DETECÇÃO DE PLÁGIO DE PARÁFRASE UTILIZANDO AS CARACTERÍSTICAS DO TEXTO

Egberto Caetano Araujo da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2019

DETECÇÃO DE PLÁGIO DE PARÁFRASE UTILIZANDO AS CARACTERÍSTICAS DO TEXTO

Egberto Caetano Araujo da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Eduardo Soares Ogasawara, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2019

Silva, Egberto Caetano Araujo da

Detecção de Plágio de Paráfrase Utilizando as Características do Texto/Egberto Caetano Araujo da Silva. – Rio de Janeiro: UFRJ/COPPE, 2019.

XVII, 105 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2019.

Referências Bibliográficas: p. 76 – 83.

1. paráfrase. 2. características. 3. documento. 4. RST. 5. RAE. 6. POSTagging. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico esse trabalho à Deus, o autor e consumador da minha fé.

Agradecimentos

Primeiramente, agradeço a Deus pela oportunidade de fazer o mestrado em um dos melhores programas de pós-graduação do país, e também pela força e vigor para, até esse presente momento, materializar o trabalho desenvolvido durante esse período acadêmico. Sou grato a Ele pelo sustento e por toda ajuda concedido a mim antes, durante e depois da produção dessa pesquisa. Pois não foi fácil, houve momentos que a vontade de desistir foi muito forte, e em outros, onde, quando faltava-me força, Ele sustentava com sua destra poderosa; quando não encontrava soluções para os meus problemas, Ele concedia-me sabedoria para transpor as dificuldades. Nunca me abondou durante todo esse processo. Ao Deus vivo seja dada a glória, a honra e louvor pois Ele é digno.

Agradeço a minha família pelo apoio e compreensão dado a mim durante esse período que tive voltar toda a minha atenção para desenvolver essa pesquisa. Trago agradecimento em especial a minha mãe pelas palavras de motivação e pelo cuidado em proporcionar um ambiente propício para eu poder pesquisar e fazer essa dissertação em casa. Quero agradecer também a minha linda moça que tanto amo, por acompanhar, por apoiar, por compreender, por ser paciente e carinhosa comigo nos meus momentos de extremo estresse e de angústia, e até nas minhas recaídas.

Agradeço aos presentes membros da banca primeiramente ao meu orientador e professor Geraldo Bonorino Xexéo, do qual obtive conselhos de extrema importância nessa reta final para defesa, e agradeço também ao professor Jano Moreira de Souza e ao professor Eduardo Soares Ogasawara por disponibilizaram-se, concedendo-me tempo de suas vidas corridas para estarem presentes na banca da minha defesa.

Os meus sinceros agradecimentos ao professor Fellipe Duarte, o qual conduziu-me e orientou-me durante toda a pesquisa para essa dissertação. Em diversas momentos priorizou as reuniões necessárias para essa pesquisa e muitas vezes abriu mão do seu descanso para orientar-me remotamente, logo assim que voltava de seus compromissos. Por muitos momentos ficava horas e horas comigo ajudando-me a resolver problemas que enfrentava durante a pesquisa. Com muita alegria, por ele atender-me a noite, o esperava terminar de jantar para conversarmos sobre a pesquisa.

Agradeço aos meus amigos que acompanharam-me durante todas essas jornada, em específico a Michel Dias de Arruda por ser um companheiro sempre presente

em diversos momentos desse mestrado, também a Ricardo Luiz por seus conselhos técnicos que sempre ajudaram-me em minhas implementações. Agradeço também a Joaquim Vianna e Hugo Rebelo por compartilharem comigo essa jornada extraordinária do mestrado.

Quero agradecer ao Programa de Engenharia de Sistema e Computação (PESC) por aceitar-me como seu aluno e proporcionar a expansão dos meus conhecimentos sobre a área da computação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - código de Financiamento 001.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

DETECÇÃO DE PLÁGIO DE PARÁFRASE UTILIZANDO AS CARACTERÍSTICAS DO TEXTO

Egberto Caetano Araujo da Silva

Setembro/2019

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Plágio é a adoção inapropriada de artefatos abstrato ou concreto tais como: textos, obras de arte, ideias ou intenções; sem fazer a devida referência ao seu autor original. Dentre as formas de cometer plágio, existe o plágio de paráfrase, o qual dá-se por meio de manipulações no texto do documento na tentativa de ofuscar a sua real origem. Para a identificação de plágio, é utilizado o *framework* Sistema de Detecção de Plágio Externo (SDPE), o qual contém a tarefa de análise detalhada, onde, dado um documento suspeito, deve identificar se há plágio ou não quando comparado com o conjunto de documentos fontes. O objetivo da pesquisa é atuar na tarefa de análise detalhada , a fim de, com as características léxica, sintática, semântica e estrutural do texto, auxiliar na identificação de plágio de paráfrase entre os documentos. Para isso, acredita-se que, quando o documento é representado por completo, levando em consideração a sua organização, as estruturas em árvores contribuem para identificação de ocorrência de plágio de paráfrase do tipo mais simples ao tipo mais complexo. Para essa tarefa, foi proposto utilizar o *Rhetorical Structure Theory* e o *Part-of-Speech Tagging* para representar as características do documento juntamente com o *Recursive Autoencoder* e o *Dynamic Pooling* detectar casos de plágio de paráfrase em documentos. Durante os experimentos, as abordagens propostas obtiveram entre 83% e 89% de acurácia no *data set* de plágio de paráfrase em documentos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

PARAPHRASE PLAGIARISM DETECTION THROUGH TEXT FEATURES

Egberto Caetano Araujo da Silva

September/2019

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Plagiarism is the improper adoption of abstract or concrete artifacts such as: texts, artwork, ideas or intentions without proper reference to their original author. The ways to commit plagiarism, there is paraphrase plagiarism, which occurs through manipulations in the document text trying to obscure its real source. For the identification of plagiarism, we use the External Plagiarism Detection System (EPDS) framework, which contains the detailed analysis task, where, given a suspicious document, it should identify whether or not plagiarism when compared to the set of document source. The objective of the research is to perform the detailed analysis task in order to, with the lexical, syntactic, semantic and structural characteristics of the text, assist in the identification of paraphrase plagiarism between documents. For this, it is believed that when the document is fully represented, taking into consideration its organization, tree structures contribute to the identification of paraphrase plagiarism from the simplest to the most complex type. For this task, it was proposed to use Rhetorical Structure Theory and Part-of-Speech Tagging to represent document characteristics along with Recursive Autoencoder and Dynamic Pooling to detect cases of paraphrase plagiarism in documents. During the experiments, the proposed approaches obtained between 83% and 89% accuracy in the paraphrase plagiarism data set.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Plágio é a adoção inapropriada de artefatos abstratos ou concretos tais como: textos, obras de arte, ideias ou intenções; sem fazer a devida referência ao seu autor original (MARTIN, 1994; OXFORD, 2019b). Dentre as formas de cometer plágio, existe a plágio de paráfrase, a qual dá-se por meio de manipulações no texto do documento, alterando palavras e estruturas do trecho textual mantendo o sentindo original, e essas modificações têm como objetivo ofuscar a real fonte e o autor do trecho plagiado (MAURER *et al.*, 2006).

Existem diversas formas de identificar plágio, uma é utilizando o *framework* Sistema de Detecção de Plágio Externo (SDPE), o qual é composto por três tarefas e uma delas é a análise detalhada, onde, dado um documento suspeito, deve identificar se há plágio ou não quando comparado com o conjunto de documentos fontes (EISELT & ROSSO, 2009). Nessa tarefa, muitas técnicas são utilizadas para detectar plágio, no entanto, quando trata-se em reconhecer casos de plágio de paráfrase falham (ALZAHRANI *et al.*, 2011). A fim de detectar plágio de paráfrase pelo SDPE através da análise detalhada, é necessário uma forma de representar os documentos de modo a capturar informações que vão além das sequências de palavras e sentenças, é preciso assimilar características intrínsecas ao texto do documento (ALZAHRANI *et al.*, 2011).

1.1 Motivação

Existe um aumento na criação de artigo científicos nos últimos anos em países emergentes, fato que levanta uma discussão sobre a necessidade de aumentar o número de publicações, que por sua vez implica na qualidade das contribuições publicadas (BAUERLEIN *et al.*, 2010; CASATI *et al.*, 2006). Esse crescimento deve-se a pressão vinda dos governos para que o número de publicações seja maior e, como resultado, há uma expansão de publicações poucas referenciadas, o que contribui

para o crescimento de artigo poucos citados, que por sua vez pode contribuir para ocorrência de plágio (BAUERLEIN *et al.*, 2010).

Visto que essa pressão pelo aumento do número de publicações pode funcionar como um motivador para o crescimento das ocorrências de casos de plágio no mundo e como também o desconhecimento sobre publicações científicas, estudos estão apurando o aumento de casos de plágio dentro das instituições de ensino e pesquisa, e MAURER *et al.* (2006) é uma dessas, onde compara dois *surveys* do projeto do *Center os Academic Integrity* (CAI) e afirma que, em 2005, 40% dos alunos admitiram envolvimento com plágio, e a mesma pesquisa feita em 2009 apresenta uma redução na taxa de envolvimento com plágio para 10%. Essa redução comprova que, se metidas forem tomadas para reprimir o envolvimento com plágio, as ocorrências desse tipo podem diminuir. Para isso é necessário meios para detectar esse tipo de prática.

A expansão de incidências de plágio em universidades afetam no crescimento de casos de plágio em artigos submetidos e aceitos em meios de publicações de material científico (DUARTE, 2017). BAKER *et al.* (2008) cita casos de identificação de plágio em trabalhos científicos e os desgostos pela necessidade de uma retratação pública. Mais uma vez esse fato demonstra a necessidade de conseguir detectar plágio antes mesmo que consolide-se no âmbito cientifico ou qualquer outra área, pois gera um grande desgaste para os envolvidos e situações problemáticas para os afetados pela prática.

Existe uma necessidade crescente por sistemas de detecção de plágio automatizados sejam eficazes em sua função, pois pesquisas demonstram que a porcentagem de ocorrência de plágio é grande, como a descoberta feita por INSTITUTE (2012) (citado por DUARTE (2017)), que revelou que dentre 23.000 estudantes dos E.U.A. 74% já copiaram tarefas de outros estudantes e 23% já cometeram plágio utilizando conteúdo da internet. Dado o grande volume de conteúdo disponível na internet torna um tanto impraticável a identificação de plágio de forma manual, reforçando a necessidade de um sistema de detecção automatizado (DUARTE, 2017).

1.2 Definição do problema

O ato de plagiar pode ser feito de diversas formas e a partir do jeito como é feito vai definir o quão difícil é identificá-lo. ALZAHRANI *et al.* (2011) e MAURER *et al.* (2006) apresentam algumas definições e meios para cometer plágio em textos. ALZAHRANI *et al.* (2011) demonstra dois grupos que agrega as formas de cometer plágio, uma delas é o plágio literal, isto é, fazer uma cópia exata do texto, ou reordenação das frases e entre outras; a outra forma é o plágio inteligente que é feito por meio de cópia textual ofuscada usando técnicas mais elaboradas como a paráfrase,

sumarização, tradução e dentre outras. A descoberta de plágio inteligente é difícil pois a forma como é feita pode ser desde alteração de algumas palavras do trecho plagiado ou até mesmo modificações na estrutura textual alterando totalmente a composição original mas sempre mantendo o mesmo sentido e significado da fonte copiada (ALZAHRANI *et al.*, 2011).

De acordo com BARRÓN-CEDEÑO (2012), é difícil para um sistema de identificação de plágio automatizado julgar se há plágio ou não entre dois documentos, ele pode indicar uma possível ocorrência de plágio, no entanto cabe o especialista julgar se indicação informada pelo sistema é de fato uma situação de plágio. Segundo CLOUGH *et al.* (2003), os sistemas de detecção de plágio tem por finalidade: "auxiliar a detecção manual através da redução da quantidade de tempo comparando documentos, possibilitando a comparação de grandes quantidades de documentos e encontrando documentos fontes em recursos eletrônicos disponíveis ao sistema".

Os sistemas de identificação de plágio têm beneficiado-se dos avanços computacionais alcançados nas áreas que lidam com texto, tais como Processamento de Linguagem Natural (PNL), Recuperação de Informação (RI) e Recuperação de Informação Multilíngue (RIM) (ALZAHRANI *et al.*, 2011). Por conta disso, bons resultados tem sido obtidos para prevenção e detecção de plágio. No entanto, alguns dos tipos de plágio são difíceis de identificar, como: paráfrase, sumarização, tradução e mudança na estrutura do texto (ALZAHRANI *et al.*, 2011).

Nos sistemas de detecção de plágio, existe uma tarefa que consiste em fazer uma análise detalhada entre os documentos (EISELT & ROSSO, 2009) e informar se ocorrência de plágio. Nessa tarefa, ALZAHRANI *et al.* (2011) destaca algumas técnicas utilizadas para detectar plágio, como: técnicas baseadas em verificação de caractere, técnicas que utilizam vetores espacias, usam sintaxe, ou usam semânticas, estilometria do texto, a estrutura do texto e entre outras. Por conta de algumas limitações que essas técnicas apresentam os casos de plágio de paráfrase passam desapercebidos como este caso retirado de DOLAN *et al.* (2004):

1. *In only 14 days, US researchers have created an artificial bacteria-eating virus from synthetic genes.*

2. *An artificial bacteria-eating virus has been made from synthetic genes in the record time of just two weeks.*

As duas sentenças transmitem a mesma mensagem mas utilizando estruturas diferente e com bastantes palavras distintas. Para uma técnica de detecção de plágio baseada em verificação de caractere ou vetores espacias podem afirmar que não há uma possível ocorrência de plágio entre essas duas sentenças. Esse presente trabalho visa atuar nessa área detecção plágio, a identificação de plágio de paráfrase em documentos.

1.3 Objetivo e contribuições

O objetivo desse trabalho é atuar dentro do sistema de detecção de plágio, na tarefa de análise detalhada entre documentos, a fim de, com as características do texto (léxica, sintática, semântica e estrutural) informada por ALZAHRANI *et al.* (2011), auxiliar na identificação de plágio de paráfrase em documentos. De acordo com ALZAHRANI *et al.* (2011), muitas abordagens de detecção de plágio não obtém êxito em identificar plágio inteligente, o grupo o qual paráfrase esta inserido, na análise detalhada entre documentos por não levar em consideração a estrutura organizacional do documento, o qual descreve os diversos contextos distribuídos no texto. Com isso, são propostas duas técnicas para representar a estrutura do documento por inteiro levando em consideração todas as suas características textuais, uma utilizando o *Rhetorical Structure Theory* e a outra é por meio da adaptação da abordagem para detecção paráfrase em sentenças feita por SOCHER *et al.* (2011a).

Em resumo, dado um par de documentos, o objetivo desse trabalho é: conseguir representar todo os documentos considerando a sua estrutura organizacional juntamente com suas características textuais, para assimilar informações que vão além de caracteres, palavras e sentenças. A partir desse representação, avaliar a sua capacidade em auxiliar na tarefa de análise detalhada para identificar plágio de paráfrase entre os documentos no idioma em inglês.

As contribuições desse trabalho são estruturas capazes de representar todo o documento facilitando a detecção paráfrase. E, dentre essas estruturas, uma outra contribuição é adaptação da abordagem do SOCHER *et al.* (2011a) para ser aplicada em textos de documentos por inteiro.

1.4 Organização do trabalho

Esse trabalho esta estruturado em sete capítulos: o capítulo 2 apresenta os conceitos sobre plágio, plágio de paráfrase e sistema de detecção de plágio; no capítulo 3 exibe os conceitos teóricos que embasaram esse trabalho; o capítulo 4 apresenta os trabalhos relacionados a pesquisa desenvolvida; o capítulo 5 demonstra como funcionam as propostas desse trabalho; o capítulo 6 demonstra como os experimentos foram conduzidos, avaliados e os resultados alcançados; por fim, o capítulo 7 relata as conclusões obtidas através dos experimentos e quais são as demandas para os trabalhos futuros.

Capítulo 2

Plágio de Paráfrase

Este capitulo segue esta divisão: a seção 2.1 descreve e define plágio; a seção 2.2 apresenta o sistema de detecção de plágio externo com seus componentes; a seção 2.3.1 discorre sobre paráfrase exibindo suas definições e o seu papel no âmbito de plágio; e por fim, a seção 2.4 demonstra a forma como é possível detectar plágio de paráfrase.

2.1 Plágio

Nessa seção serão demonstrados: as definições de plágio, os motivos que impulsionam a prática e também alguns modos de como cometer plágio.

2.1.1 Definições

Plágio é o ato de apropriar-se de trabalho ou de ideias pertencentes à terceiros e fazendo-as aparentar serem suas (OXFORD, 2019b), ou a prática de exibir uma obra intelectual de outra pessoa como se fosse de sua própria autoria (DICIO, 2019b). Ainda de acordo com (DICIO, 2019b), agora sobre o prisma jurídico, trata-se da apresentação que um indivíduo faz de algo, de maneira que, pareça ser de própria autoria, quando na verdade foi criado ou pertence a outrem.

Além das definições citadas nos parágrafo anterior, existem mais definições sobre plágio, como afirma (MERRIAM-WEBSTER, 2019) que diz, ao usar de palavras ou produto intelectual sem dar o devido crédito ao autor; "Cometer furto literário, apresentando como sua ideia ou obra, literária ou científica, de outrem"(MERRIAM-WEBSTER, 2019); "Usar obra de outrem como fonte sem mencioná-la"(MERRIAM-WEBSTER, 2019).

Ainda a respeito de plágio, temos as seguintes definições(DICTIONARY, 2019b):

1. Um ato ou instância de utilização, ou de imitação próxima, da linguagem e pensamento de outro autor, sem autorização, e a repre-

sentação do trabalho de outrem como seu próprio sem o devido crédito ao real autor;

2. Um pedaço de escrita, ou de trabalho, que reflete o uso ou imitação não autorizada.

Essa última afirmação é que mais aproxima-se da intenção desse trabalho. Vale ressaltar que, a partir das definições apresentadas nos parágrafos anteriores, a abrangência das definições sobre plágio, vão além de artefatos visíveis ou tangíveis atingindo o campo abstrato das ideias e das concepções humanas as quais traduzem as reais intenções de quem as detém.

2.1.2 Motivos para o plágio

Segundo MAURER *et al.* (2006), na ocorrência de plágio nem sempre há intenção de o fazer ou furtar alguma coisa de alguém; pode ser por falta de conhecimento sobre trabalhos existentes, ou acidental. Ainda de acordo com MAURER *et al.* (2006), o plágio é tipificado de acordo com a intenção do plagiador, podendo ser dividido em quatro categorias: o **plágio acidental** é gerado pela falta de conhecimento sobre o que é plágio ou de compreender as regras de referências adotadas em um instituto; o **plágio não intencional** tem a sua produção por meio da ampla disponibilização dos conhecimentos, os quais podem induzir os pensamentos de pessoas distintas a convergirem para mesma concepção; o **plágio intencional** é a ação premeditada de copiar parte ou a obra completa de alguém sem atribuir os devidos reconhecimentos ao seu autor primário; e por último, o auto **auto plágio**, que usa um ou mais trabalhos previamente publicado sem referenciá-los.

Para COMAS & SUREDA (2008), o plágio é um fenômeno cultural complexo, o qual influencia as pessoas o praticarem. Ainda de acordo com COMAS & SUREDA (2008), e com foco voltado para instituições educacionais, afirma que existem fatores externos que induzem a prática do plágio, como os anunciados a seguir: a ideia da alta disponibilidade e acessibilidade dos diversos conhecimentos distribuídos pela internet, tornando-os de todos, dando a sensação que, qualquer um pode pode pegar emprestado, usar, apoderar-se e disseminar à vontade; a cultura difundida de que é mais fácil copiar do que reproduzir com criação; constantes exemplos negativos com falta de ética em diversas esferas como corrupção política, fraude acadêmica e entre outras.

Ainda no âmbito educacional, COMAS & SUREDA (2008) declara que existem condições internas relacionados ao sistema de educação que estimulam o ato de plágio pelos alunos, que devem ser punidas. De acordo com BARRÓN-CEDEÑO (2012) (citado por DUARTE (2017)), essas condições internas são divididas em quatro categorias: **orientados ao professor** em que o problema reside nas estratégias de

ensino e no modelo de atribuição de tarefas; **orientados ao aluno** onde o problema se encontra nas atitudes do aluno com relação a escola e ao processo de aprendizagem; **orientados ao sistema educacional** em que o problema está na falta de regras, políticas e instruções claras por parte da instituição educacional.

2.1.3 Formas de plagiar

Observados na seção anterior sobre os motivos que impulsionam o plágio, nessa seção será descrito algumas formas de como o ocorre. Segundo LEUNG & CHAN (2007), existem três grupos macros que enquadram as maneiras de cometer plágio: a cópia de uma fonte que não tem formato eletrônico; a cópia direta de uma fonte com versão eletrônica; e por fim a cópia de uma fonte com versão eletrônica e conteúdo modificado intencionalmente. Dentre esses grupos, esse presente trabalho visa focar em fontes textuais com versões eletrônicas e que podem conter conteúdo modificado intencionalmente.

Dadas as condições em que podem ocorrer o ato de plágio, a seguir são listadas algumas das formas de o fazer:

1. Copiar e colar: copiar palavra por palavra do conteúdo textual (MARTIN, 1994; MAURER *et al.*, 2006) (citado por DUARTE (2017)).

2. Plágio de fonte secundárias: o autor cita a fonte original do trabalho sem informar a fonte secundária de obteve a informação sobre a citação (MARTIN, 1994) (citado por DUARTE (2017)).

3. Plágio da forma da fonte: o plagiador usa a estrutura de argumentação da fonte secundária, olhando e citando a fonte primária do texto, mas não indica que existe uma dependência das citações da segunda fonte (MARTIN, 1994) (citado por DUARTE (2017)).

4. Plágio de ideia: tem a forma abstrata pois usa o pensamento original de outra pessoa sem dependência ou creditar ao autor os devidos reconhecimentos (MARTIN, 1994) (citado por DUARTE (2017)).

