
cachorro	**galinha**	**coelho**	pato	n	s	n
cachorro	**gato**	**pássaro**	gavião	n	s	s
cachorro	**gato**	**rato**	cavalo	n	s	n
cachorro	**gato**	**tigre**	macaco	n	s	n
cachorro	**vaca**	**galinha**	pato	n	s	s
calango	**cobra**	**macaco**	peixe	n	s	n
camelo	**carneiro**	**veado**	ovelha	n	s	n
carneiro	**bode**	**galo**	papagaio	n	s	s
carneiro	**camelo**	**elefante**	vaca	n	s	n
cavalo	**cachorro**	**porco**	burro	n	s	n
cavalo	**cachorro**	**porco**	cavalo	n	s	n
cavalo	**jumento**	**zebra**	girafa	n	s	s
cavalo	**porco**	**cachorro**	galinha	n	s	n
coelho	**gato**	**onça**	zebra	n	s	s
coelho	**gato**	**tigre**	elefante	n	s	n
elefante	**girafa**	**avestruz**	coelho	n	s	n
elefante	**onça**	**jacaré**	#	n	s	s
galinha	**cachorro**	**cavalo**	cachorro	n	s	n
galinha	**galo**	**carneiro**	cabrito	n	s	n
galinha	**porco**	**pato**	marreco	n	s	s
gato	**tigre**	**elefante**	cobra	n	s	s
jacaré	**crocodilo**	**tubarão**	baleia	n	s	s
leão	**cobra**	**onça**	cachorro	n	s	n
macaco	**mico**	**tamanduá**	#	n	s	n
macaco	**porco**	**peru**	tigre	n	s	n
onça	**hipopótamo**	**capivara**	tatu	n	s	n
onça	**leão**	**macaco**	jacaré	n	s	s
paca	**quati**	**macaco**	mono	n	s	n
paca	**tatu**	**cutia**	veado	n	s	n
pássaro	**cobra**	**onça**	leão	n	s	n
pássaro	**jacaré**	**onça**	porco	n	s	s
porco	**cachorro**	**cavalo**	burro	n	s	n
quati	**macaco**	**mono**	guaxinim	n	s	s
sapo	**cobra**	**onça**	lagarta	n	s	n
tatu	**gambá**	**coelho**	veado	n	s	n
tigre	**onça**	**cobra**	jacaré	n	s	n
vaca	**burro**	**camelo**	ovelha	n	s	s
vaca	**cavalo**	**camelo**	bezerro	n	s	n

Tabela A.2: Performance de todos os descritores, combinações e detectores de switches. Células em cinza apresentam $AUC < 0.6$. A variância dos valores fica entre 0.17 e 0.32.