5. Plágio de tradução: trata-se do uso do resultado de uma tradução sem fazer as devidas referências ao trabalho original (MAURER *et al.*, 2006) (citado por DUARTE (2017)).

6. Paráfrase: funciona modificando certas palavras, mas não tanto, sem que fonte original seja citada (MARTIN, 1994) (citado por DUARTE (2017)).

7. Coleções misturadas e coladas: consiste na cópia de diversos parágrafos de fontes distintas aparentando que os parágrafos foram colocados em uma sacola

e sacudidos, em seguida colocados aleatoriamente (WEBER-WULFF, 2010) (citado por DUARTE (2017)).

8. Plágio Estrutural: semelhante ao plágio da forma da fonte, pode incluir o uso da estrutura argumentativa, as fontes, as configurações experimentais e até mesmo a pesquisa (WEBER-WULFF, 2010) (citado por (DUARTE, 2017)).

Existem outras formas de cometer plágio, os quais não estão descritas na lista anterior.

Apresentados alguns meios de cometer plágio, existem também contramedidas para os identificar em textos. Para isso, existem estratégias que permitem identificar vestígios que possam indicar se o documento suspeito contém ocorrência de plágio. Uma delas trata-se da **análise de estilo ou estilometria**, a qual baseia-se no estilo de escrita único de cada autor (MAURER *et al.*, 2006). A outra técnica é a **comparação de documento fonte**, que compara o documento suspeito d_s com um conjunto de candidatos a documentos fontes (BARRÓN-CEDEÑO, 2012) (como cita (DUARTE, 2017)). E por último, **a busca de fragmentos**, a qual seleciona de d_s fragmentos de textos que o caracterizam submetendo-os a diversos buscadores (BARRÓN-CEDEÑO, 2012; MAURER *et al.*, 2006). Essas estratégias são comumente utilizadas nos processos contidos no Sistema de Detecção de Plágio.

Um Sistema de Detecção de Plágio pode ser feito de duas maneiras para identificar plágio em textos: um é a **detecção de plágio externo**, o qual consiste na comparação do documento d_s com um ou mais documentos fontes; a outra maneira é a **detecção de plágio interno**, o qual consiste na análise isolada do documento suspeito, não levando qualquer outro documento em consideração (ALZAHRANI *et al.*, 2011).

Um sistema de detecção de plágio externo depende de características externas extraídas do conjunto dos documentos fontes, as quais são utilizadas para comparar com o documento suspeito, a fim de encontrar trechos emprestados (BARRÓN-CEDEÑO, 2012) (citado por DUARTE (2017)). Já o sistema detecção de plágio interno está relacionado com a cadência como o texto evolui no documento suspeito, de modo a perceber diferenças não esperadas entre as partes do texto podendo ser causadas por adição de texto não adaptado, podendo ser de uma fonte externa.

2.2 Sistema de detecção de plágio externo

A figura 2.1 apresenta o fluxo do Sistema de Detecção de Plágio Externo, o qual o processo inicia com a entrada do documentos suspeito d_s e com o conjunto de documentos D, a fim de verificar se há plágio em d_s.

Dado o grande tamanho do conjunto de documentos fontes D, a *tarefa de recuperação heurística* tem por meta escolher os documentos fontes mais relevantes dado o documento suspeito, a fim de reduzir a quantidade de comparações detalhada entre pares que não apresentam nenhuma similaridade (ALZAHRANI *et al.*, 2011). Comumente, Eles são selecionados através de técnicas semelhantes as utilizadas em recuperação da informação, tais como: modelo booleano, vetor de termos, *fingirprint*, *hash-base* e entre outros (ALZAHRANI *et al.*, 2011). Ao fim desse passo, é criado o conjunto dos documentos candidatos D_x que seguem para análise detalhada com o documento suspeito (EISELT & ROSSO, 2009).

A análise detalhada (ou a comparação detalhada) consiste em fazer comparações par a par entre os documentos candidatos e o documento suspeito (EISELT & ROSSO, 2009). O processo de análise da-se por meio de verificações entre as partes do texto do par de documentos, com intuito de encontrar trechos textuais similares, indicando um possível caso de plágio (EISELT & ROSSO, 2009). Esses trechos semelhantes são detectados por intermédio de métodos de comparação que podem atuar desde do nível léxico (caracteres, palavras) até a estrutura do documento (ALZAHRANI *et al.*, 2011). A seção 2.2.1 irá detalhar com mais profundidade sobre essas técnicas.

Ao fim da comparação detalhada, as seções em comum entre d_s e os documentos candidatos, vão para o **pós-processamento**, onde ocorre a filtragem por meio de critérios como: fez corretamente a citação ou discurso literal. Restando, ao fim, apenas as seções suspeitas, as quais são transmitidas para um analista avaliar.

2.2.1 Análise detalhada

A tarefa de análise detalhada é onde concentra-se a área de atuação da proposta desse trabalho, então iremos adentrar com maior profundidade nessa tarefa.

A análise detalhada é responsável por detectar se há plágio ou não entre os pares de documentos (EISELT & ROSSO, 2009). Para isso, é necessário saber lidar com as características do texto, as quais o representam e o definem diante da análise em relação aos demais (ALZAHRANI *et al.*, 2011). Segundo ALZAHRANI *et al.* (2011), para o sistema de detecção de plágio externo, existem quatro características textuais que devem ser consideradas: a **característica léxica** atuante em nível caractere ou palavra, *n-gram* de caracteres ou *n-gram* de palavras; a **característica sintática** age nos blocos textuais, como sentenças, frases; a **característica semântica** lida com significado das palavras, sinônimos, antônimos; e por fim a **característica estrutural** que leva em consideração o contexto de acordo com a localidade no texto. Vale ressaltar, que, essas características compreendem a granularidade do

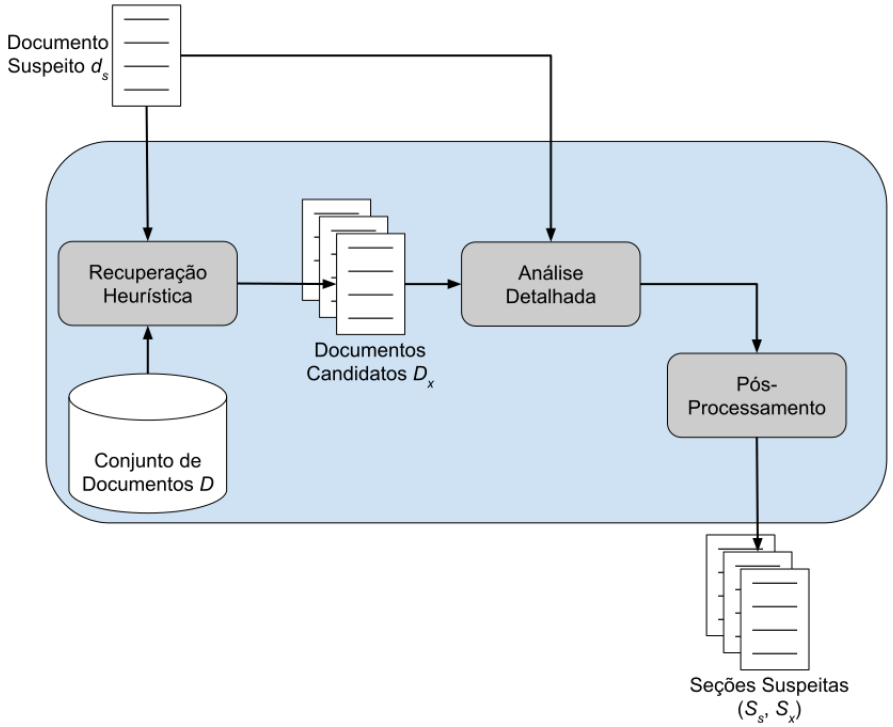


Figura 2.1 – Sistema de Detecção de Plágio Externo. Reprodução da imagem em EISELT & ROSSO (2009)

texto contido no documento, sendo o léxico o mais granular e o estrutural mais amplo e geral (CHOW & RAHMAN, 2009).

Para identificar plágio em documentos, torna-se necessário saber comparar, manipular e avaliar as características dos textos (ALZAHRANI *et al.*, 2011). Muitas propostas para a tarefa de comparação detalhada utilizam apenas uma característica do texto, como os métodos baseados em *string match*, os quais utilizam apenas informações léxicas do texto. Esse tipo de abordagem é facilmente ludibriada por meio da paráfrase que substitui determinadas palavras em determinado trecho do texto, ou por meio do plágio de ideia que absorve apenas a semântica contido no trecho e o reescrever de forma totalmente diferente (ALZAHRANI *et al.*, 2011).

Outras abordagens conseguem combinar mais de uma característica textual para executar a tarefa de análise detalhada, como os métodos que utilizam modelos vetoriais para representar *n-gram* ou sentenças ou até documentos inteiros. Contudo, não é a forma mais eficiente de representar o documento inteiro pois, as relações contextuais atreladas as palavras que auxiliam na interpretação semântica do uso das mesmas, são perdidas porque elas não são mantidas no vetor de representação de um documento (ALZAHRANI *et al.*, 2011).

Existem métodos que atuam com a característica sintática a fim de conseguir informações semânticas sobre as palavras ou até as sentenças (SOCHER *et al.*, 2011a). Essas técnicas costumam utilizar *Part-Of-Speech Tagging* para fazer as

marcações das relações gramaticais entre as palavras pertencente ao trecho textual (ALZAHRANI *et al.*, 2011). A partir da estrutura fornecida pelo *Part-Of-Speech Tagging*, tenta-se verificar a ocorrência de plágio por meio da árvore sintática com nós em comum entre o par de documentos. As árvores sintáticas são representações estruturais para as sentenças (ADHVARYU & BALANI, 2015), sendo necessário dividir o documento em sentenças, na maioria das vezes, sem demarcar as relações estabelecidas entre elas.

Há uma necessidade de reter e ser capaz de representar as informações sobre as relações entre as partes do texto (palavras, sentenças, parágrafos) de modo conectado, não desconexo, pois o conjunto dessas relações fornece o conhecimento sobre a estrutura do texto, que por sua vez contribui para identificar plágio estrutural (CHOW & RAHMAN, 2009).

ALZAHRANI *et al.* (2011) afirma, que as características do texto são essenciais para detectar diferentes tipos de plágio e recomenda o uso combinado da característica estrutural com as demais características para aumentar a detecção de tipos diferentes de plágio.

2.3 Paráfrase

2.3.1 Definição

Paráfrase é a interpretação de um texto através das próprias palavras, de modo a manter o mesmo pensamento do original; ou a reprodução de ideias e conteúdos de um texto, livro ou narrativa, dando-lhes uma nova interpretação, tornando-os mais perceptivos, atribuindo-lhes uma perspectiva, sem alterar seu sentido inicial (DICIO, 2019a).

De acordo com (OXFORD, 2019a), expressa o significado de (algo escrito ou falado) usando palavras diferentes, especialmente para obter maior clareza, ou uma reformulação de algo escrito ou falado mantendo o seu sentido original.

Para BHAGAT & HOVY (2013), paráfrases são sentenças ou frases que transmitem o mesmo ou o quase igual significado usando palavras diferentes. Um exemplo simples de paráfrase, retirado de BHAGAT & HOVY (2013):

1. *The school said that their buses **seat** 40 students each.*

2. *The school said that their buses **cram** in 40 students each.*

Observando as sentenças 1 e 2, note-se que as palavras *seat* e *cram*, embora diferentes e com significados não tão semelhantes, a substituição de uma por outra não modificou o sentido mais amplo da sentença, que é informar que os ônibus (buses) conseguem comportar quarenta estudantes.

Agora um exemplo mais complexo, retirado de DOLAN *et al.* (2004):

3. *There was no chance it would endanger our planet, astronomers said.*

4. *NASA emphasized that there was never danger of a collision.*

A primeira complexidade para identificar a paráfrase entre as sentenças 3 e 4, inicia com o não compartilhamento de palavras em comuns; segundo, há uma necessidade em ter o conhecimento do contexto por de trás das sentenças para que seja possível identificar a paráfrase. Ambas as sentenças explicitam que o planeta Terra não corre perigo de colisão com um asteroide.

2.3.2 Paráfrase no plágio

Como descrito na seção 2.1.3, paráfrase é uma forma de plagiar quando os reconhecimentos necessários para o autor não são apresentados de modo correto. ALZAHRANI *et al.* (2011) apresenta uma taxonomia sobre plágio como demonstra a tabela 2.1. Essa taxonomia é dividida em dois grupos maiores, que são: o **plágio literal**, o qual é o delito mais cometido e mais fácil de executar; o **plágio inteligente** tenta ocultar, ofuscar, e mudar o trabalho original de forma elaborada, por meio de manipulação de texto, tradução e adoção de ideias (ALZAHRANI *et al.*, 2011).

Plágio	Literal	Cópia exata	Documento inteiro
			Partes do documento
		Cópia aproximada	Inserir
			Remover
			Substituir
			Divisão ou junção das sentenças
		Cópia modificada	Reordenação da sentença
			Sintaxe
	Inteligente	Manipulação de texto	Paráfrase
			Sumarização
		Tradução	Manual
			Automática
		Adoção de ideia	Baseado em significado semântico
			Baseado em seção
			Adaptação de contexto

Tabela 2.1 – Taxonomia de plágio. Reprodução baseada em ALZAHRANI *et al.* (2011)

O plágio de paráfrase, de acordo com ALZAHRANI *et al.* (2011), é considerado um tipo de plágio inteligente que é feito através da manipulação do texto, podendo alterar entre uma ou duas palavras, ou reescrevendo todo o trecho mantendo o sentido original.

2.4 Identificando plágio de paráfrase

De acordo com ALZAHRANI *et al.* (2011), diversos métodos para identificação de plágio em documentos falham em detectar paráfrase, sumarização, adoção de ideia, porque focam apenas em lidar com as formas mais fáceis de plágio, cópia exata, por exemplo. Ainda de acordo com ALZAHRANI *et al.* (2011), muitos modelos levam em consideração somente as características das sentenças (léxica, sintática e semântica), prejudicando a detecção de casos de plágio mais complexo que precisam, também, das informações do contexto os quais as características locais estão inseridas.

ALZAHRANI *et al.* (2011); CHOW & RAHMAN (2009) afirmam sobre a necessidade de haver uma estrutura capaz de representar o documento mantendo a sua organização, consiga assimilar todas as suas características textuais locais, e seja eficiente em reconhecer as relações estabelecidas entre as partes do texto. Pois, conseguindo localizar a ocorrência de um possível caso de plágio dentro do documento, com a estrutura organizacional do texto, torna viável fazer uma análise contextual levando em consideração os trechos de texto ao redor do possível caso de plágio.

Baseado nas afirmações de ALZAHRANI *et al.* (2011); CHOW & RAHMAN (2009), a identificação de plágio de paráfrase não deve ser feita somente no âmbito da característica léxica ou da sintática, torna-se necessário o contexto, o qual vem por meio da estrutura organizacional do texto, que retém as relações entre os segmentos textuais e as suas localizações. Isso permite inferir a semântica dos segmentos de acordo com o contexto o qual está inserido. Desse modo, é possível identificar casos de paráfrase complexas tais como o descrito na seção 2.3.1 com as sentenças 3 e 4.

Capítulo 3

Representação das Características do Texto

Neste capítulo são descritas as teorias que embasaram esse trabalho. Alguns conceitos são tratados com maior detalhamento por não serem tão disseminados no meio científico. O capítulo está dividido da seguinte forma: a seção 3.1 descreve as características inerentes ao texto, as quais são necessárias para detectar plágio de paráfrase; a seção 3.2 apresenta a *Rhetorical Structure Theory*; a seção 3.3 faz uma breve definição do *Part-Of-Speech Tagging*; a seção 3.4 demonstra a descrição sobre *word embedding* e sobre a rede neural *Recursive Autoencoder*.

3.1 Características do texto

Como foi mencionado na seção 2.2, no sistema de detecção de plágio externo, a tarefa de comparação detalhada é responsável pela identificação de quais documentos candidatos tiveram o conteúdo reutilizado pelo documento suspeito de cometer plágio. A fim de alcançar o resultado esperado por essa tarefa, é necessário extrair características representativas dos documentos, permitindo-os que sejam analisados uns com outros. ALZAHRANI *et al.* (2011) separa essas características em quatro grupos: características léxicas, características sintáticas, características semânticas e características estruturais.

As **características léxicas** atuam no nível caractere ou palavra, onde, respectivamente, o documento é representado por um conjunto de caracteres ou por um conjunto de palavras. Por exemplo, um documento pode ser representado por uma sequência de caracteres $d = \{(c_1, d), (c_2, d), ..., (c_n, d)\}$ ou representado por uma sequência de palavras $d = \{(w_1, d), (w_2, d), ..., (w_n, d)\}$, onde (c_i, d) e (w_i, d) são o *n-ésimo* caractere e a *n-ésima* palavra do documento d, respectivamente.

As **características sintáticas** abrangem as frases e as palavras em diversas declarações dentro do documento, podendo classificá-las como verbo, substantivo, adjetivo, advérbio e em outras classes gramaticais. A representação por características sintáticas dá-se quando o documento é dividido em sentenças, em seguida, elas são submetidas a classificação gramatical construindo o conjunto de sentenças classificadas gramaticalmente $d = \{(s_1, d), (s_2, d), ..., (s_n, d)\}$, onde (s_i, d) é a *n-ésima* sentença do documento d.

As **características semânticas** buscam representar os significados adequados para os segmentos de qualquer tamanho (palavras, sentenças, parágrafos) dentro de um contexto expresso no documento. Esses significados podem ser capturado pela observação do uso das classes gramaticais, sinônimos, antônimos, hiperônimos e hipônimos[1].

As **características estruturais** reproduzem a organização do texto e capturam o significado ou entendimento que o autor deseja transmitir ao leitor. A representação pode ser por meio de conjuntos de blocos textuais (palavras, sentenças, parágrafos ou segmentos de qualquer tamanho) que respeitem a estrutura organizacional do texto. Por exemplo, $d = \{(b_1, p_1, d), (b_2, p_2, d), ..., (b_n, p_m, d)\}$, onde (b_i, p_j, d) é o *n-ésimo* bloco em uma determinada posição m no documento d.

Saber representar o significado ou entendimento do texto é importante tanto para identificar as diferentes formas de plágio inteligente, que vão desde manipulação de texto até a adoção de ideias, quanto para a detecção de plágio de paráfrase (ALZAHRANI *et al.*, 2011).

De acordo com CHOW & RAHMAN (2009) e ALZAHRANI *et al.* (2011), utilizar apenas os modelos clássicos (*string match*, vetores espaciais termo-documento, alinhamento textual, entre outros) para representar os documentos para detecção de plágio é rudimentar, pois informações contextuais não são levadas em consideração, por exemplo: dois documentos podem ter a frequência de termos similar mas podem divergirem em relação ao contexto quando a distribuição espacial dos termos são diferentes, isto é, *information* e *retrieval* assumem outros significados em diversas partes do documento comparado a *information retrieval* no momento em que aparecem juntos (CHOW & RAHMAN, 2009). Por esse fato torna-se necessário dar a devida importância a similaridade entre os contextos do documento, levando em consideração a forma como as palavras são utilizadas através do texto, ou seja, quais são os papéis delas dentro de uma sentença, uma seção, um parágrafo (ALZAHRANI *et al.*, 2011). Para inserir essa informação espacial nas representações dos documentos, segundo CHOW & RAHMAN (2009), a estrutura em árvore é a

[1]hiperônimo: é uma palavra que pertence ao mesmo campo semântico de outra mas com o sentido mais abrangente, exemplo, legume é mais abrangente que cenoura; hipônimo: análogo ao hiperônimo mas com sentido mais restrito, exemplo, rosa é mais especifica que flor

mais adequada, pois dependendo da forma como ela é construída oferece informações detalhadas por nível.

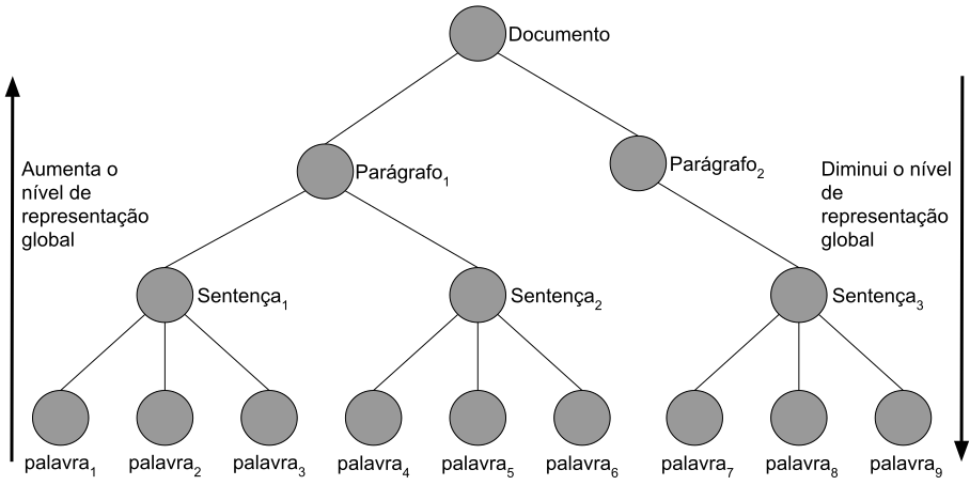


Figura 3.1 – Representação do documento pela estrutura em árvores. Baseado em CHOW & RAHMAN (2009)

A figura 3.1 demonstra a capacidade de representação da estrutura em árvore, começando pelas folhas que são as palavras chegando até raiz que pode funcionar como uma representação, sem muito detalhamento, para o todo o documento. Conforme sobe os níveis da árvore a representação do documento adquire característica mais globais perdendo detalhes, já descendo obtém maior detalhamento, chegando até as palavras, que mantém a ordem conforme aparecem no documento (CHOW & RAHMAN, 2009). Observando a figura em questão é possível perceber os recursos oferecido por essa estrutura, como: a ordem das palavras; as relações entre palavras, sentenças e parágrafos; e a hierarquia entre as partes do documento (MANN & THOMPSON, 1987).