		ALZ			aMCI			a+mdMCI		
		NB	RF	RL	NB	RF	RL	NB	RF	RL
Taxonomia	switches	0.788	0.771	0.795	0.581	0.646	0.617	0.770	0.708	0.712
	cadeia média	0.469	0.639	0.484	0.623	0.640	0.502	0.580	0.723	0.586
	cadeia min	0.474	0.669	0.493	0.616	0.650	0.501	0.571	0.731	0.584
	cadeia max	0.596	0.634	0.637	0.798	0.779	0.526	0.591	0.461	0.469
	sw + cadeia media	0.755	0.769	0.806	0.626	0.584	0.499	0.725	0.762	0.667
	sw + cadeia min	0.756	0.779	0.806	0.626	0.593	0.497	0.721	0.761	0.668
	sw + cadeia max	0.816	0.747	0.824	0.728	0.635	0.527	0.715	0.644	0.660
	cadeia média + max	0.624	0.646	0.687	0.714	0.638	0.409	0.534	0.689	0.565
	cadeia média + min	0.463	0.672	0.594	0.625	0.640	0.395	0.565	0.726	0.554
	cadeia max + min	0.680	0.640	0.689	0.756	0.739	0.396	0.535	0.421	0.385
	todas	0.782	0.800	0.837	0.696	0.625	0.362	0.653	0.735	0.552
$RF_{GloVe/WN}$	switches	0.839	0.805	0.837	0.629	0.613	0.653	0.685	0.614	0.713
	cadeia média	0.600	0.612	0.421	0.550	0.490	0.567	0.594	0.665	0.580
	cadeia min	0.600	0.613	0.423	0.558	0.495	0.566	0.596	0.668	0.588
	cadeia max	0.548	0.552	0.603	0.554	0.499	0.568	0.537	0.471	0.524
	sw + cadeia media	0.804	0.828	0.849	0.616	0.548	0.649	0.681	0.764	0.699
	sw + cadeia min	0.805	0.826	0.854	0.621	0.550	0.655	0.685	0.764	0.692
	sw + cadeia max	0.837	0.780	0.841	0.642	0.577	0.636	0.665	0.639	0.696
	cadeia média + max	0.566	0.610	0.620	0.553	0.370	0.510	0.563	0.593	0.528
	cadeia média + min	0.542	0.627	0.496	0.558	0.560	0.449	0.608	0.600	0.572
	cadeia max + min	0.509	0.564	0.560	0.530	0.522	0.475	0.548	0.432	0.528
	todas	0.773	0.780	0.804	0.605	0.495	0.560	0.645	0.676	0.628
GloVe	switches	0.838	0.834	0.836	0.601	0.591	0.657	0.698	0.599	0.724
	cadeia média	0.498	0.541	0.308	0.511	0.455	0.608	0.531	0.402	0.570
	cadeia min	0.497	0.519	0.315	0.525	0.462	0.601	0.523	0.400	0.585
	cadeia max	0.739	0.726	0.721	0.626	0.629	0.424	0.617	0.589	0.645
	sw + cadeia media	0.796	0.793	0.853	0.582	0.638	0.614	0.654	0.484	0.702
	sw + cadeia min	0.808	0.792	0.853	0.578	0.628	0.615	0.651	0.506	0.701
	sw + cadeia max	0.862	0.839	0.827	0.652	0.653	0.637	0.643	0.577	0.709
	cadeia média + max	0.712	0.672	0.723	0.560	0.529	0.480	0.631	0.515	0.678
	cadeia média + min	0.524	0.518	0.218	0.425	0.411	0.513	0.599	0.401	0.599
	cadeia max + min	0.724	0.613	0.680	0.540	0.424	0.397	0.616	0.474	0.639
	todas	0.862	0.789	0.819	0.582	0.579	0.551	0.615	0.550	0.654
WordNet	switches	0.766	0.665	0.766	0.446	0.449	0.477	0.447	0.527	0.543
	cadeia média	0.683	0.631	0.684	0.422	0.655	0.457	0.646	0.614	0.692
	cadeia min	0.676	0.628	0.689	0.425	0.681	0.501	0.646	0.629	0.691
	cadeia max	0.689	0.660	0.663	0.596	0.479	0.303	0.498	0.446	0.546
	sw + cadeia media	0.781	0.842	0.822	0.400	0.478	0.560	0.658	0.635	0.753
	sw + cadeia min	0.780	0.831	0.811	0.392	0.479	0.542	0.661	0.648	0.745
	sw + cadeia max	0.763	0.783	0.741	0.485	0.484	0.332	0.438	0.621	0.531
	cadeia média + max	0.719	0.607	0.619	0.488	0.456	0.443	0.611	0.538	0.691
	cadeia média + min	0.714	0.696	0.727	0.502	0.478	0.554	0.731	0.567	0.713
	cadeia max + min	0.757	0.677	0.756	0.555	0.481	0.561	0.667	0.590	0.717
	todas	0.795	0.864	0.883	0.475	0.455	0.504	0.687	0.621	0.756

Tabela A.3: Modelo com detector de switches aleatório.

	ALZ			aMCI			a+mdMCI		
	NB	RF	RL	NB	RF	RL	NB	RF	RL
switches	.86	.83	.85	.67	.67	.66	.75	.70	.66
cadeia média	.56	.65	.49	.43	.53	.51	.55	.60	.52
cadeia min	.57	.65	.49	.42	.53	.51	.57	.63	.52
cadeia max	.66	.61	.67	.42	.42	.36	.59	.42	.61
sw + cadeia media	.86	.83	.84	.62	.59	.62	.74	.63	.61
sw + cadeia min	.85	.83	.86	.62	.58	.61	.75	.63	.64
sw + cadeia max	.85	.82	.86	.58	.57	.63	.70	.57	.65
cadeia média + max	.62	.55	.70	.36	.37	.49	.55	.55	.58
cadeia média + min	.49	.53	.45	.65	.64	.54	.50	.50	.47
cadeia max + min	.62	.57	.67	.54	.62	.43	.59	.44	.60
todas	.86	.83	.84	.61	.63	.57	.69	.62	.69