3.2 *Rhetorical Structure Theory*

Rhetorical Structure Theory (RST) é um *framework* descritivo focado na organização do texto, provendo uma forma genérica de descrição para as relações entre os segmentos do texto, por meio de demarcações gramáticas ou léxicas (MANN & THOMPSON, 1988); e é útil para identificar o significado das conjunções, das combinações dos segmentos gramaticais, e parataxes[2] não sinalizados.

RÖSNER & MANFRED (1986) e MANN & THOMPSON (1987) comentam que o RST demonstra ser profícuo na análise da narrativa do discurso, como também serve para descrever as propriedades gramaticais e retóricas da narrativa, e, MANN

[2]Maneira de construir as frases, própria das crianças e das línguas primitivas, a qual consiste em empregar apenas proposições principais, sem conjunções coordenativas ou subordinativas (https://dicionariodoaurelio.com/parataxe, https://pt.wikipedia.org/wiki/Parataxe)

& THOMPSON (1986) diz que, o RST provê uma estrutura hierárquica para investigar as proposições relacionais, nas quais não são declaradas mas são proposições inferidas que surgem da estrutura do texto no processo de interpretação textual. A coerência textual, em partes, depende das proposições relacionais, e o RST apresenta meios para realizar o estudo sobre ela (MANN & THOMPSON, 1988).

MANN & THOMPSON (1988) afirmam que, além das características citadas nos parágrafos anteriores, a aplicação do RST não se limita ao tamanho do texto podendo ser utilizado em textos com tamanhos variados. Porém, apresenta algumas limitações, como somente sendo aplicável à monólogos textuais não sendo aplicável à diálogos textuais e também não se estende à textos multilíngual (MANN & THOMPSON, 1988).

3.2.1 Elementos do RST

Antes de conhecer os elementos do RST, algumas terminologias serão apresentadas. Segundo MANN & THOMPSON (1987) uma *text span* é um intervalo linear ininterrupto do texto. O termo escritor refere ao autor do texto em análise; leitor faz referência ao plúblico pretendido do texto (MANN & THOMPSON, 1987). E e L representam escritor e leitor respectivamente (MANN & THOMPSON, 1987). O analista (A) é a pessoa que faz o julgamento para produzir a análise sobre o texto (MANN & THOMPSON, 1987).

O RST contém quatro elementos importantes que são: Relações, *Schemas*, *Schema Applications* e Estruturas (MANN & THOMPSON, 1988).

Em resumo, as relações identificam o relacionamento mantido entre duas *text spans* (MANN & THOMPSON, 1988). Com base nas relações, os *schemas* definem qual padrão será aplicado entre uma parte específica do texto em relação a outra parte (MANN & THOMPSON, 1988). O *Schema Applications* convenciona a forma como um *schema* pode ser instanciando (MANN & THOMPSON, 1988). Estrutura é definido pela composição de *schema applications* que unidos devem contemplar todo o texto (MANN & THOMPSON, 1988).

3.2.2 Relações

As Relações definem o tipo de interação entre duas *text spans* não sobrepostas; essas partes doravante conhecidas como núcleo e satélite, N e S respectivamente (MANN & THOMPSON, 1987). O núcleo é considerado a parte mais importantes em um análise, enquanto que o satélite colabora para enfatização do assunto exposto no núcleo (TABOADA & MANN, 2006). O núcleo é o mais essencial que o satélite para a proposta do escritor. O satélite comumente não é entendido sem o núcleo (TABOADA & MANN, 2006). Logo, para a definição de uma relação, TABOADA

& MANN (2006) propõe quatro considerações que devem ser observadas: **restrição no núcleo**; **restrição no satélite**; **restrição na combinação núcleo e satélite**; **e efeito**.

Cada um dos quatro itens mencionados por TABOADA & MANN (2006), especificam um julgamento particular que o analista do texto deve fazer na construção da Estrutura do RST. TABOADA & MANN (2006) afirmam que, sabendo a natureza da análise de texto, esses itens são julgamentos de plausibilidade ao invés de certeza. No caso do item efeito, o analista esta julgando se é plausível o intento do escritor, expresso na condição entre as *text spans*, seja especificada (TABOADA & MANN, 2006).

3.2.2.1 Relações e suas definições

Nesta seção serão apresentadas algumas relações com os seus nomes e suas definições. Todas as demais relações podem ser encontradas no site do *Rhetorical Structure Theory*, na url: `https://www.sfu.ca/rst/01intro/definitions.html`.

Evidência: Conforme afirmam MANN & THOMPSON (1988), o satélite dessa relação tem a intenção de aumentar a credibilidade do conteúdo expresso no núcleo.

Restrição no N:	O leitor pode não acreditar em N em um grau satisfaróio para o escritor
Restrição no S:	O leitor acredita em S ou achará crível
Restrição em N e S:	O leitor compreende que S aumenta a crença do leitor em N
Efeito:	A crença do leitor em N é aumentada

Circunstância: Nessa relação, MANN & THOMPSON (1987) afirmam que o satélite funciona como um enquadramento temporal ou espacial, dentro do qual o leitor consegue interpretar o núcleo.

Restrição no N:	-
Restrição no S:	S apresenta uma situação
Restrição em N e S:	S define o contexto do assunto dentro do qual o leitor destina-se a interpretar a situação apresentada em N
Efeito:	O leitor reconhece que a situação apresentada em S fornece o contexto necessário para interpretar N

Antítese: O efeito desejado dessa relação é causar uma consideração positiva no leitor sobre o núcleo por meio do contraste (MANN & THOMPSON, 1988).

Restrição no N:	O escritor tem uma consideração positiva da situação apresentada em N
Restrição no S:	-
Restrição em N e S:	A situação apresentada em N e S contrastam entre si; por causa da incompatibilidade que surge do contraste entra elas, não há como as duas situações serem apresentadas com considerações positivas; logo, S, com sua incompatibilidade com a situação apresentada em N, aumenta a consideração positiva do leitor em relação a N
Efeito:	Aumenta a consideração positiva do leitor em relação a N

Habilitação: Essa relação estimula os leitores a agirem por meio de convites, pedidos, ofertas, comandos ou sugestões, feitos de maneira a fornecer informações necessária para aumentar a vontade do leitor em executar a ação proposta (MANN & THOMPSON, 1988).

Restrição no N:	Apresenta uma ação do leitor (incluindo a aceitação de oferta) não realizada no contexto apresentado por N
Restrição no S:	-
Restrição em N e S:	O leitor compreende que S aumenta o seu potencial em executar a ação apresentada em N
Efeito:	O potencial do leitor em executar a ação apresentada em N é aumentada

3.2.3 Esquemas

Esquemas (*Schemas*) definem o arranjo dos componentes estruturais do texto (MANN & THOMPSON, 1987). Eles são padrões abstratos formados por um pequeno número de *text spans*, uma relação específica entre essas partes, e uma definição de como certas *text spans*, núcleos, estão relacionados com toda a coleção dos componentes estruturais do texto (TABOADA & MANN, 2006).

Schemas, definidos no que diz respeito a relações, determinam como as *text spans* coocorrem; acompanhados por *schema applications*, eles definem as possíveis Estruturas do RST para o texto (MANN & THOMPSON, 1988).

RST reconhece cinco tipos de *schemas*, os quais são apresentados na figura 3.2. As curvas representam as relações entre as *text spans*, as retas identificam os núcleos

entre as partes envolvidas (TABOADA & MANN, 2006). Os *schemas* para relações não apresentados na figura 3.2, quase todos eles seguem o padrão apresentado pela circunstância: uma única relação com núcleo e satélite. O nome dos *schemas* são os mesmos das relações correspondentes (MANN & THOMPSON, 1987).

Os *schemas* multinuclear são utilizados para representar uma porção do texto que não segue o padrão apresentado anteriormente (MANN & THOMPSON, 1988). Por exemplo, contraste *schema* sempre há exatamente dois núcleos (MANN & THOMPSON, 1988). Sequência e junção têm inúmeros núcleos, um para cada elemento da sequência, e uma relação sucessora para o núcleo adjacente (TABOADA & MANN, 2006). Observa-se que, para relações junção e sequência, os seus *text span* são considerados núcleos por convenção, desde que não haja satélite correspondente (MANN & THOMPSON, 1987).

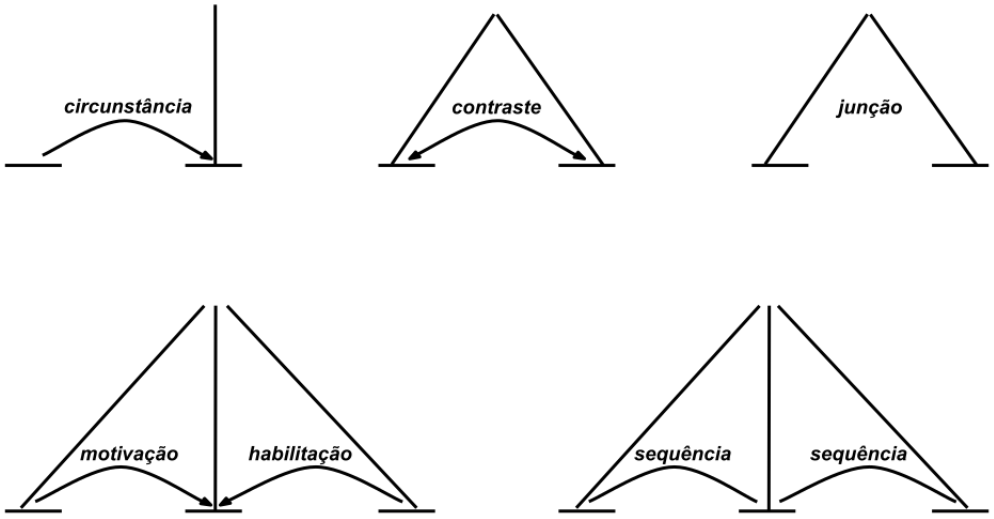


Figura 3.2 – Cinco tipos de *schemas*. Reprodução da imagem em MANN & THOMPSON (1987)

3.2.4 Aplicação de esquemas

Aplicação de esquemas (*Schema applications*) definem como serão utilizados os *schemas* na estrutura do texto (MANN & THOMPSON, 1987). Existem três convenções que determinam a possibilidade da aplicação dos *schemas*:

1. os ***text spans* desordenados** definem que os *schemas* não devem restringir a ordem dos núcleos ou satélites no *text span* na qual o *schema* esta sendo aplicado (MANN & THOMPSON, 1988);

2. as **relações opcionais** afirmam que, para todo o *schema multi-relations*, todas as relações individuais são opcionais, mas pelo menos uma dessas deve ser mantida (MANN & THOMPSON, 1988);

3. segundo as **relações repetidas**, uma relação que faz parte de um *schema* pode ser utilizado quantas vezes for necessário na aplicação desse mesmo *schema* (MANN & THOMPSON, 1988).

3.2.5 Estruturas

Nesta são apresentando as Estruturas, que são: estrutura da análise e diagrama da estrutura.

3.2.5.1 Estrutura da análise

Segundo MANN & THOMPSON (1987) a estrutura da análise é um conjunto de *schema applications* que devem seguir as seguintes condições: a **completude** diz que o conjunto contém um *schema application* que contém um conjunto das *text spans* que por sua vez constituem o texto inteiro (MANN & THOMPSON, 1988); a **conectividade** declara que, exceto pelo texto inteiro como uma *text span*, cada *text span* na análise é uma unidade mínima e também um constituinte de um outro *schema application* da análise (MANN & THOMPSON, 1988); a **unicidade** diz que cada *schema application* é composto por um conjunto diferente de *text spans*, e dentro de um *schema multi-relation* cada relação deve ser aplicado para um conjunto diferente de *text spans* (MANN & THOMPSON, 1988); a **adjacência** declara como as *text spans* de cada *schema application* constituem uma *text span* em relação a outra *text span* mais abrangente (MANN & THOMPSON, 1988).

3.2.5.2 Diagrama da estrutura

O diagrama de estrutura representa a organização estrutural do RST para o texto como mencionam MANN & THOMPSON (1987). Nesse diagrama: os arcos, rotulados com os nomes das relações, conectam fragmentos de uma estrutura para a qual a relação se mantém (TABOADA & MANN, 2006); cada linha vertical proveniente da *text span* será descomposta por *schema application* até o núcleo do mesmo (TABOADA & MANN, 2006); os números representam a sequência não descomposta das unidades da estrutura (MANN & THOMPSON, 1988). A seção a seguir, com a figura 3.3 é demonstrado um exemplo do diagrama da estrutura da análise.

3.2.6 Exemplo de uma análise usando RST

Temos o seguinte trecho retirado de um programa de rádio chamado "*Meet the Announcers*"(MANN & THOMPSON, 1988): *P. M. has been with KUSC longer than any other staff member. While attending Occidental College, where he majored*

in philosophy, he volunteered to work at the station as a classical music announcer. That was in 1970.

Norteado por um dos princípios do RST, o qual afirma que o texto deve ser dividido em unidades com integridade funcional independente, o trecho foi segmentado da seguinte maneira:

1. *P. M. has been with KUSC longer than any other staff member.*
2. *While attending Occidental College,*
3. *where he majored in philosophy,*
4. *he volunteered to work at the station as a classical music announcer.*
5. *That was in 1970.*

Após a segmentação, é dado início a análise entre os *text spans*. Comparando os dois *text spans* 4 e 5 pode ser percebido que o evento descrito no *text span* 4 apresenta maior importâcia que o *text span* 5, pois a intenção do escritor é evidenciar o início do relacionamento entre duas entidades. Logo, podemos classificar os dois como: o *text span* 4 como núcleo enquanto que o *text span* 5 é o satélite. Observando a forma como esse dois *text spans* interagem, é possível observar que o *text span* 5 temporaliza o *text span* 4, alocando o evento em uma linha temporal. Então podemos inserir uma relação do tipo circunstância entre eles.

Seguindo para o próximo par de *text spans*, 2 e 3. Para definir quem é o núcleo entre esse dois *text spans*, precisa-se captar a intenção do escritor nesse texto. Nota-se que o escritor quer evidenciar o início de um longo relacionamento entre duas entidades do trecho, apresentando local e o momento. Com isso em mente, e observando o par de *text spans* em questão, é possível perceber que o *text span* 2 mostra o local que o escritor deseja transmitir ao leitor. Conclui-se então, que o *text span* 2 é o núcleo e o *text span* 3 é o satélite. Como o *text span* 2 apresenta o local do início do relacionamento entre essas duas entidades do trecho, o *text span* 3 adiciona mais informações dando maior detalhes no fato descrito no *text span* 2, essa interação caracteriza uma relação de elaboração.

Observando para os dois pares de *text span* que acabaram de serem analisados, agora será avaliado como eles relacionam-se. O par 4-5 contém a informação central que é quando as duas entidades começaram a relacionar-se. Enquanto que o par 2-3 apresenta o local de onde começou esse relacionamento. Então pode-se classificar o par de *text spans* 4-5 como núcleo e o par de *text spans* 2-3 como satélite. Como o par de *text span* 2-3 está dando a localização do envento importante descrito no par 4-5, pode-se classificar a relação entre esse pares como elaboração.

O *text span* 1 é o ultimo a ser incorporado à estrutura do texto e será analisado em relação a todo conjunto já descrito anteriormente. Como o desejo do escritor é demonstrar a longa duração de um relacionamento entre as entidades apresentadas, isso fica evidente no *text span* 1, o tornando núcleo dessa comparação. Enquanto que o conjunto torna-se o satélite. Sabendo que o *text span* 1 é o núcleo, pode observar que o conjunto das outras *text spans* serviu para dar maior credibilidade a afirmativa apresentada no núcleo. Esse aumento de credibilidade veio por meio de um detalhamento descritos nos outros *text spans*. A relação que melhor define essa interação é elaboração.

Ao fim dessa análise é gerado o seguinte diagrama da estrutura, como é demonstrado na figura 3.3.

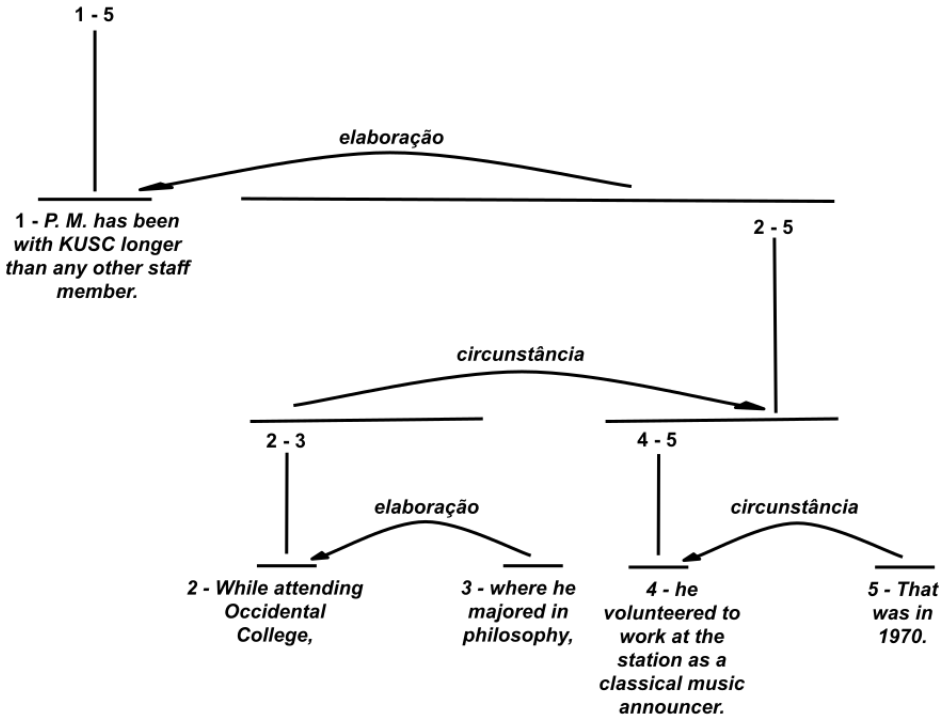


Figura 3.3 – Exemplo da Estrutura do trecho "*Meet the Announcers*". Reprodução da imagem em MANN & THOMPSON (1987)

É possível notar que o diagrama gerado segue as normativas descritas anteriormente. As disposições dos *schemas* estão alinhados com a seção 3.2.3. Assim como também estão seguindo as exigências descritas na seção 3.2.4. Por fim o diagrama de estrutura está adequada aos requisitos da seção 3.2.5.1 solicita.

3.2.7 Unidade elementar de discurso

Segundo MANN & THOMPSON (1988), é necessário saber que o RST parte do princípio do qual o texto precisa ser dividido em unidades, do qual o tamanho da unidade é arbitrário mas a divisão deve estar fundamentada em alguma classificação

teórica neutra. Para isso, a fim de alcançar resultados interessantes, as unidades devem ter integridade funcional independente (MANN & THOMPSON, 1987).

Executar essa divisão torna-se complexa quando tenta manter informações importantes para uma análise do texto feita por um computador, afirma MARCU (1999). Ainda MARCU (1999) continua dizendo que, essa é uma tarefa extremamente difícil pois os limites entre o sintático, o semântico e a informação retórica não são tão claros. Segundo STEDE *et al.* (2017), a segmentação deve gerar unidades mínimas do texto, ou seja, unidades mínimas com integridade funcional independente. Na tentativa de atender esses princípios, MARCU (1999) introduz o termo unidade elementar de discurso (*Elementary Discourse Unit*), mais conhecido como EDU. Pois, como afirma STEDE *et al.* (2017), o termo *text span* é muito abrangente para descrever qualquer unidade de discurso, simples (uma simples sentença) ou complexa (uma sentença complexa ou a combinação de sentenças), por conta desse fato, o conceito EDU é introduzido no universo do RST.

A respeito de sentenças, MARCU (1999) diz não ser uma boa escolha para serem EDUs no RST. Pois ao escolhê-las, muitas informações retóricas ficarão fora do escopo da análise. Ele demonstra esse fato através do exemplo a seguir. Tem-se a seguinte sentença:

> *Only the midday sun at tropical latitudes is warm enough to thaw ice on ocasion, but any liquid water formed in this way evaporate almost instantly because the atmospheric pressure is low.* (3.1)

Caso fosse escolhida sentenças para serem as EDUs do texto, após a segmentação duas EDUs seriam geradas como no exemplo a seguir:

> *[Only the midday sun at tropical latitudes is warm enough to thaw ice on ocasion*[1]*], [but any liquid water formed in this way evaporate almost instantly because the atmospheric pressure is low.*[2]*]* (3.2)

Quando analisadas essas duas EDUs, percebe-se que há uma relação do tipo contraste entre elas. Algumas informações retóricas não são capturadas quando utiliza-se sentenças para serem EDUs. Agora veja no próximo exemplo como fica a análise do mesmo texto ao utilizar outra forma para segmentar o texto.

[Only the midday sun at tropical latitudes is warm enough[1]] [to thaw ice on ocasion[2]], [but any liquid water formed in this way evaporate almost instantly[3]] [because the atmospheric pressure is low.[4]] (3.3)

Escolhendo cláusulas[3] como EDUs, a segmentação do mesmo texto gera 4 EDUs. Fazendo a análise dessas EDUs, observa-se que entre a unidade 1 e a unidade 2 há uma relação do tipo propósito. Enquanto que entre a unidade 3 e a unidade 4 há uma relação do tipo explicação. E fazendo uma análise entre esses dois conjuntos é verificado a mesma relação identificada no exemplo de segmentação anterior, contraste.

Existem outras propostas para escolhas de EDUs para fragmentar o texto além da descrita em MARCU (1999). Como em TOFILOSKI *et al.* (2009) e em STEDE *et al.* (2017). Todos têm como base os princípios básicos do RST descritos em MANN & THOMPSON (1987). Nesse trabalho a abordagem utilizada para fragmentar o texto é a descrito por MARCU (1999).

3.3 *Post-of-seepch Tagging*

Part-of-speech Tagging (*POS Tagging*) é uma tarefa da área de processamento de linguagem natural que consiste em determinar qual a classe gramatical para cada palavra em uma sentença (ADHVARYU & BALANI, 2015). Segundo JURAFSKY & MARTIN (2014), as palavras da língua inglesa podem ser classificadas nas seguintes classes gramaticais: substantivo, verbo, adjetivo, advérbio, preposição, pronome, conjunção, interjeição.

Seja utilizada a sentença a seguir para exemplificar o uso do *POS Tagging*: *The boy kicked the ball that he bought.*

The(DT) boy(NN) kicked(VDB) the(DT) ball(NN) that(WDT) he(PRP) bought(VDB). (3.4)

A siglas representam: DT → Determinante; NN → Substantivo (*noun*); VDB → Pretérito perfeito (verb pasttense); PRP → Pronome pessoal (personal pronoun); WDT → Determinante (wh-determiner).

[3]Conjunto de palavras ordenadas de acordo com normas gramaticais, com sentido completo (`https://www.dicio.com.br/clausula/`).

3.3.1 Tipos de POS Tagging

ADHVARYU & BALANI (2015) afirmam que o POS Tagging pode ser dividido em dois métodos: supervisionado e não supervisionado. No método supervisionado, existe um *data set* de textos com suas palavras anotadas com suas respectivas *tags* a fim de serem utilizadas para treinar modelos de predição. Já no método não supervisionado, treina o modelo em um *data set* não anotado para detectar padrões, após essa etapa é feito o processo de *tagging* em novos textos. Esses métodos podem ser subdivididos em três categorias: baseadas em regras, estocásticas e baseadas em transformação.

3.3.1.1 Técnica baseada em regras

As técnicas baseadas em regras utilizam normas da língua, como as definidas pelas classes gramaticais, para atribuir as *tags* corretas para as palavras nas sentenças (GARG *et al.*, 2012). Essas técnicas costumam utiliza grandes *data set* anotados manualmente para eliminar ambiguidade das palavras (KUMAR & JOSAN, 2010).

3.3.1.2 Técnica estocástica

Técnicas estocásticas fundamentam-se na probabilidade condicional para uma sequência de palavras inseridas em um mesmo contexto (ADHVARYU & BALANI, 2015). Esse sequenciamento de palavras ordenado, são organizados em estruturas chamadas *n-grams*, que são formados a partir das n palavras adjacentes a uma em destaque (ADHVARYU & BALANI, 2015). Algumas técnicas são consideradas estocásticas como a *Hidden Markov Model* (HMM), a *Maximum Entropy Markov Model* (MEMM) e a *Conditional Random Fields* (CRFs).

3.3.1.3 Técnica baseada em transformação

As técnicas baseadas em transformações consiste em encontrar melhorias a cada iteração do algoritmo. No inicio do algoritmo, cada palavra recebe uma *tag* temporária, então a cada iteração aplica regras de transformações às palavras para obter um melhor *score* de *tagging* (ADHVARYU & BALANI, 2015). Essas transformações são extraídas de um *data set* anotado. As iterações param quando o *score* atinge um valor menor que o limiar pré-definido ADHVARYU & BALANI (2015).

3.4 Embeddings

Embeddings são representações numéricas no espaço continuo que contém informações sobre os dados (palavras, imagens, categorias, tempo e outros) que corres-

pondem (GLOBERSON *et al.*, 2007). A partir de funções ou de métodos, os dados são mapeados para o espaço continuo R^n de modo que possam ser compreendidos por outras aplicações (GLOBERSON *et al.*, 2007).

As *embeddings* aplicada à área de Processamento de Linguagem Natural (PLN), são conhecidas como *language model*, ou *text representation*, ou *distributed representation*, que tem por objetivo capturar informações léxica e/ou sintática e/ou semântica dentro de um contexto (LI & YANG, 2018). Ainda no âmbito da PLN, as *embedings* são utilizadas para representar palavras, sentenças, parágrafos e até documentos (LE & MIKOLOV, 2014; LI & YANG, 2018).

Em PLN existem tarefas que consistem em gerar *embeddings*, essas por sua vez utilizam alguns métodos para as criar, como os a seguir: *one-hot model* (TURIAN *et al.*, 2010); *vector space model* (SALTON *et al.*, 1975); *co-ocorrence model* (LUND & BURGESS, 1996); *neural network language model* (BENGIO *et al.*, 2003). Existem outras técnicas além das citadas nesse parágrafo.

3.4.1 *Word Embedding*

Segundo LI & YANG (2018), *word embedding* é o conjunto de métodos de modelo de linguagem ou métodos de seleção de características, onde a sua principal função é mapear palavras ou frases para o espaço contínuo de baixa dimensionalidade (GLOBERSON *et al.*, 2007). LAI *et al.* (2016) afirma que o *Word embedding* é capaz de capturar informações sintáticas e semânticas de uma palavra, onde a informação semântica esta relacionada com o significado da palavra, e a informação sintática refere as suas regras estruturais.

De acordo HINTON *et al.* (1986), que as palavras com conceitos parecidos são propensas a apresentarem comportamento semelhante no espaço continuo, ou seja, os *word embeddings* dessas palavras tendem a estarem próximos no espaço continuo R^n. SUN *et al.* (2015) conta que as palavras encontradas em contextos semelhantes são propicias a apresentarem significados parecidos.

Segundo TURIAN *et al.* (2010), os *word embeddings* são comumente vetores, onde cada dimensão corresponde a uma característica da palavra mapeada podendo ser informação semântica ou informação sintática. Essas informações para serem capturadas precisam de contextos associados a elas, para isso, SUN *et al.* (2015) diz que essas relações podem ser capturadas de duas maneiras: paradigmático e sintagmático.

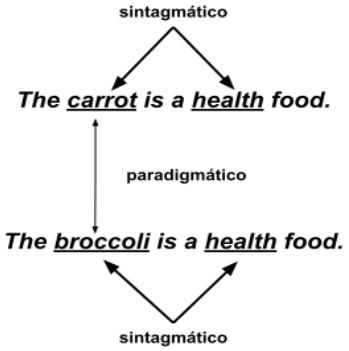


Figura 3.4 – Demonstração de paradigmático e sintagmático. Reprodução baseada em SUN *et al.* (2015)

As relações sintagmáticas são expressas nas palavras que co-ocorrem na mesma região do texto (SUN *et al.*, 2015). Na figura 3.4, as palavras *carrot* e *health* apresentam essa relação podendo estarem próximas quando forem mapeadas para o espaço continuo, desde que ocorra com frequência em textos.

As relações paradigmáticas são palavras que ocorrem em contextos semelhantes mas não no mesmo texto (SUN *et al.*, 2015). Na figura 3.4 essa relação é exposta nas palavras *carrot* e *broccoli*.

3.4.2 Modelo de linguagem de rede neural

BENGIO *et al.* (2003) propôs uma Modelo de linguagem de rede neural (*neural network language model*, NNLM) no qual é capaz de assimilar, simultaneamente, a *word embedding* para cada palavra como também a função probabilística para uma sequência de palavras. Esse modelo pode ser representado pela probabilidade condicional da próxima palavra dadas as anteriores, como é descrito na equação a seguir:

$$P(w_1^T) \;=\; \prod_{t=1}^{T} P(w_t|w_1,\; \ldots\; , w_{t-1}), \tag{3.5}$$

onde w_t é a palavra a ser prevista e $w_1,\; \ldots\; , w_{t-1}$ representa a sequência das probabilidades das palavras anteriores.

BENGIO *et al.* (2003) leva em consideração a ordem das palavras, pois as palavras mais próximas são estatisticamente mais dependentes, podendo diminuir a dificuldade desse modelo por meio da redução da quantidade de combinações possíveis quando a sequência de palavras é grande. A fim de reduzir esse número de combinações, BENGIO *et al.* (2003) utiliza *n-gram model*[4], o qual auxilia na construção da tabela de probabilidade condicional da próxima palavra; então até $n-1$

[4]n-gram model é uma sequências de palavras contíguas dada uma amostra do texto (BROWN *et al.*, 1992)

palavras serão combinadas como informação do contexto. Com essa nova abordagem, a equação passa a ser:

$$P(w_1^t|w_1^{t-1}) \approx P(w_1^t|w_{t-n+1}^{t-1}), \tag{3.6}$$

onde n é o número de palavras a serem consideradas no cálculo da probabilidade condicional.

A NNLM é considerada uma abordagem não-supervisionada, e LI & YANG (2018) afirmam que as seguintes ténicas também são: *Restrict Boltzmann Machine* (RBM), *Convolutional Neural Network* (CNN), *Long-Short Term Memory* (LSTM), *Log-Bilinear* (LBL) e *Recursive Autoenconder* (RAE).

3.4.3 *Recursive Autoencoders*

Segundo SOCHER *et al.* (2011a), *Recursive Autoencoder* (RAE) tem por meta encontrar representações vetoriais para textos de tamanhos variados os quais estão incorporados em uma *parse tree*. Ainda SOCHER *et al.* (2011b), o RAE é capaz de explorar a estrutura hierárquica contida na *parse tree* e assim capturar informações intrínsecas ao texto da qual está sendo gerada a representação. Após geradas essas representações vetoriais, elas são utilizadas para outras tarefas de PLN, como: análise de sentimentos, tradução, sumarização e detecção de plágio de paráfrase (SOCHER *et al.*, 2011a).

3.4.3.1 *Recursive Autoencoder*

Como dito na seção 3.4.3, para a execução do RAE é necessário como entrada uma *parse tree* do texto e que ela tenha as características de uma árvore binária de Chomsky, além disso, o RAE precisa da lista das *word embeddings* $x = (x_1, ..., x_m)$ das palavras contidas no texto (SOCHER *et al.*, 2011a). O formato da entrada para o RAE é uma tupla composta por um nó pai com seus dois nós filhos: $(p \rightarrow f_1 f_2)$, juntamente com o vetor de representação dos nós filho podendo ser uma folha ou o nó não terminal (SOCHER *et al.*, 2011a). Como demonstra a figura 3.5, as tuplas de entrada são: $((y_1 \rightarrow x_2 x_3), (y_2 \rightarrow x_1 y_1)), \forall x, y \in \mathbb{R}^n$ (SOCHER *et al.*, 2011a).

Com a estrutura em árvore binária, os nós filho são utilizados para gerar a representação para os nós pai (SOCHER *et al.*, 2011a), iniciando pela primeira tupla, $((y_1 \rightarrow x_2 x_3)$, onde $p = y_1$, a representação do pai é calculada a partir de seus filhos $(f_1, f_2) = (x_2, x_3)$ por uma camada padrão de uma *neural network*:

$$p = func(W_e[f_1; f_2] + b), \tag{3.7}$$

onde $[f_1; f_2]$ é a concatenação das *word embeddings* dos nós filho, $func$ é uma função de ativação tal como *tahn* e $W_e \in \mathbb{R}^{n \times 2n}$ é o *encoding* que deseja-se aprender (SOCHER *et al.*, 2011a). De acordo com SOCHER *et al.* (2011a), para verificar a qualidade do vetor de representação do nó pai, uma verificação pode ser feita por meio do *decoding* do mesmo para uma camada de reconstrução dos nós filhos e calcular a distância Euclidiana entre os os nós reconstruídos e os nós originais, essa distância representa o erro de reconstrução que é expresso na equação a seguir:

$$[f_1'; f_2'] = func(W_d p + b_d) \qquad E_{rec}(p) = ||[f_1; f_2] - [f_1'; f_2']||^2. \tag{3.8}$$

Os passos são repetidos recursivamente, logo as mesmas ações feitas anteriormente são aplicadas à próxima tupla (SOCHER *et al.*, 2011a). Com a representação vetorial para y_1, a equação 3.7 é aplicada para encontrar um vetor de representação para o nó pai y_2, então os nós filhos passam a ser: $(f_1, f_2) = (x_1, y_1)$ (SOCHER *et al.*, 2013). Após calculado o vetor de representação para o nó pai $p = y_2$, é verificado a qualidade do seu vetor por meio da equação 3.8. Esse processo repete até que a árvore esteja construída e com todos os seus nós associados à um error de reconstrução (SOCHER *et al.*, 2011a).

SOCHER *et al.* (2011a) declara que durante o treino o foco principal é minimizar o erro de reconstrução de todos pares de entrada para os nós não terminais p da *parse tree* τ:

$$E_{rec}(\tau) = \sum_{p \in \tau} E_{rec}(p), \tag{3.9}$$

como é apresentado na figura 3.5, durante a execução, há a tentativa de minimizar $E_{rec}(\tau) = E_{rec}(y_1) + E_{rec}(y_2)$.

3.4.3.2 *Unfolding Recursive Autoencoder*

O processo de *encoding* do *Unfolding Recursive Autoencoder* funciona da mesma maneira que o processo de *encoding* do RAE, onde a principal diferença entre os dois é a tarefa de *decoding*, a qual o *Unfolding* RAE tenta reconstruir toda sub-árvore abaixo do nó o qual a sua representação esta sendo construída, como apresenta a figura 3.6 (SOCHER *et al.*, 2011a). Agora, no nó $p = y_2$, o cálculo do erro de reconstrução levará em consideração todos os nós folha $[x_1, x_2, x_3]$ e as suas representações reconstruídas $[x_1', x_2', x_3']$ (SOCHER *et al.*, 2011a). Então, recursivamente, reconstruirá a sub-árvore começando pelo nó y_2:

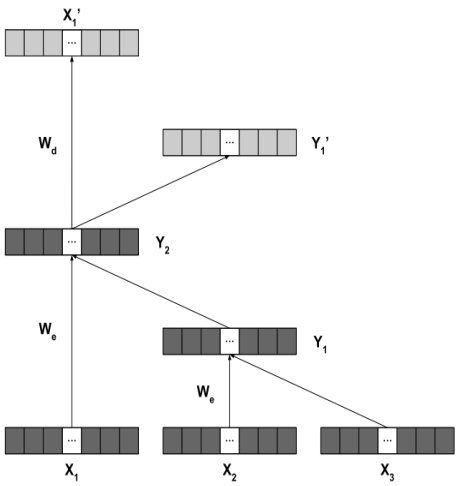


Figura 3.5 – Instância de um *Recursive Autoencoder*. Reprodução baseada em SOCHER *et al.* (2011a)

$$[x_1'; y_1'] = func(W_d y_2 + b_d), \tag{3.10}$$

depois o nó y_1':

$$[x_2'; x_3'] = func(W_d y_1' + b_d). \tag{3.11}$$

O erro de reconstrução é calculado a partir da concatenação das representações que estão abaixo do nó em questão. O nó y que abrange i até j palavras, teria o seu erro calculado da seguinte forma:

$$E_{rec}(p) = ||[x_i; ...; x_j] - [x_i'; ...; x_j']||^2. \tag{3.12}$$

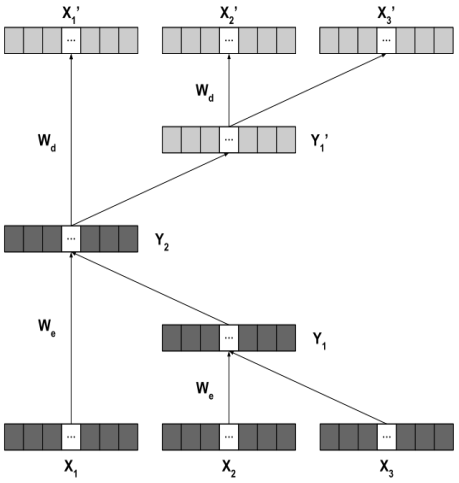


Figura 3.6 – Instância de um *Unfolding Recursive Autoencoder*. Reprodução baseada em SOCHER *et al.* (2011a)

Segundo SOCHER *et al.* (2011a), o *Unfolding* RAE tenta encontrar a melhor representação dos nós quando considera a reconstrução das sub-árvores abaixo deles. Esse comportamento permite capturar o aumento da importância do nó filho quando o mesmo representa uma sub-árvore grande (SOCHER *et al.*, 2011a).

3.4.3.3 Exemplo prático de uso do RAE

Têm as seguintes sentenças as quais deseja-se detectar plágio de paráfrase (demonstração baseado em SOCHER *et al.* (2011a)):

1. *The cats catch mice.*
2. *Cats eat mice.*

As sentenças 1 e sentença 2 são submetidas pelo processo de *POS-Tagging* gerando suas árvores sintáticas, essas árvores por sua vez são "binarizadas"[5]. Após esse pré-processamento as árvores resultantes estão descritas na figura 3.7, onde os nós não-terminas representam as relações sintáticas entre as palavras das sentenças.

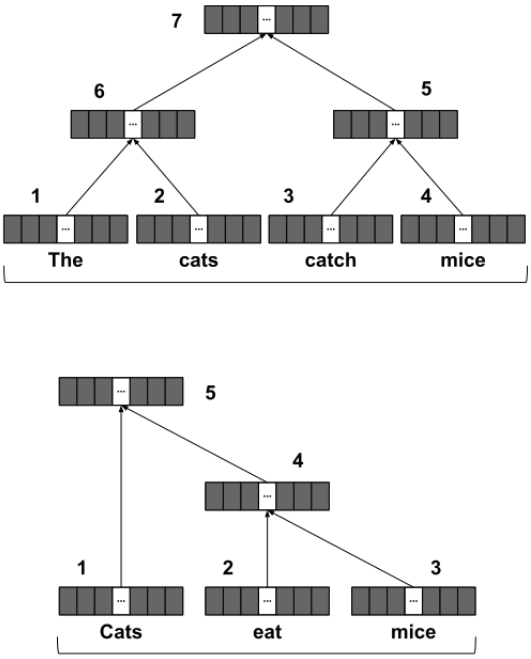


Figura 3.7 – Árvores binarizadas. A árvore superior é referente a sentença 1, e a árvore inferior é referente a sentença 2. Reprodução baseada em (SOCHER *et al.*, 2011a)

A fim de comparar as sentenças e verificar se há plágio de paráfrase, as árvores binárias passam pelo RAE para gerar representações aos nós não-terminais, desse modo, cada nó das árvores poderão ser comparados uns com outros originando uma

[5]são transformadas para a atender as regras da Forma Normal de Chomsky https://www.nltk.org/_modules/nltk/treetransforms.html

matriz de similaridade A, onde $A \in \mathbb{R}^{(2n-1)\times(2m-1)}$, n e m são os nós da sentença 1 e sentença 2 respectivamente, como demonstra a figura 3.8. Cada célula dessa matriz representa a distância euclideana entre os nós das duas árvores. A similaridade é definida pela distância, ou seja, quanto menor for a distância entre os nós mais semelhantes eles são, analogamente para maior distância indicando que os nós são não tão semelhantes. Observando a célula $a_{2,1}$, onde $a \in A$, percebe-se que a mesma apresenta um alto grau de similaridade, isso por causa do fato de os nós em comparação representarem a mesma palavra (*cats*). Em contrapartida, a célula $a_{7,5}$ revela que não há similaridade entre os nós, esse caso ocorre devido os nós destoarem nas informações que as suas sub-árvores transmitiram a eles (SOCHER *et al.*, 2011a).

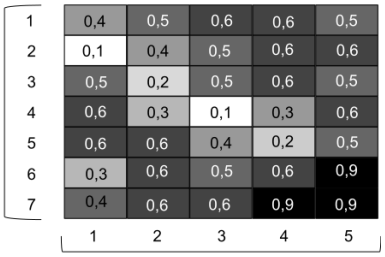


Figura 3.8 – Matriz de distância A das sentenças 1 e 2. Reprodução baseada em (SOCHER *et al.*, 2011a)

Por conta da variação dos tamanhos dos textos comparados, as matrizes de similaridade tendem a terem tamanhos heterogêneos e esse fato torna-se um problema quando considera-se usar classificadores que trabalham apenas com conjuntos de dados com vetores de característica de mesmo tamanho SOCHER *et al.* (2011a). A fim de lidar com essa dificuldade, SOCHER *et al.* (2011a) propôs o *dynamic pooling*.

Considerando a matriz similaridade A, a função do *dynamic pooling* é mapea-lá para uma matriz A_{pooled} de tamanho fixo $n_p \times n_p$, desde que as dimensões de A sejam divisíveis por n_p, assim permitindo fazer o janelamento da matriz SOCHER *et al.* (2011a). As dimensões da janela para o *pooling*: $w \in \mathbb{R}^{r\times c}$, onde $r = (2n-1)/n_p$ e $c = (2m-1)/n_p$. Quando a divisão não é exata, sobrando linhas ou colunas, essas linhas e colunas extras são distribuídas uniformemente para os janelamentos finais da matriz (SOCHER *et al.*, 2011a). Conforme é feito o janelamento, os valores dentro da janela, um ou mais são escolhidos para caracterizar aquela região na matriz *pooled*. O critério de seleção varia de acordo com a finalidade o qual está sendo feito, nesse caso que esta sendo demonstrado, a condição é escolher o valor mínimo dentro da região a qual janela esta localizada, conforme a figura 3.9 retrata.

Segundo SOCHER *et al.* (2011a), a matriz A_{pooled} perde algumas informações contida na original mas ainda consegue capturar informações gerais da estrutura

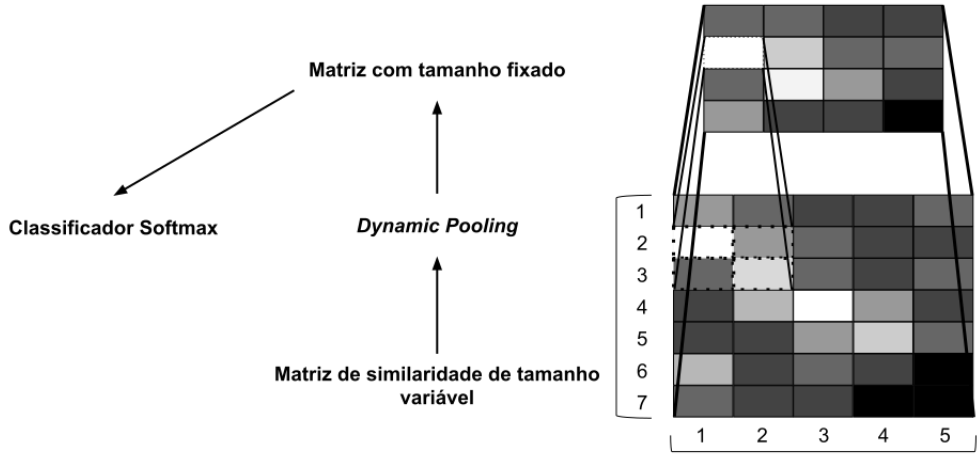


Figura 3.9 – Matriz A_{pooled} da matriz A. Reprodução baseada em (SOCHER *et al.*, 2011a)

como também informações semânticas que permitem detectar plágio de paráfrase. Após todo esse processamento a matriz A_{pooled} é submetida ao classificador.

3.5 Capacidade de representação das características do texto

Retornando a discussão mencionada no início desse capitulo, sobre a necessidade de retratar todas características importantes para detecção de plágio de paráfrase, a tabela 3.1 apresenta a abrangência das representações exibidas neste capítulo. Dentre elas, o RST foi o que demonstrou maior possibilidade de capturar todas características, porque durante a construção da sua estrutura leva em consideração as relações léxicas e gramaticais para definir as interações entre as partes do texto, essas interações por sua vez contribuem para esclarecer as intenções do autor diante do leitor que contribuem para compreender a semântica do texto, e por fim todas essas informações estão organizadas hierarquicamente em uma estrutura em árvore ((MANN & THOMPSON, 1987, 1988; TABOADA & MANN, 2006)).

O POS-Tagging, limita-se a retratar sentenças, não se adequando a representar um documento por completo, pois a sua estrutura representativa consegue expressar informações léxicas e sintáticas podendo fazer inferência sobre a característica semântica. O *word embedding* foca representar os possíveis significados e posicionamentos da palavra dentro de um contexto.

O RAE gera a representação da estrutura, ou seja, dado um conjunto de nós pertencentes à árvore binarizada, para cada nó dessa árvore o RAE cria uma representação adequada permitindo a comparação deles. Logo, a representação gerada pelo RAE dependente das informações que a estrutura de dados esta carregando,

por exemplo: caso seja uma estrutura oriundo do processo de POS-Tagging, só poderá representar as informações que essa estrutura carrega; caso seja originária do RST, representará as informações que o RST é capaz de oferecer. Por isso, a sua marcação na tabela 3.1 está entre parênteses.

	Léxica	Sintática	Semântica	Estrutura
RST	X	X	X	X
POS-Tagging	X	X		
Word Embedding	X		X	
RAE	(X)	(X)	(X)	(X)

Tabela 3.1 – Abrangência das representação sobre as características do texto

Capítulo 4

Trabalhos Relacionados

Neste capítulo são apresentadas as técnicas aplicadas à detecção de plágio e também à detecção de plágio de paráfrase. Algumas das técnicas descritas nesse capítulo tiveram maior detalhamento por algumas razões como: a abordagem descrita na seção 4.3 foi escolhido por ser o atual estado da arte na detecção de paráfrase no *data set* da *Microsoft Research Paraphrase Corpus* (MSRPC) (DOLAN *et al.*, 2004); o método detalhado na seção 4.1 por ser umas das abordagem que já foi aplicado ao *data set* utilizado nos experimentos dessa dissertação; a técnica descrita na seção 4.2 para corroborar com a proposta apresentada no capítulo 5. Os demais métodos foram retratados de forma superficial na seção 4.4.

4.1 Alinhamento de texto

O alinhamento de texto é uma técnica na qual, a partir de um par de documentos, um suspeito e outro fonte, tem por finalidade identificar todas as passagens contiguas que são compartilhadas entre eles (POTTHAST *et al.*, 2013). Esse é um algoritmo de alinhamento de sequência que funciona da seguinte maneira: primeiro, a lista de *n-grams* de palavras, que são iguais ou correspondes entre os dois documentos, são extraídas como sementes; segundo, essas sementes são unidas gradativamente obedecendo um conjunto de regras de união pré-estabelecidas, que decidem se as sementes devem ser unidas ou não, criando passagens alinhadas; terceiro, a coleção de passagens alinhas é retornada para a análise de um especialista (POTTHAST *et al.*, 2013).

O alinhamento de texto era uma técnica bastante utilizado na competição de detecção de plágio externo que ocorria durante a *Conference on Multilingual and Multimodal Information Access Evaluation* (POTTHAST *et al.*, 2013). Dentre as diversas abordagens utilizadas nessa conferência, a apresentada por SANCHEZ-PEREZ *et al.* (2014) foi a que apresentou desempenho mais consistente no ano de 2014 e em comparação com anos anteriores (POTTHAST *et al.*, 2014). Como a

maioria das abordagens, SANCHEZ-PEREZ *et al.* (2014) faz uso de quatro técnicas comuns:

- semeadura (*seeding*): dado um documento fonte e um documento suspeito, essa tarefa consiste em separar um conjunto de pequenos candidatos a caso de plágio, os quais são chamados de sementes. Cada candidato é um par constituído por um pequeno fragmento do documento suspeito e um pequeno fragmento do documento fonte, os quais apresentam alguma similaridade entre eles. SANCHEZ-PEREZ *et al.* (2014) verificava a similaridade entre os fragmentos por meio de vetores extraídos da matriz *tf-idf*[1].

- extensão (*extension*): a partir do conjunto dos pares de fragmentos similares encontrados na tarefa anterior, esse próximo passo tem por finalidade construir os maiores fragmentos similares possível entre os documentos (POTTHAST *et al.*, 2014). A fim de atingir esse objetivo, utilizando o conjunto de sementes, os fragmentos do documento suspeito são unidos progressivamente e essa mesma ação é aplicada aos fragmentos do documento fonte. Após a união eles são verificados se continuam similares, caso continuem, uma nova tentativa é feita para expandir esse fragmento contiguo. SANCHEZ-PEREZ *et al.* (2014) faz uso de variáveis e abordagens para continuar a tentar juntar os fragmentos dos documentos, essas variáveis e abordagens são para lidar com partes dos fragmentos que são possíveis sequências de palavras que geram divergência de similaridade entre os trechos comparados dos documentos.

- filtragem (*filtering*): no passo de filtragem, as sobreposições são removidas assim como também, os já unidos, fragmentos muito pequenos (SANCHEZ-PEREZ *et al.*, 2014).

- comportamento adaptável: SANCHEZ-PEREZ *et al.* (2014) para cada *data set* da competição, o algoritmo tinha dois ou mais conjuntos de parâmetros os quais, após o pré-processamento e o passo de semeadura, é feito duas execuções com alguns elementos da coleção de pares de fragmentos a fim de decidir quais dos conjuntos de parâmetros era o melhor para ser aplicado à coleção.

Após todos esses processos, o conjunto contendo os casos de possível plágio é retornado a fim de que o especialista possa analisar e decidir se há plágio entre os documentos ou não.

[1]em recuperação de informação, é a abreviação da expressão *term frequency-inverse document frequency*, em português frequência do termo-inverso da frequência nos documentos, que é uma medida para representar a importância da palavra em um documento

A abordagem proposta por SANCHEZ-PEREZ *et al.* (2014) tem o seu foco na captura das características léxicas e na ordem da sequência das palavras, apresentando uma deficiência em considerar as demais características do texto que também são importantes para detecção de plágio em documentos.

4.2 Utilização da estrutura do texto

Uma outra abordagem para identificação de plágio é a apresentada por CHOW & RAHMAN (2009), na qual consiste na utilização da estrutura em árvore para representar a organização do documento html em níveis, combinada-a com uma rede neural para gerar um vetor de características para cada nó dessa árvore. CHOW & RAHMAN (2009) usa a estrutura em árvore para representar três níveis dos documentos html: sendo o primeiro nível para representar os parágrafos, onde esses são as folhas da árvore que têm o tamanho limitado por um limiar pré-definido; o segundo nível representa as páginas, nós não terminais que são a junção dos blocos (parágrafos) também respeitando um limite pré-estabelecido por página; e por fim, o terceiro nível representa todo o documento sendo composto por apenas um nó.

Para gerar a representação das características do documento, CHOW & RAHMAN (2009) utiliza a rede neural conhecida como *MultiLayer Self-Organization Map* (MLSOM). O principal objetivo dessa rede neural é gerar neurônios de representação de característica para cada nó da árvore de acordo com os seus níveis, ou seja, a quantidade de camadas em MLSOM acompanha a quantidade de níveis que existem na árvore, dessa forma, os neurônios criados carregam informações das características dos nós que representam como também as informações dos seus filhos, caso tenham, e, dentro do contexto inseridos, informações dos níveis (CHOW & RAHMAN, 2009). Segundo CHOW & RAHMAN (2009), as características contidas nos parágrafos podem ser mapeadas até um último nível da árvore, o qual contém informações sobre toda a estrutura do documento html.

Na momento da identificação de plágio, é utilizada a técnica de detecção por localidade, onde todos os neurônios que representam os parágrafos dos documentos fontes são *clusterizados*, comparados e analisados, que por sua vez são comparados com os neurônios que representam os parágrafos do documento suspeito para assim identificar quais parágrafos dos documentos fontes foram plagiados. Uma limitação dessa abordagem para detecção de plágio de paráfrase em documentos é a sua extrema dependência dos marcadores html para identificar os parágrafos. Outra possível deficiência é o nível de detalhamento, o qual não aparenta capturar as relações estabelecidas entres as palavras dentro do parágrafo perdendo informações léxicas e possíveis informações sintáticas.

4.3 TF-KLD mapeado no espaço latente

A abordagem de JI & EISENSTEIN (2013) foi aplicado ao *data set* MSRPC, que é composto por casos de plágio de paráfrase (DOLAN *et al.*, 2004). A fim de detectar paráfrase nesse *data set*, JI & EISENSTEIN (2013) utiliza a similaridade no espaço latente no qual afirma que as informações semânticas estão relacionadas. Para isso, utiliza *Term Frequency Kullback-Leibler Divergence* (TF-KLD) para substituir o TF-IDF para ajustar os pesos da matriz de características e a combina com a fatorização de matrizes, com a adição de características granulares e com o classificador SVM.

Dada a matriz de termo-contexto ou matriz instância-características A, o TF-KLD é a forma de ajustar os pesos dessa matriz, que contém as representações das instâncias (sentenças, parágrafos, documentos), o qual tem por objetivo aumentar a expressão das características que descrevem melhor a instância e diminuir as que não descrevem. Assumindo o par de instâncias $(\vec{a}_i^{(1)}, \vec{a}_i^{(2)}, r_i)$, onde $\vec{a}_i^{(1)}$ é o vetor de características para a primeira instância, $\vec{a}_i^{(2)}$ é o vetor de caraterísticas para a segunda instância, e $r_i \in 0, 1$, que indica se há paráfrase ou não entre o par. JI & EISENSTEIN (2013) define duas distribuição de Bernoulli:

- $p_k = P(a_{ik}^{(1)} | a_{ik}^{(2)} = 1, r_i = 1)$. Essa equação descreve a probabilidade da instância $a_{ik}^{(1)}$ conter a característica k, dado que k aparece em $a_{ik}^{(2)}$ e as instâncias marcadas como paráfrase, $r_i = 1$.

- $q_k = P(a_{ik}^{(1)} | a_{ik}^{(2)} = 1, r_i = 0)$. Essa equação descreve a probabilidade da instância $a_{ik}^{(1)}$ conter a característica k, dado que k aparece em $a_{ik}^{(2)}$ e as instâncias marcadas como não paráfrase, $r_i = 0$.

A partir da divergência de Kullback-Leibler, tem-se $KL(p_k||q_k) = \sum_x p_k(x) \log \frac{p_k(x)}{q_k(x)}$ que é a discriminabilidade da característica k nas duas instâncias. Essa divergência é utilizada para ajustar os pesos da matriz A antes do processo de fatorização, onde o efeito de aumento do valor da característica na matriz é manifesto quando a probabilidade da sua ocorrência no par de instância está extremamente influenciada pela relação de paráfrase entre elas. Por outro lado, o efeito de diminuição do valor da característica na matriz A acontece quando a probabilidade de p_k e q_k são quase iguais fazendo com que a divergência KL tenda à zero.

Após o TF-KLD, a fatorização é aplicada à matriz A, sendo utilizados os métodos Decomposição de Valores Singulares ou Matrizes Não-Negativas. Em seguida, com a representação latente oriunda da fatorização, os vetores que representam o par das

instâncias $(\vec{v}_1, \vec{v}_1)$ são transformados da seguinte forma:

$$\vec{f}(\vec{v}_1, \vec{v}_2) = [\vec{v}_1 + \vec{v}_2; |\vec{v}_1 - \vec{v}_2|] \tag{4.1}$$

onde $\vec{f}(\vec{v}_1, \vec{v}_2)$ é a concatenação da soma de $\vec{v}_1 + \vec{v}_2$ e o valor absoluto de $|\vec{v}_1 - \vec{v}_2|$. Ao fim da concatenação da operações no vetores, as dez primeiras características granulares descritas por WAN *et al.* (2006) são concatenadas ao vetor resultante e por fim submetido ao SVM.

Por ser uma abordagem voltada para detectar paráfrase em sentenças, as características estruturais do documento, os quais remetem aos diversos contextos inseridos no documento (CHOW & RAHMAN, 2009), ou a relação entre a suas partes texto não são capturadas deixando de assimilar as informações que são úteis para detectar paráfrase entre documentos (ALZAHRANI *et al.*, 2011).

4.4 Outras abordagens para detecção de plágio de paráfrase

Outras abordagens utilizadas para detecção de paráfrase somente em sentenças no *data set Microsoft Research Paraphrase Corpus* (DOLAN *et al.*, 2004), são: já explicada na seção 3.4.3.3, é a proposta por SOCHER *et al.* (2011a); CHENG & KARTSAKLIS (2015) propõem uma arquitetura de construção unificada do modelo composicional e do conjunto de *word embedding* de modo a permitir a indução dinâmica do sentido de cada palavra durante o processo de aprendizagem, a fim de não somente predizer a ocorrência de uma palavra dentro do contexto como também a sua posição; FILICE *et al.* (2015) utiliza estruturas de representação do texto, como árvore sintática e grafo, para as tarefas de *Text Entailments* e *Question Answering*, além da detecção de paráfrase; HE *et al.* (2015) propõe um método, utilizando rede neural convolucional, a qual cria representações para as sentenças que permitam comparações em múltiplas perspectivas; WANG *et al.* (2016) demonstra uma forma que leva em consideração a similaridade como também a dissimilaridade através da decomposição e da composição da semântica lexical das sentenças; por fim, MADNANI *et al.* (2012) redireciona a aplicação das técnicas de *machine translation* para detecção de paráfrase.

Outra forma para detectar plágio, semelhante a relatada na seção de alinhamento de texto, é abordagem descrita por SU *et al.* (2008) que usa a distância de Levenshtein e uma versão simplificada do algoritmo de Smith-Waterman para identificar plágio.

Capítulo 5

Identificação de Plágio de Paráfrase com Representações Estruturais do Documento

A análise detalhada entre pares de documentos para detecção de paráfrase é complexa pois apenas comparar caractere à caractere ou apenas em nível de sentenças pode não ser a forma mais eficaz para a tarefa, pois a paráfrase enquadra-se na classe de plágio inteligente, que vai desde manipulação de texto até a adoção de ideias (ALZAHRANI *et al.*, 2011).

O capítulo 4 apresentou um conjunto de técnicas que detectam plágio de paráfrase em nível de sentença não sendo adequadas para serem aplicadas em documentos; outro conjunto identifica plágio em documentos utilizando apenas as características léxicas do texto não sendo eficaz no reconhecimento de casos de plágio de paráfrase mais complexo, o qual vai além das características léxica; uma técnica usa a característica estrutural dependente de linguagem de marcação para sua construção, não sendo aplicável apenas a texto. Todos essas abordagem apresentam algum tipo de limitação que as impedem de detectar plágio de paráfrase em um documento inteiro levando em consideração a sua característica léxica, sintática, semântica e estrutural. Por conta desse fato, torna-se necessário transcrever os documentos em uma representação capaz de assimilar o seu conteúdo intrínseco permitindo contrapô-los e avaliar se compartilham as mesmas informações que possam ser classificadas como possível caso de plágio de paráfrase.

O objetivo dos métodos propostos neste capítulo é representar o documento contemplando as suas características léxica, sintática, semântica e estrutural, a fim de permitir a detecção de plágio de paráfrase quando comparado com um outro documento suspeito. A seção 5.1 mostra uma adaptação da abordagem do SOCHER *et al.* (2011a) para ser utilizada em documentos, a seção 5.2 demonstra a abor-

dagem utilizando o RST aliado à *word embedding*, RAE e *dynamic pooling* para gerar a representação do texto e a seção 5.3 apresenta o fluxo geral para geração da representação do documento e também a classificação.

Para as propostas que serão descritas nas próximas seções, considere $D = [d_1, d_2, ..., d_h]$ o conjunto de todos os documentos fonte, dessa coleção é selecionado o documento d_1 para ser comparado com o documento suspeito d_s a fim de detectar se há plágio de paráfrase.

5.1 Representação com uso do POS-Tagging

O método proposto nesta seção visa suprir as deficiências encontradas nas atuais abordagens em detectar plágio de paráfrase. A abordagem apresentado é capaz de representar todo o documento levando em consideração as quatro características intrínsecas do texto, como descreve ALZAHRANI *et al.* (2011).

A abordagem dessa seção é baseada no método proposto por SOCHER *et al.* (2011a), no qual consiste no uso de uma rede neural, chamada *Recursive Autoencoder*, que tem por objetivo aprender o vetor de características para as sentenças em uma árvore sintática, como já explicado na seção 3.4.3.3.

Dada a necessidade de detectar plágio de paráfrase em documentos e a limitação da proposta de SOCHER *et al.* (2011a) para essa tarefa, torna-se necessário adaptar o método para ser aplicável à textos com nível estrutural acima do nível de sentenças, de modo a capturar a informação organizacional do documento permitindo assimilar as variações semânticas das palavras de acordo com os diversos contextos distribuídos pelo texto.

Documentos são comumente estruturados em níveis acima do nível de sentença, como combinações de sentenças (linhas), parágrafos e o texto por inteiro e sabendo que a abordagem de SOCHER *et al.* (2011a) trabalha apenas com sentença por conta do POS-Tagging, torna-se necessário dividir os documentos neste nível de detalhamento.

Dados os documentos d_1 e d_s que são compostos por texto bruto, para criar as suas representações com a estrutura em árvore sintática, é necessário "granularizar"o texto em sentenças. Toma-se como exemplo o trecho de texto "*Meet the Announcers*"(MANN & THOMPSON, 1988), utilizado na seção 3.2.5.2:

P. M. has been with KUSC longer than any other staff member. While attending Occidental College, where he majored in philosophy, he volunteered to work at the station as a classical music announcer. That was in 1970. (5.1)

Ao dividir esse trecho em sentenças, é criado um conjunto S com três elementos como a seguir:

- s_1 = *P. M. has been with KUSC longer than any other staff member.*
- s_2 = *While attending Occidental College, where he majored in philosophy, he volunteered to work at the station as a classical music announcer.*
- s_3 = *That was in 1970.*

Os documentos d_1 e d_s ao passarem pelo processo de divisão dão origem aos seguintes conjuntos: $S_{d_1} = [s_1, s_2, ..., s_n]$, onde $S_{d_1} \subseteq d_1$; $S_{d_s} = [s_1, s_2, ..., s_m]$, onde $S_{d_s} \subseteq d_s$. Com os documentos "granularizados" em sentenças, é possível aplicar o POS-Tagging *parser* ao conjunto de sentenças em S_{d_1} e S_{d_s}, o qual o resultado desse *parser* nesse dois conjuntos são: $P_{d_1} = [p_1, p_2, ..., p_n]$, onde P_{d_1} detém todos as árvores sintáticas das sentenças de S_{d_1}; e $P_{d_s} = [p_1, p_2, ..., p_m]$, onde P_{d_s} contém todas as árvores sintáticas das sentenças em S_{d_s}.

Para exemplificar como é uma árvore sintática gerada pelo POS-Tagging, toma-se a sentença s_1 do trecho descrito anteriormente. A figura 5.1 demonstra a sentença após passar pelo POS-Tagging *parser*. As folhas são as palavras, os nós não terminais demarcam as relações sintáticas entre as folhas e S representa a raiz da árvore sintática. Vale notar que esta árvore sintática não está na Forma Normal de Chomsky.

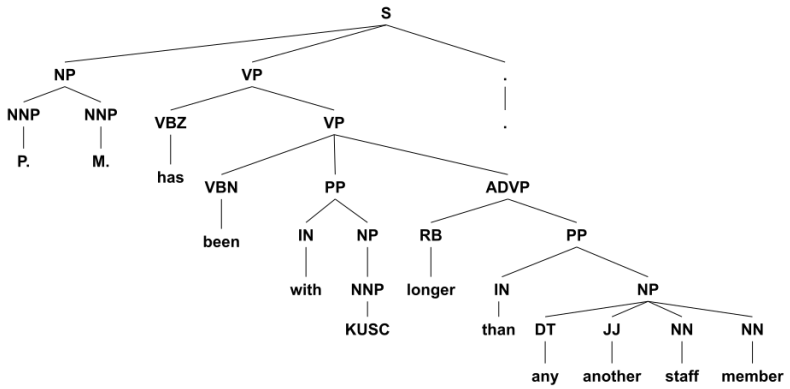


Figura 5.1 – Exemplo do POS-Tagging para a sentença "*P. M. has been with KUSC longer than any other staff member*"

Criada as árvores de relações sintática para cada sentença, agora é necessário organizá-las de modo que possam representar a característica estrutural do documento. Com isso, foi pensado em uni-las através de suas raízes obedecendo a ordem em que elas surgem no decorrer todo texto. Por exemplo:

$$b_1 = (p_1, p_2), \tag{5.2}$$

onde b_1 é a junção de p_1 e p_2, que são unidas por meio de suas raízes ligadas à uma nova raiz (b_1) em um nível acima delas. logo em sequência, a p_3 é unida pelo mesmo processo como demonstrar a seguir:

$$b_2 = (b_1, p_3), \tag{5.3}$$

onde b_2 é união de b_1 e p_3. Esse processo de junção é repetido até a árvore gerada contemplar todo o conjunto de POS-Tagging de P_{d_1} e P_{d_s} seguindo o princípio de completude descrito por MANN & THOMPSON (1988). As junções feitas entre as árvores seguem a ordem delas dentro do conjunto, ou seja, p_1 fica à esquerda de b_1 enquanto que p_2 fica à direita, e na próxima junção b_1 ficam a esquerda de b_2 enquanto que p_3 fica à direita. Digno de ressaltar, que a ordem das árvores sintáticas nos conjunto P_{d_1} e P_{d_s} seguem a ordem de aparição de suas respectivas sentenças no documento. Ao fim do processo de junção das árvores sintáticas dão origem a árvore T_{d_1} que representa todo o documento d_1 estruturado em árvore, e de semelhante modo, quando é submetida a mesma transformação, é gerada a árvore T_{d_s} que representa o documento d_s.

A figura 5.2 demonstra o processo de junção das sentenças para representar as características do texto. Os nós na cor cinza representam as folhas que são as palavras, enquanto que os nós na cor branca refletem as relações sintáticas entre as palavras das sentenças e os nós não terminais, e por fim os nós na cor preta, que são pontos de ligação entre as árvores sintáticas do conjunto de sentenças de P_{d_1}.

Quanto finalizada junção entre as sub-árvores dos documentos, é necessário "binarizar" as árvores T_{d_1} e T_{d_s}, pois, o POS-Tagging não cria as suas árvores sintáticas na Forma Normal de Chomsky. Por conta desse fato, o processo de binarização é aplicado sobre T_{d_1} e T_{d_s} gerando T'_{d_1} e T'_{d_s}, respectivamente. A binarização da árvore sintática consiste em duas tarefas: compressão unária (*collapse unary*) e ajustar a árvore para adequar-se as regras que são requeridas pela Forma Normal de Chomsky.

Veja a figura 5.3 que demonstra os pontos da árvore que precisam ser ajustados para a mesma está apta ao RAE. Os retângulos tracejados identificam as regiões da árvore que precisam sofrer a compressão unário. Já as elipses tracejadas, marcam os pontos que necessitam ser reajustados para se adequarem a Forma Normal de Chomsky.

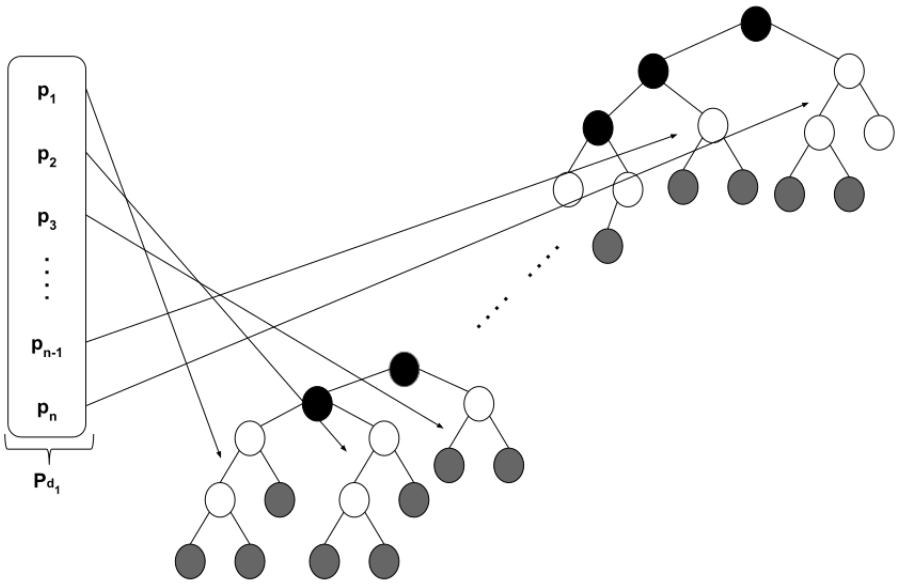


Figura 5.2 – O processo de junção das árvores sintáticas de P_{d_1}

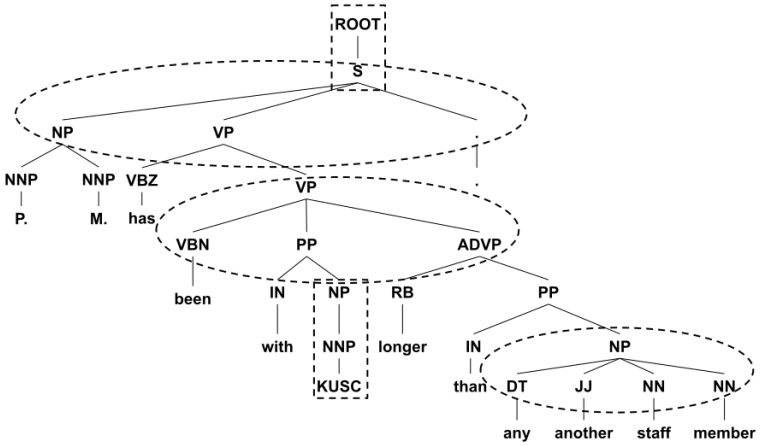


Figura 5.3 – A árvore sintática da sentença "*P. M. has been with KUSC longer than any other staff member*" com as regiões que precisam de ajuste demarcadas

A compressão unária identifica o nó não terminal que tenha apenas um filho o junto com o seu nó pai criando um novo nó. A figura 5.3 com os retângulos tracejados demonstram esses nós que serão comprimidos. A figura 5.4 apresenta a árvore após a compressão unária.

A gramática livre de contexto normalmente utilizada em formalizações sintática de linguagens de programação, onde as suas produções são, geralmente, representadas em árvores de derivações para a criação de um reconhecedor automático de linguagem (HOPCROFT, 2006). Como nesse trabalho esta sendo utilizado a estrutura em árvores e o requerimento do RAE exige uma árvore binária como insumo para poder gerar representações para os nós não terminais a partir do seus nós filhos, as regras da Forma Normal de Chomsky (FNC) é adequada para ajustar a estrutura de acordo com a demanda, pois a relação nó pai com dois nós filhos remete

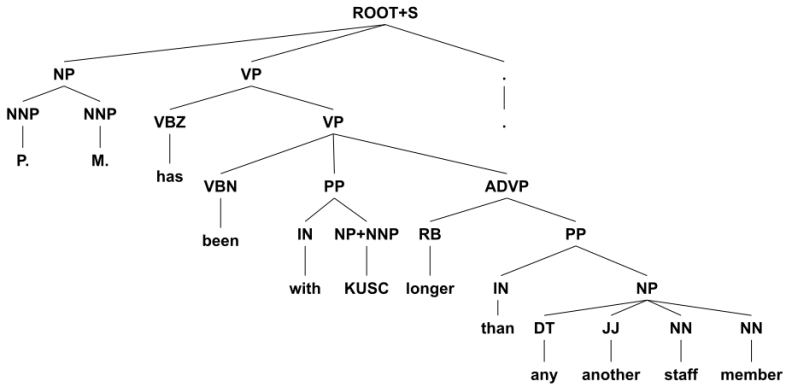


Figura 5.4 – A árvore sintática da sentença "*P. M. has been with KUSC longer than any other staff member*" após a compressão unária

a teoria proposta pela FNC para uma linguagem, o qual propõe que cada produção tenha duas produções não terminais ou seja produção terminal (HOPCROFT, 2006). Aplicando a FNC na árvore da sentença "*P. M. has been with KUSC longer than any other staff member*", ela assume a organizaçãoS como apresenta a figura 5.5.

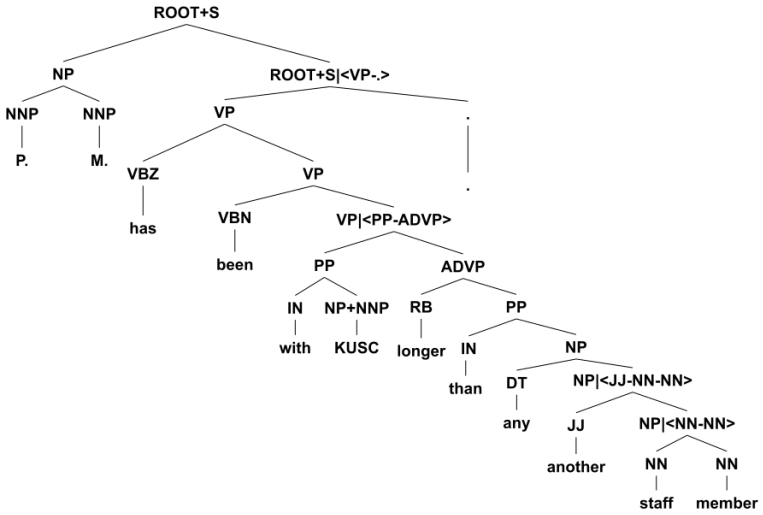


Figura 5.5 – A árvore sintática da sentença "*P. M. has been with KUSC longer than any other staff member*" após ser ajustada para atender as regras da FNC

O tamanho l das árvores T'_{d_1} e T'_{d_s} é definido pela quantidade de nós que pertencem a cada árvore. A equação a seguir demonstra como chegar ao valor de l:

$$l = 2w - 1, \tag{5.4}$$

onde: w é igual ao número de folhas (palavras) da árvore T'_d.

Com a criação das árvores T'_{d_1} e T'_{d_s}, agora é necessário capturar as informações que essas representações de características oferece e avaliar se é possível executar a tarefa de detecção de plágio de paráfrase. Para isso, cada palavra das árvores dos

documentos será substituída por uma *word embedding* a fim de auxiliar na produção dos *embeddings* para os nós não terminais como também para a raiz. Essa geração de *embeddings* é feita através do *Recursivo Autoencoder*, o qual tem como entrada para o seu processamento a árvore "binarizada". Com os nós das árvores com seus respectivos *embeddings*, gera-se a matriz de similaridade $A \in \mathbb{R}^{z \times y}$, onde $z = l_{T'_{d_1}}$ e $y = l_{T'_{d_s}}$, que compara nó-a-nó das árvores T'_{d_1} e T'_{d_s}. A matriz A é preenchida pela distância entre os nós em comparação. Em seguida, é aplicada o *dynamic pooling* para fixar o seu tamanho para servir como insumo para o classificador de detecção de plágio de paráfrase.

A proposta apresentada nessa seção consegue capturar as características léxica, sintática, semântica e, com adaptação, estrutural do documento. No entanto, a forma como é construída a representação da estrutura do texto dependendo apenas da ordem de aparição das sentenças no documento, não levando em consideração as relações que possam haver entre elas, pode não ser suficiente para auxiliar na assimilação das características intrínsecas do texto.

5.2 Representação com uso RST

Dadas as limitações das atuais abordagens em conseguir assimilar todas as características inerente ao documento e a fim de detectar plágio de paráfrase, é proposto uma abordagem capaz de representar o documento desde do nível mais detalhado (palavras) ao nível mais geral (texto inteiro). O RST é um método descritivo focado na organização do texto, oferecendo um meio genérico para a descrição das relações entre as partes do texto, mais as *word embeddings*, RAE e o *dynamic pooling*, deseja-se detectar plágio de paráfrase em documentos.

A detecção de plágio de inteligente, o qual o plágio de paráfrase esta incluso, faz uso de manipulação de texto, sumarização, uso da estrutura argumentativa, tradução e entre outras, é a classe mais difícil de detectar plágio (ALZAHRANI *et al.*, 2011; WEBER-WULFF, 2010). Para detectar paráfrase em documentos torna-se necessário o uso de técnicas hábeis em capturar e representar as informações dos caracteres, das palavras, das sentenças, dos parágrafos e outras que vão além desses níveis como estruturais, semânticas e retóricas. Dentre essas informações a que transmite a intenção do autor do texto como também a mensagem que ele deseja divulgar ao leitor é a retórica[1] que engloba as demais citadas MANN & THOMPSON (1987); O'REILLY & PAUROBALLY (2010).

[1]Segundo Aristóteles, é habilidade de usar os meios disponíveis de persuasão, como: ordem e estrutura das ideias; escolhas das palavras e a forma de usá-las; e outras. (BORCHERS & HUNDLEY, 2018)

Com o propósito de representar a estrutura de um documento para assimilar sua organização hierárquica de modo que permita obter suas diversas informações contextuais distribuídas em seu corpo textual (ALZAHRANI *et al.*, 2011), o RST foi escolhido para retratar a estrutura do documento, pois o RST é capaz de capturar a retórica contida no texto por meio da análise das proposições relacional inferidas através da estrutura do texto durante o processo de interpretação, além de ser *framework* descritivo focado na identificação da estrutura hierárquica do texto, provendo uma forma genérica de descrição para as relações entre os segmentos do texto (EDUs), por meio de demarcações gramáticas ou léxicas (MANN & THOMPSON, 1987).

Com o potencial do RST em representar a estrutura do texto, alinhando-o com o quê foi dito por ALZAHRANI *et al.* (2011), que a combinação da característica estrutural com a característica léxica ou a sintática ou a semântica contribui para a detecção de plágio de paráfrase em textos, acredita-se que uso do RST aliado à um ou mais dessas outras características poderá identificar se há ou não plágio de paráfrase entre dois documentos.

Para representar os documentos d_1 e d_s utilizando o RST é necessário segmentar o texto em EDUs, relembrando que as EDUs são unidades elementares de discurso que tem integridade funcional independente. Quando d_1 e d_s são seccionados em EDUs origina-se os conjuntos $E_{d_1} = [e_1, e_2, ..., e_n]$ e $E_{d_s} = [e_1, e_2, ..., e_m]$, respectivamente.

Quando os documentos estão segmentados, a construção da árvore dá-se por meio de demarcações das relações entre as EDUs contiguas no texto, sempre respeitando as regras estabelecidas pelo *schema application* e a estrutura de análise como descrito nas seções 3.2.4 e 3.2.5.1, respectivamente. Ao fim do processo de construção da árvore sobre os conjuntos E_{d_1} e E_{d_s}, a saída é uma árvore "binarizada"T_{d_1} e T_{d_s}, nessa ordem.

Relembrando do exemplo de uso do RST que utiliza o trecho de texto "*Meet the Announcers*"(MANN & THOMPSON, 1988) na seção 3.2.5.2, a figura 5.6 retrata a estrutura RST desse trecho na árvore binária. A identificação do núcleo e satélite são representados por [N] e [S] respectivamente; a ordem deles ao lado da relação indica quem dos nós é o núcleo e o satélite (FENG, 2015). As folhas são as EDUs, os nós não terminais são as relações estabelecidas entre as partes do texto, sendo uma EDU ou um *schema application*.

Semelhante ao processo descrito por SOCHER *et al.* (2011a), contudo, ao invés de usar a árvore gerado pelo POS-Tagging como entrada para o RAE, a ideia é utilizar a estrutura do RST, conforme apresentada na figura 5.6, como insumo para o RAE a fim de que ele gere as representações para cada nó da árvore como descrito na seção 3.4.3. Porém, essa abordagem apresenta uma dificuldade. O RAE espera como entrada a árvore binária e as *embeddings* das suas folhas, e atualmente não

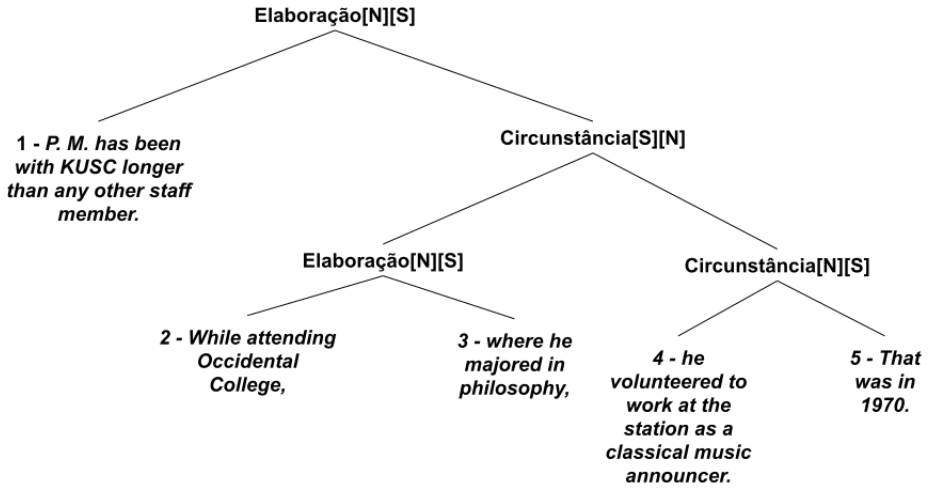


Figura 5.6 – Representação em árvore binária do resultado da análise feita na seção 3.2.6 e expressa na figura 3.3. Árvore gerada pela aplicação de FENG (2015)

há *embeddings* para EDUs (*EDUs embeddings*). Por conta dessa complicação, foi feita uma adaptação na estrutura da árvore do RST para ser trabalhada com *word embeddings* ao invés de *EDUs embeddings* quando submetida ao RAE. Para ser possível utilizar as *word embeddings* em conjunto com a estrutura RST, foi necessário aplicar o POS-Tagging nas EDUs para que as folhas da árvore RST tornassem em palavras, criando $P_{d_1} = [p_1, p_2, ..., p_n]$, $P_{d_s} = [p_1, p_2, ..., p_m]$ que são os POS-tagging das EDUs contidas em E_{d_1} e E_{d_s}, respectivamente. A imagem 5.7 demonstra como a árvore aparenta após a transformação.

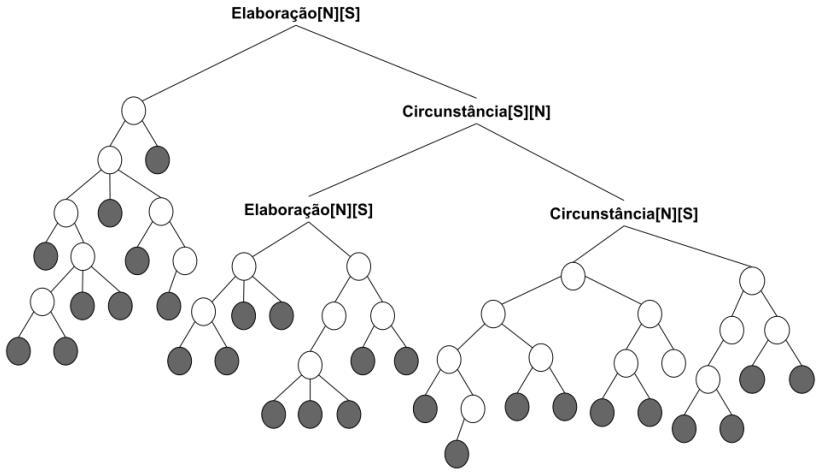


Figura 5.7 – Árvore RST do *Meet the Announcers* com POS-Tagging nas EDUs

Após a transformação, são geradas as árvores T'_{d_1} e T'_{d_s} que representam os documentos d_1 e d_s, respectivamente. A ação do POS-Tagging sobre as árvores do RST causa a "desbinarização" delas, tornando necessário a aplicação da compressão unária e da FNC sobre elas, as "binarizando" novamente.

A quantidade de nós existentes nas árvores é definida utilizando a equação 5.4.

Com as árvores preparadas para trabalharem com as *embeddings* das palavras, logo estão aptas a serem submetidas ao RAE. Quando submetidas, são geradas as representações para os nós não terminais como também para a raiz da árvore.

Após o processamento do RAE, os documentos estão prontos para serem comparados nó-a-nó pela matriz de similaridade $A \in \mathbb{R}^{z \times y}$, onde $z = l_{T'_{d_1}}$ e $y = l_{T'_{d_s}}$. A matriz é preenchida com a distância entre os nós em comparação. Uma vez pronta essa matriz, que pode variar de dimensionalidade de acordo com o par de documentos, o *dynamic pooling* é aplicado sobre ela para gera uma matriz com tamanho fixo e servir como insumo ao classificador para detectar se há plágio de paráfrase.

Embora o RST seja muito promissor no auxílio à detecção de plágio de paráfrase, a impossibilidade atual em trabalhar com EDUs *embeddings* forçando a utilização do POS-Tagging para usar os *embeddings* das palavras, pode afetar na eficácia na execução da tarefa de detecção.

5.3 Fluxo da geração das representação

Nesta seção são explicados as tarefas executadas durante a geração das representações e a detecção de plágio de paráfrase. A figura 5.8 contém o fluxo de execução para os as abordagens POS-Tagging/RAE e RST/RAE.

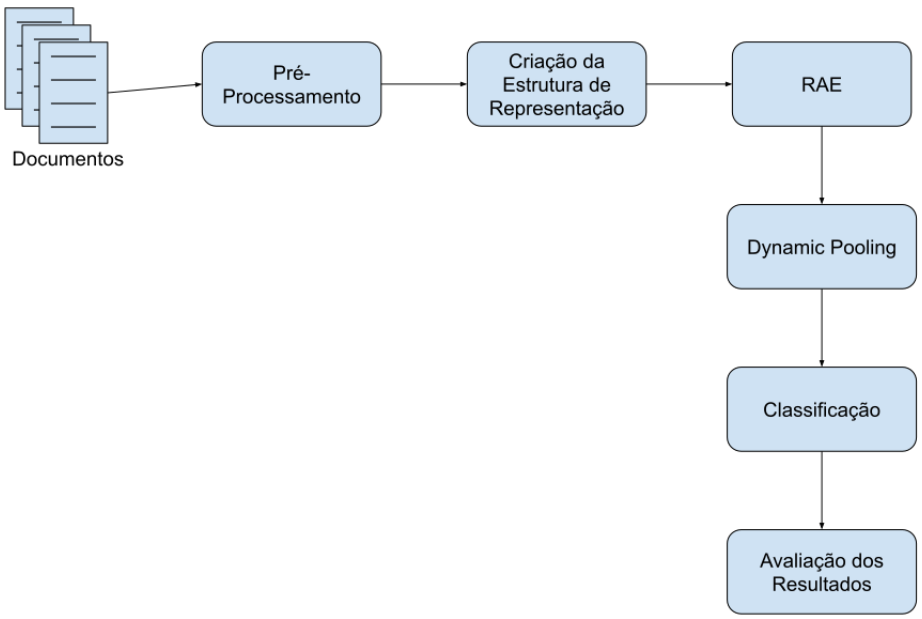


Figura 5.8 – Fluxo de Execução para

Com uma lista dos pares dos documentos, segue o seguinte fluxo:

1. A tarefa de pré-processamento é responsável por remover e normalizar alguns dados e preparar os documentos para a próxima tarefa do fluxo. O

pré-processamento faz: a substituição de todos os números, inteiros ou fracionários, por zero; a substituição de todas as ocorrências de três ou mais espaço em branco, quebra de linha, pontuação do mesmo tipo, para uma ocorrência de cada uma respectivamente. Para o POS-Tagging/RAE, o documento é dividido em sentenças.

2. O próximo passo é a geração das estruturas que representam os documentos, para esse fluxo a estrutura utilizada é a árvore binária gerada pelo POS-Tagging com adaptação proposta ou pelo RST.

3. A tarefa do RAE, com a estrutura em árvore e *word embeddings*, é gerar recursivamente *embeddings* para nós não terminais que não têm representações até chegar a raiz.

4. As tarefas desempenhadas pelo *dynamic pooling* são: fazer a comparação cartesiana nó a nó entre as árvores dos documentos; e reduzir a dimensionalidade da matriz de comparação para que sirvam como insumo para os classificadores.

5. A classificação da-se por meio da verificação se há plágio de paráfrase ou não entre os par de documentos.

6. A avaliação e análise utilizam os dados gerados durante o processo de classificação determinar como foi o eficácia das representações das características por meio do uso da árvore binária.

No início do fluxo de geração de representações, durante o pré-processamento, o par de documentos seguem até a tarefa 3 de forma independente, ou seja, são processados separadamente; no passo 4 eles são unidos novamente para as tarefas 4 e 5.

Capítulo 6

Experimentos

Neste capítulo são explicitados o objetivo dos experimentos na seção 6.1 e os resultados obtidos por eles na seção 6.4. Da mesma forma, a descrição da coleção de dados está exposta na seção 6.2. A seção 6.3 contém informações sobre como os experimentos foram conduzidos, como determinadas decisões de ajustes foram feitas, e também quais métricas e parametrizações foram utilizadas para avaliar a eficácia das representações das características dos documentos.

6.1 Objetivo dos Experimentos

Os experimentos descritos nas próximas seções, têm por finalidade avaliar a capacidade de representação das características dos documentos pelas abordagens RST/RAE e adaptação SOCHER *et al.* (2011a), POS-Tagging/RAE pela observação do quanto os classificadores utilizados durantes os experimentos são influenciados pelas representações. A avaliação da-se através da classificação dos pares de documentos identificando se há ocorrência de plágio de paráfrase ou não, sem indicar a posição exata do trecho textual que os pares compartilham entre si.

6.2 Coleção de Dados

6.2.1 *Microsoft Research Paraphrase Corpus* (MSRPC)

O *data set Microsoft Research Paraphrase Corpus* (MSRPC) construído por DOLAN *et al.* (2004) composto por 5801 pares de sentenças sendo 3900 casos de paráfrase. O *corpus* é divido em conjunto de treino contendo 4076 pares contendo 2753 ocorrências de paráfrase, e o conjunto de teste quantificado em 1725 pares, desses 1147 são paráfrase.

DOLAN *et al.* (2004) construiu o *data set* sobre artigos de reportagens coletadas por um período tempo, clusterizando-os de modo a facilitar a seleção dos casos de

paráfrases, ou seja, selecionava artigos do mesmo *cluster* a fim de produzir eventos de paráfrase. A produção deu-se por meio de duas técnicas não-supervisionadas: distância de Levenshtein e uma estratégia heurística.

A **distância de Levenshtein**, a qual, durante o processo de construção dos pares de sentenças, auxiliava na rejeição dos pares que eram idênticos ou diferenciavam-se apenas na pontuação, como também aqueles com a sentença mais curta com tamanho inferior à dois terços da sentença com maior comprimento.

A **estratégia heurística** era induzida a selecionar pares de paráfrases com conjunto de palavras muitos distintos, essa escolha baseava-se em uma convenção comum no meio jornalístico, a qual afirma sobre reportagens que tratam a mesma notícia ou mesmo assunto tendem a sumarizar as informações iniciais do artigo originador (DOLAN *et al.*, 2004). Apoiado sobre essa afirmativa, eram escolhidas as duas primeiras sentenças das reportagens que tratavam do mesmo assunto. Após a seleção dos pares candidatos ao *data set*, aplicava-se o processo de filtragem, onde: o par de sentença que não compartilhava entre si pelo menos três palavras com mais de quatro caracteres era descartado; a quantidade de palavras contida na menor sentença deveria ter no mínimo a metade da quantidade da sentença maior.

O resultado oriundo da execução dos métodos sobre os artigos clusterizados são dois conjuntos de pares de sentenças: um gerado pela distância de Levenshtein, L12 *corpus* com 139K pares; e o segundo formado pela estratégia heurística, F2 *corpus* com 214K pares. DOLAN *et al.* (2004) categorizou as paráfrase em seis tipos: **elaboração** diz que o par de sentenças pode diferir no conteúdo em uma palavra, frase ou cláusula; ***"phrasal"*** utiliza uma frase para substitui uma palavra; **grafia** são as palavras que apresentam escritas diferentes mas o significado permanece o mesmo; **sinônimo** utiliza a substituição de uma ou mais palavras com o mesmo significado entre as sentenças; **anáfora** é retomada de um termo por outro mais complexo ou simples; **reordenação** faz mudança na ordem da ocorrência das palavras.

DOLAN *et al.* (2004) afirma que o *data set* produzido por L12 mostrou-se incompleto a respeito de fonte de informações sobre paráfrase, não reconhecendo alternativas complexas de paráfrase, como: alteração léxica longa, reordenações. Enquanto, o *data set* feito por F2, apresentou pares de sentenças com tipos de paráfrase mais complexas.

6.2.2 *Paraphrase for Plagiarism Corpus*

O *data set Paraphrase for Plagiarism* (P4P) construído por BARRÓN-CEDEÑO *et al.* (2013) composto por 847 pares de fragmentos de textos contendo 11243 casos de plágio de paráfrase, onde, dessa quantidade, apenas 35 não contém caso de plágio

de paráfrase. Esses casos de plágio estão distribuídos entre 639 pares de documentos distintos, sendo 617 são documentos fonte e 556 são documentos suspeitos.

BARRÓN-CEDEÑO *et al.* (2013) construiu o *data set* sobre PAN-PC-10 *corpus*, que foi construído para Competição Internacional sobre Detecção de Plágio (*International Competition on Plagiarism Detection*)(POTTHAST *et al.*, 2010). O PAN-PC-10 contém 27073 documentos contendo 65558 casos de plágio, que foram feitos por meio de estratégias: geração automática, criação manual e geração por tradução automática de textos em alemão e espanhol (POTTHAST *et al.*, 2010). Desse *corpus*, BARRÓN-CEDEÑO *et al.* (2013) escolheu apenas os pares de documentos que tinham até 50 palavras plagiadas, originando a quantidade de pares igual a 847.

Durante a construção do *corpus*, os casos de plágio foram classificados em classes, subclasses e tipos, será anunciado apenas as classes a seguir: a classe **morfo-léxica** atua diretamente nas palavras; a classe **estrutural** concentra-se em sentenças ou no discurso; a classe **semântica**; a classe **geral** que age por meio de deleção ou adição; outras que contém **cópia idêntica** e **não plágio**. A distribuição das ocorrências de plágio nas classes está dividida da seguinte maneira: morfo-léxica 62,2%, estrutura 15,5%, semântico 3% e geral 18,2% e outros 1,2%. É notório a dominância da classe morfo-léxica, que atua em grande parte somente na característica léxica. BARRÓN-CEDEÑO *et al.* (2013) afirma que a classe morfo-léxica é a prática mais comum por pessoas pelo fato de ser a mais fácil de ser executada.

6.3 Metodologia

6.3.1 Sample P4P

Dado o alto grau de desbalanceamento entre as classes de plágio de paráfrase e a não plágio no *data set*, sugeriu-se fazer um recorte no *corpus* e aumentar os casos de não plágio de paráfrase para avaliar o comportamento dos classificadores sobre as representações desenvolvidas nesse trabalho sem o desbalanceamento.

Para selecionar os pares de documentos para comporem o recorte do P4P, foi utilizado a soma dos tamanhos dos fragmentos nos pares de documentos que podem ou não conter casos de plágio de paráfrase, ou seja, o tamanho do fragmento no documento suspeito é somado ao tamanho do fragmento no documento fonte, caso houvesse mais de um par de fragmentos entre os documentos, as suas somas são agregados. A figura 6.1 demonstra a distribuição da soma dos tamanhos dos pares de fragmentos por pares de documentos.

A partir da distribuição soma dos tamanhos dos fragmentos, foram criados intervalos para, assim, selecionar os pares de documentos. Os intervalos são: $soma < 500$;

$500 \leqslant soma \leqslant 700$; e $700 < soma$. A tabela 6.1 contém a quantidade documentos por intervalo. Foram selecionado 10 pares de documento em cada intervalo e selecionados mais cinco aleatoriamente, totalizando 35 pares de documentos com casos de plágio de paráfrase. Além dos 9 pares de documentos que não contém plágio, foram selecionados mais 26 pares do PAN-10 corpus somando 35 pares de documentos com não ocorrência de plágio de paráfrase. O recorte do P4P, P4P@70, contém 70 pares de documentos com 138 documentos distintos.

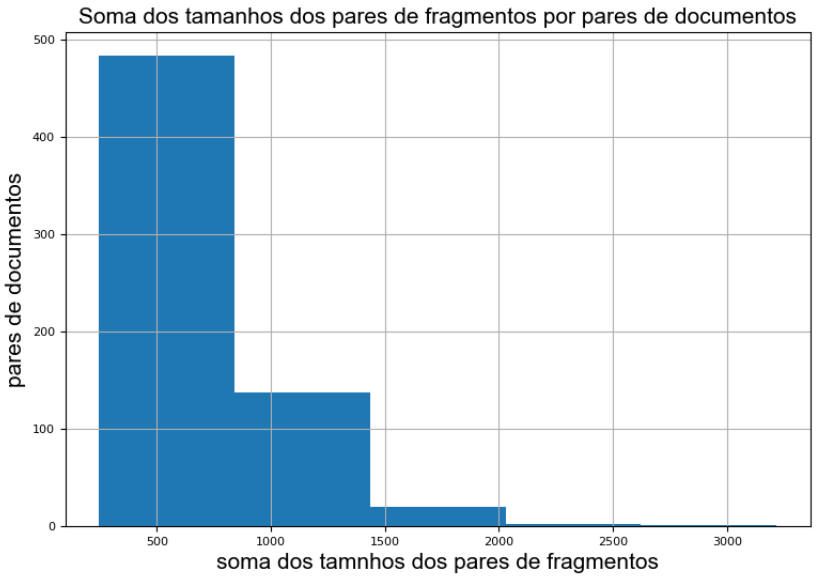


Figura 6.1 – Soma dos tamanhos dos pares de fragmentos por pares de documentos

$soma$ <500	321
500 <$soma$ <700	152
700 <$soma$	166

Tabela 6.1 – Quantidade de documentos por intervalo

6.3.2 Métricas

Para a avaliar o comportamento dos classificadores sobre as representações geradas a partir das abordagens descritas na seção 5, foram utilizadas as métricas $f1$ e acurácia.

Para explicar essas medidas será apresentando brevemente a matriz confusão como demonstra a tabela 6.2. Os valores 1 e 0 representam as classes presentes no *data set*; VN e FN são Verdadeiro Negativo e Falso Negativo, respectivamente; semelhantemente são VP e FP, respectivamente, Verdadeiro Positivo e Falso Positivo.

As medidas *precision* (P) e recall (R) são formuladas sobre os valores da matriz confusão, como descreve as equações 6.1 e 6.2, respectivamente.

	Negativo	Positivo
1	VN	FP
0	FN	VP

Tabela 6.2 – Matriz Confusão

$$P = \frac{VP}{VP + FP} \tag{6.1}$$

$$R = \frac{VP}{VP + FN} \tag{6.2}$$

A medida $f1$ é derivada das medidas *precision* e recall e tem a sua equação descrita a seguir:

$$f1 = 2\frac{PR}{P + R} \tag{6.3}$$

O valor da acurácia é obtido por meio da equação 6.4

$$acc = \frac{VP + VN}{VP + VN + FN + FN} \tag{6.4}$$

6.3.3 Ambiente Computacional

Os computadores utilizados foram: um com processador Intel Core i5-4210U, 1.70GHz com 8GB de memória RAM; outro com processador Intel Core i7-3770, 3.40GHz com 8GB de memória RAM; e o último com processador Intel Core Intel Core i7-8550U, 1,80GHz com 16GB de memória RAM. Todos utilizaram o sistema operacional Ubuntu 16.04.

Os programas foram desenvolvidos utilizando a linguagem Python 2.7 e 3.5. O código do POS-Tagging/RAE foi baseado no código original em Matlab e disponibilizado em `http://nlp.stanford.edu/~socherr/classifyParaphrases.zip`. O *parser* utilizado para gerar o POS-Tagging das sentenças foi o *Stanford Parser*, disponível em `https://nlp.stanford.edu/software/lex-parser.shtml`. A geração do RST foi por conta do *parser* desenvolvido por FENG *et al.* (2014), disponível em `http://www.cs.toronto.edu/~weifeng/software/discourse_parse-2.01.tar.gz`. As *words embeddings* utilizadas foram as disponibilizadas por TURIAN *et al.* (2010).

6.3.4 Configuração do Experimento

Nesta seção são apresentadas os conjuntos de parâmetros utilizados para as tarefas explicitadas no fluxo da seção anterior (5.3).

Para o POS-Tagging foi utilizado o método *raw_parse_sents* do nltk em conjunto com *Stanford Parser*, o qual não há alguma forma de parametrização. A aplicação disponibilizada por FENG *et al.* (2014) oferece um parâmetro que influencia na qualidade da geração da árvore RST, esse parâmetro é o *global feature*, o qual indica para o *parser* executar o processo de segmentação em EDUs duas vezes a fim de obter uma divisão com maior eficácia.

Os parâmetros utilizados durante o processo do RAE, foram os sugeridos por SOCHER *et al.* (2011a), que são: usar o *unfolding* RAE, logo todas as aparicões do termo RAE daqui em diante remetem ao *unfolding* RAE; a taxa de regularização ficou fixada em 10^{-5}; o *softmax* foi parametrizado com $0,05$; a dimensão para os *words embeddings* está definida em 100. O único parâmetro que sofreu variações foi a quantidade de épocas para a geração das *embeddings* do nós não terminais, variando em 100, 1000 e 2500 para o *data set* P4P, e variou 10, 100, 500 para o *data set* MSRPC.

O processo de *dynamic pooling* teve os seus valores variados também para a geração das matrizes, os tamanhos variaram em [15, 30, 45, 60, 75, 90, 100, 250, 500, 750, 1000] para o P4P, e variaram[15, 30, 45, 60, 75, 90, 100] para o MSRPC.

A fase de classificação, durante o treinamento, foi utilizado 3 *folds* para os dois *data sets*. Os classificadores utilizados são os da *lib scikit-learn*, que são *Naive Bayes*, *Logistic Regression*, *K-Nearst Neighbors*, *Suport Vector Machine* (SVM) e *Decision Tree*, todos os classificadores foram executados em suas configurações padrões e elas podem ser encontradas no apêndice C.

6.4 Resultados

Nesta seção são apresentados os resultados obtidos durante os experimentos das propostas, onde a subseção 6.4.1 contém os resultados sobre o *data set* MSRPC e a seção 6.4.2 expõe os dados obtido sobre o *data set* P4P. Como *baseline* para as propostas dessa dissertação, foram escolhidas duas abordagens: uma foi a de SANCHEZ-PEREZ *et al.* (2014), que foi a técnica vencedora da competição de detecção de plágio em 2014, PAN 2014, a qual utiliza o *data set* do qual foi criado o P4P; e a outra escolhida foi JI & EISENSTEIN (2013) que é o atual estado da arte na tarefa detecção plágio de paráfrase no *data set* MSRPC. Ambos disponibilizaram as suas implementações, a de SANCHEZ-PEREZ *et al.* (2014), encontra-se nesse endereço `https://www.gelbukh.com/plagiarism-detection/PAN-2014/`, e a JI & EISENSTEIN (2013) neste endereço `https://github.com/jiyfeng/tfkld`.

Com objetivo de simplificar a notação, a tabela 6.3 apresenta abreviaturas usadas durante a demonstração dos resultados.

Nome	Sigla
Acurácia	acc
Naive Bayes	NB
Logistic Regression	LR
K-Nearest Neighbors	KNN
Suport Vector Machine	SVM
Decision Tree	DT
POS-Tagging/RAE	PTRAE
RST/RAE	RSTRAE
Global Features	GF
Tamanho do Pooling	TP
Época	EP
Classificador	Clsfc

Tabela 6.3 – Abreviatura dos métodos utilizados nos experimentos

6.4.1 MSRPC

Nesta seção serão apresentados e analisados os resultados das abordagens no *data set* MSRPC. Como a abordagem do SOCHER *et al.* (2011a) já tem o seu resultado neste *corpus*, não fará parte da apresentação dos resultados. Serão apresentados resultados do RSTRAE sem e com GF.

6.4.1.1 Resultados do RSTRAE sem GF

As tabelas A.1 e A.2, que estão no apêndice A, contém todos os resultados obtidos pela abordagem RSTRAE sem GF variando os parâmetros EP do RAE e TP do *dynamic pooling*.

A figura 6.2 apresenta o comportamento, por meio da acurácia, dos classificadores utilizando as representações das sentenças geradas pelo RSTRAE sem GF. Como é possível notar, as representações parecem não exercer muito influência sobre os classificadores, nem quando alterna valores de EP e nem quando varia com TP. Os modelos que são mais influenciados pela variação desse dois parâmetros são DT e KNN, no entanto, não são os modelos que detém a maior variação[1] de 29% de *acc* o qual é alcançado pelo modelo NB.

[1] diferença entre valor máximo e o valor mínimo obtido pelo modelo

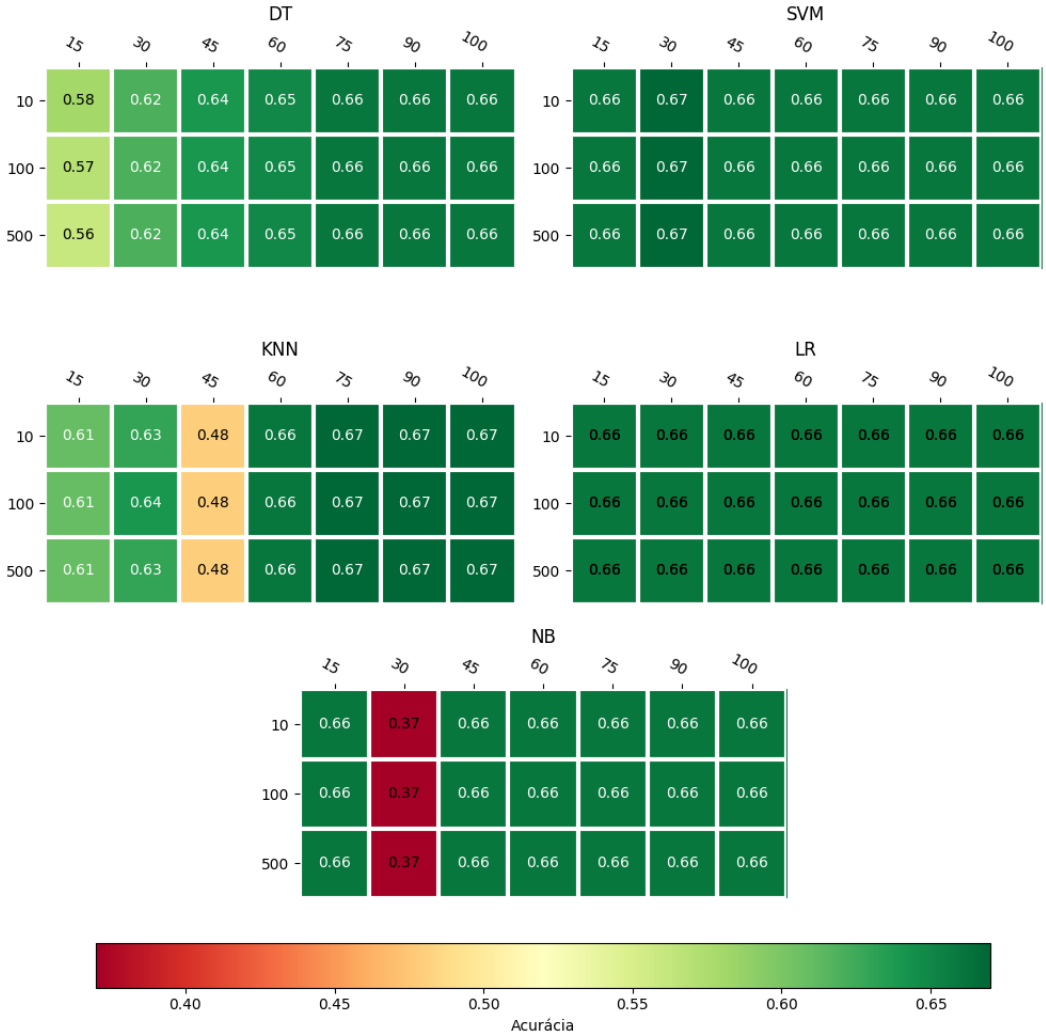


Figura 6.2 – Mapa de calor demonstrando a influência do RSTRAE sem GF e dos parâmetros EP (vertical) e TP (horizontal) no comportamento dos classificadores

As figuras 6.3 e 6.4 apresentam o gráfico com os comportamentos da *acc* e do $f1$ para os classificadores, respectivamente. No gráfico 6.3 é notável o baixo valor alcançado por NB *acc* igual à 37% com EP=[10, 100, 500] e TP=30, e também do modelo KNN atingindo *acc* de 48%. Já observando o gráfico 6.4 comprova o comportamento dos modelos KNN e NB, os quais apresentam o valor de $f1$ igual à 46% e 11%, respectivamente. No demais modelos, em ambos os gráficos, apresentam comportamentos quase idênticos, demonstrando algumas pequenas diferenças.

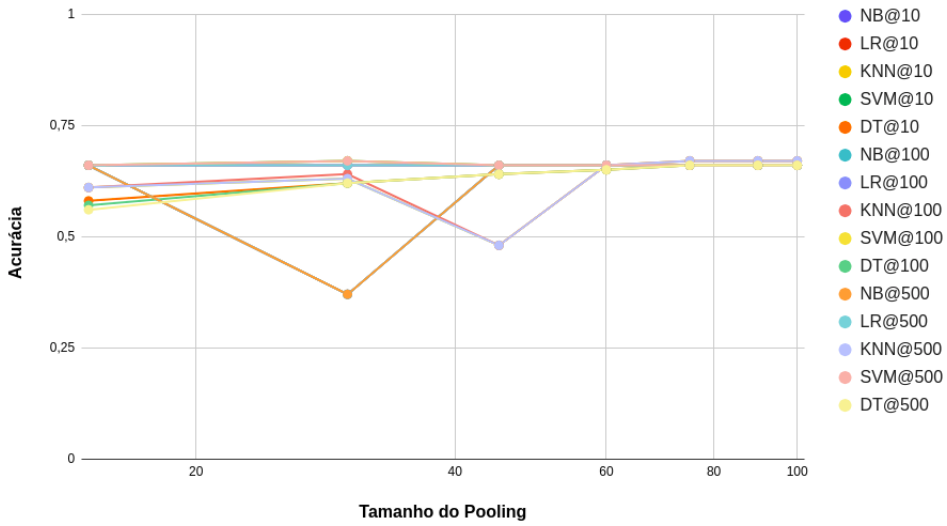


Figura 6.3 – Comportamento da *acc* dos classificadores quando usam as representações geradas pelo RSTRAE sem GF de acordo com os parâmetros EP e TP

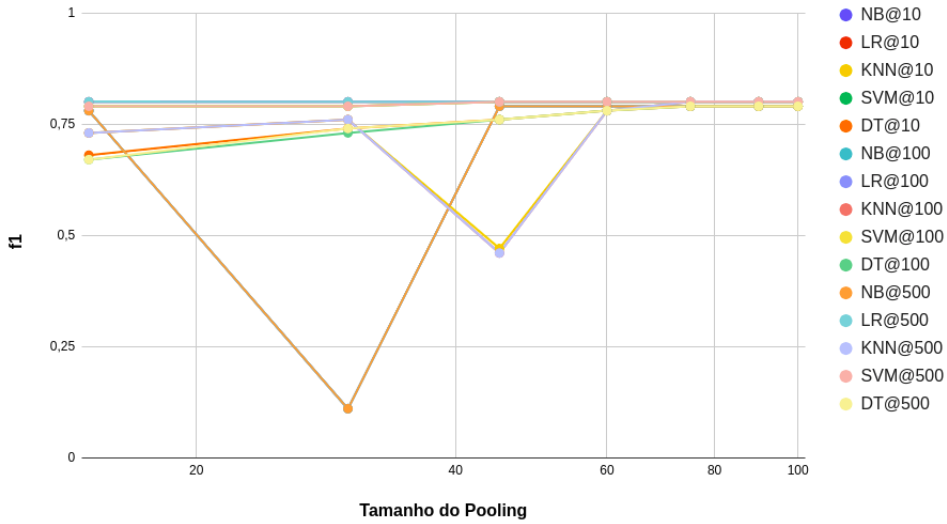


Figura 6.4 – Comportamento do $f1$ dos classificadores quando usam as representações geradas pelo RSTRAE sem GF de acordo com os parâmetros EP e TP

A tabela 6.4 apresenta os melhores resultados alcançados utilizando as representações geradas pelo RSTRAE sem GF. O melhor resultado é obtido pelo KNN com EP=[10, 100, 500] e TP=[75, 90, 100] atingindo *acc* de 67% e $f1$ de 80%.

6.4.1.2 Resultados do RSTRAE com GF

As tabelas A.3 e A.4, que estão no apêndice A, contém todos os resultados obtidos pela abordagem RSTRAE com GF variando os parâmetros EP do RAE e TP do *dynamic pooling*.

A figura 6.5 apresenta o comportamento, por meio da acurácia, dos classificadores utilizando as representações das sentenças geradas pelo RSTRAE com GF.

EP	10			100			500		
TP	Clsfc	acc	f1	Clsfc	acc	f1	Clsfc	acc	f1
15	LR	0,66	0,8	LR	0,66	0,8	LR	0,66	0,8
30	SVM	0,67	0,79	SVM	0,67	0,79	SVM	0,67	0,79
45	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8
60	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8
75	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**
90	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**
100	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**

Tabela 6.4 – Os melhores resultados obtidos durante os experimentos para a abordagem RSTRAE sem GF

Os comportamentos dos classificadores com RSTRAE com GF são muito semelhantes aos comportamentos apresentados pelos classificadores com RSTRAE sem GF, havendo uma pequena diferença ou outra entre os valores.

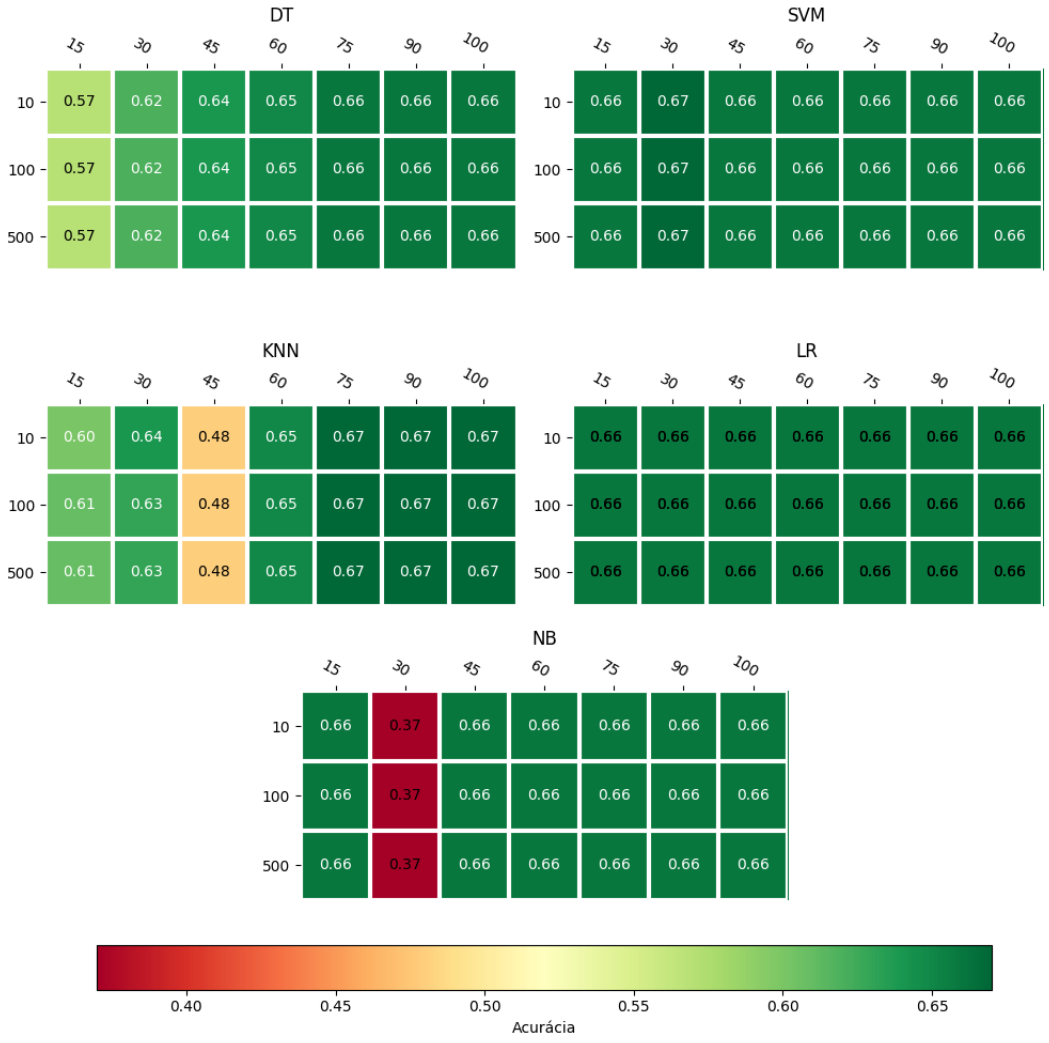


Figura 6.5 – Mapa de calor demonstrando a influência do RSTRAE com GF e dos parâmetros EP (vertical) e TP (horizontal) no comportamento dos classificadores

As figuras 6.6 e 6.7 apresentam o gráfico com os comportamentos da *acc* e do $f1$ para os classificadores, respectivamente. Os comportamentos demonstrados nos gráficos são muito semelhantes aos apresentados para a RSTRAE sem GF.

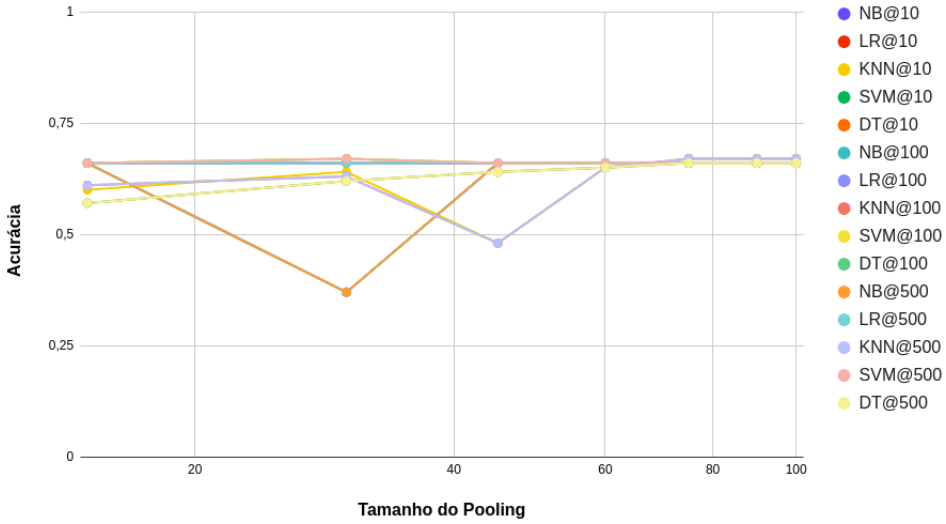


Figura 6.6 – Comportamento da *acc* dos classificadores quando usam as representações geradas pelo RSTRAE com GF de acordo com os parâmetros EP e TP

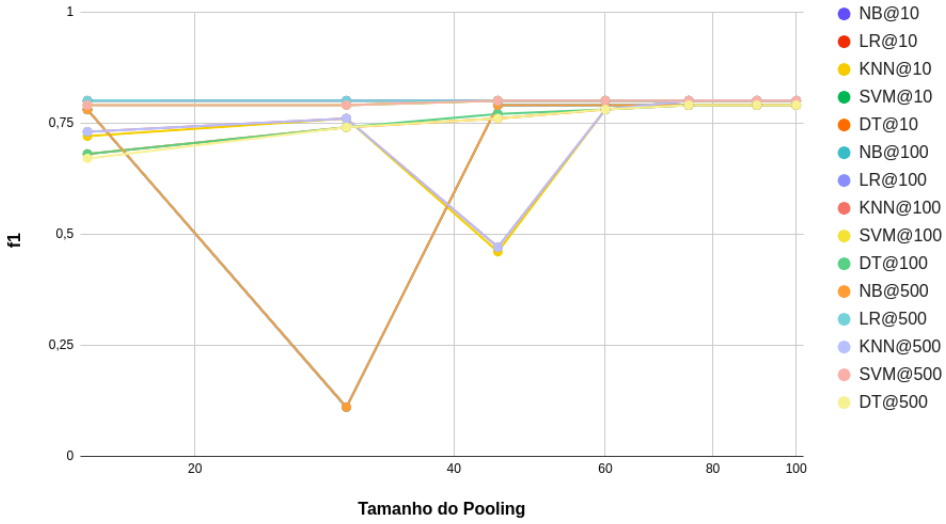


Figura 6.7 – Comportamento do $f1$ dos classificadores quando usam as representações geradas pelo RSTRAE com GF de acordo com os parâmetros EP e TP

A tabela 6.5 apresenta os melhores resultados alcançados utilizando as representações geradas pelo RSTRAE sem GF. Igual ao RSTRAE sem GF, o com GF obtém os mesmo melhores valores com KNN com EP=[10, 100, 500] e TP=[75, 90, 100] atingindo *acc* de 67% e $f1$ de 80%.

EP	10			100			500		
TP	Clsfc	acc	f1	Clsfc	acc	f1	Clsfc	acc	f1
15	LR	0,66	0,8	LR	0,66	0,8	LR	0,66	0,8
30	SVM	0,67	0,79	SVM	0,67	0,79	SVM	0,67	0,79
45	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8
60	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8	LR/SVM	0,66	0,8
75	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**
90	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**
100	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**	**KNN**	**0,67**	**0,8**

Tabela 6.5 – Os melhores resultados obtidos durante os experimentos para a abordagem RSTRAE com GF

6.4.1.3 Comparação com as outras abordagens

A tabela 6.6 demonstra os resultados obtidos por JI & EISENSTEIN (2013), SANCHEZ-PEREZ *et al.* (2014) e também pela abordagem de SOCHER *et al.* (2011a) no *data set* MSRPC em comparação com os métodos propostos nesse trabalho.

A abordagem de SANCHEZ-PEREZ *et al.* (2014) apresentou um resultado muito baixo, atingindo 34% de *acc* e 2% de $f1$. Esse resultado deve-se ao fato do método basear-se apenas nas características léxicas das sentenças, como a paráfrase pode mudar todas as palavras de um trecho plagiado, isso faz com que a abordagem de SANCHEZ-PEREZ *et al.* (2014) não tenha um êxito em detectar casos de plágio de paráfrase onde há grande modificação nas palavras entre os trechos textuais em comparação.

Os métodos JI & EISENSTEIN (2013) e SOCHER *et al.* (2011a) obtiveram os maiores resultados nesse comparação, onde JI & EISENSTEIN (2013) alcançou 80% de *acc* e 86% de $f1$, enquanto que SOCHER *et al.* (2011a) conseguiu 77% de *acc* e 83% de $f1$.

As abordagens utilizando o RSTRAE ficaram atrás das JI & EISENSTEIN (2013) e SOCHER *et al.* (2011a), apresentando uma certa deficiência quando é utilizada em textos apenas em nível de sentença. Esse baixo desempenho, de 67% de *acc* e 80% de $f1$, deve-se pelo fato das sentenças nem sempre serem separadas em EDUs, o que acarreta na não definições das relações do RST para o trecho analisado, dificultando a identificação de ocorrência de plágio de paráfrase em sentenças. Por exemplo, um par de sentenças, onde uma delas não é separada em EDUs, faz com que o processo de comparação de similaridade tenha uma certa dificuldade de identificar pontos em comum entre as sentenças, que por sua vez faz o *dynamic pooling* coletar ruído na matriz de similaridade e por fim, faz o classificador ter menos visibilidade em detectar casos de paráfrase.

Métodos	acc	f1
Sanches	0,34	0,02
Ji & Eisenstein	0,80	0,86
Socher (PTRAE)	0,77	0,83
RSTRAE com GF	0,67	0,80
RSTRAE sem GF	0,67	0,80

Tabela 6.6 – Comparação entre as abordagens no *data set* MSRPC

6.4.2 P4P *Sample*

Nesta seção serão apresentados e analisados os resultados das abordagens no *data set* P4P.

6.4.2.1 Resultados do PTRAE

As tabelas A.5, A.6, A.7 e A.8, que estão no apêndice A, contém todos os resultados obtidos pela abordagem PTRAE variando os parâmetros EP do RAE e TP do *dynamic pooling*.

A figura 6.8 demonstra a influência exercida pelo PTRAE, aliado à variação do parâmetro EP do RAE e a variação do TP do *Dynamic Pooling*, sobre os classificadores na detecção de caso plágio de paráfrase. Como é perceptível, os classificadores tem o seu desempenho afetado pela alternância dos valores dos parâmetros do RAE e do *Dynamic Pooling*. A *Decision Tree* (DT) é o classificador mais afetado pela combinação dos parâmetros EP e TP, tendo variação entre 1% e 39% no valor da sua acurácia. Enquanto que os outros modelos mantém-se quase que constantes, ocorrendo uma alteração ou outra em sua acurácia. A tabela 6.9 apresenta os melhores resultados obtidos pelo PTRAE para cada combinação dos parâmetros EP e TP, dentre os desempenhos descritos, o KNN foi o que alcançou melhor resultado com 89% de *acc* e 90% de $f1$, na configuração EP=100 ou EP=1000 junto com TP=750.

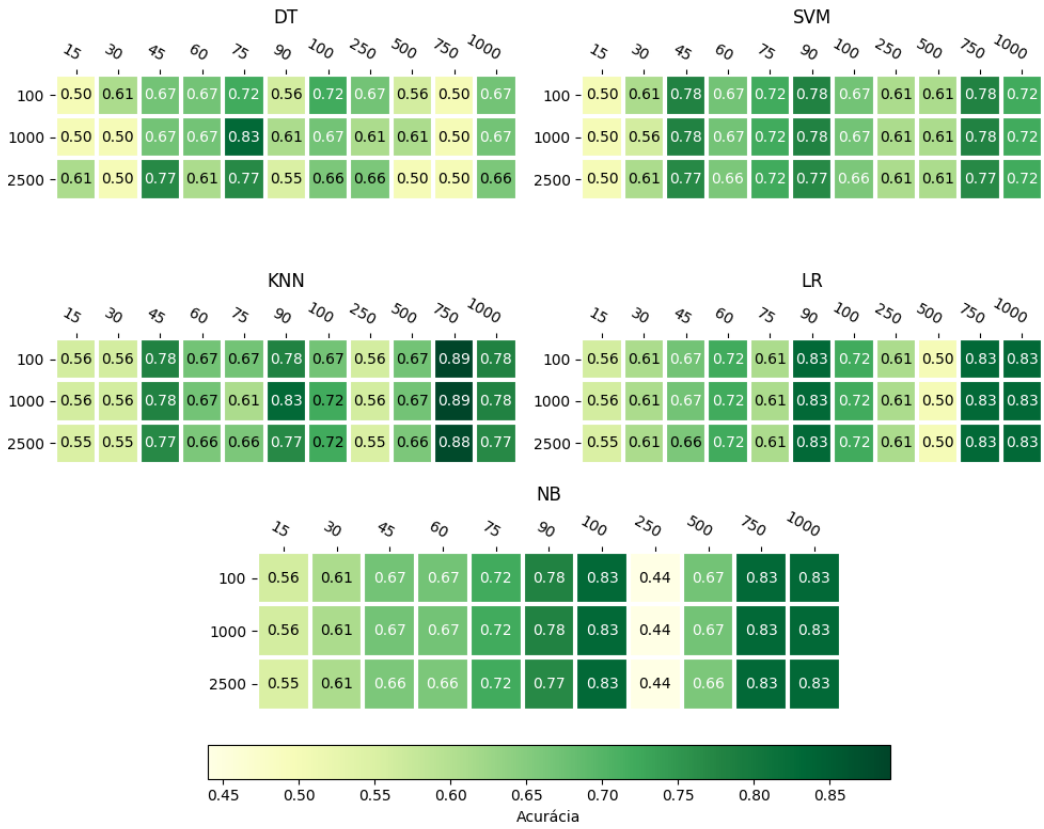


Figura 6.8 – Mapa de calor demonstrando a influência do PTRAE e dos parâmetros EP (vertical) e TP (horizontal) no comportamento dos classificadores

As figuras 6.9 e 6.10 indicam o gráfico do comportamento da acurácia e do f1, respectivamente, dos classificadores utilizando as representações geradas pelo PTRAE. No gráfico 6.9 vale destacar a baixa performance apresentada pelo NB com EP=[100, 1000, 2500] e TP=250, atingindo 44% de *acc* e 62% de $f1$, esse resultado é devido ao modelo quase sempre indicar que os casos de teste são ocorrência de plágio de paráfrase como demonstra a matriz confusão 6.7. Uma outra observação a ser feita no gráfico 6.10 é o comportamento de $f1$ do LR com EP=2500 e TP=500, o valor de $f1$ iguala-se a zero. Como demonstra a matriz de confusão 6.8, esse comportamento deve-se ao fato de LR sempre escolher a classe que não há plágio de paráfrase, por conta disso e levando em consideração a equação 6.3, faz o $f1$ igualar-se a zero e *acc* seja de 50%.

	Negativo	Positivo
1	0	9
0	1	8

Tabela 6.7 – Matriz Confusão do NB para EP=[100, 1000, 2500] e TP=250

	Negativo	Positivo
1	9	0
0	9	0

Tabela 6.8 – Matriz Confusão do LR para EP=[100, 1000, 2500] e TP=500

EP	100			1000			2500		
TP	Clsfc	acc	f1	Clsfc	acc	f1	Clsfc	acc	f1
15	LR*	0,56	0,67	KNN*	0,56	0,67	DT	0,61	0,67
30	LR*	0,61	0,7	LR*	0,61	0,7	NB	0,61	0,7
45	KNN*	0,78	0,78	KNN*	0,78	0,78	KNN*	0,77	0,78
60	LR	0,72	0,78	LR	0,72	0,78	LR	0,72	0,78
75	DT	0,72	0,78	DT	0,83	0,86	DT	0,77	0,82
90	LR	0,83	0,86	KNN*	0,83	0,86	LR	0,83	0,86
100	NB	0,83	0,86	NB	0,83	0,86	NB	0,83	0,86
250	DT	0,67	0,73	KNN*	0,61	0,7	DT	0,66	0,73
500	KNN*	0,67	0,73	KNN*	0,67	0,73	KNN*	0,66	0,73
750	**KNN**	**0,89**	**0,9**	**KNN**	**0,89**	**0,9**	KNN	0,88	0,9
1000	LR*	0,83	0,82	LR*	0,83	0,82	LR*	0,83	0,82

Tabela 6.9 – Os melhores resultados obtidos durante os experimentos para a abordagem PTRAE. O * indica que há mais de um modelo com os mesmos valores do qual está sendo apresentado na tabela.

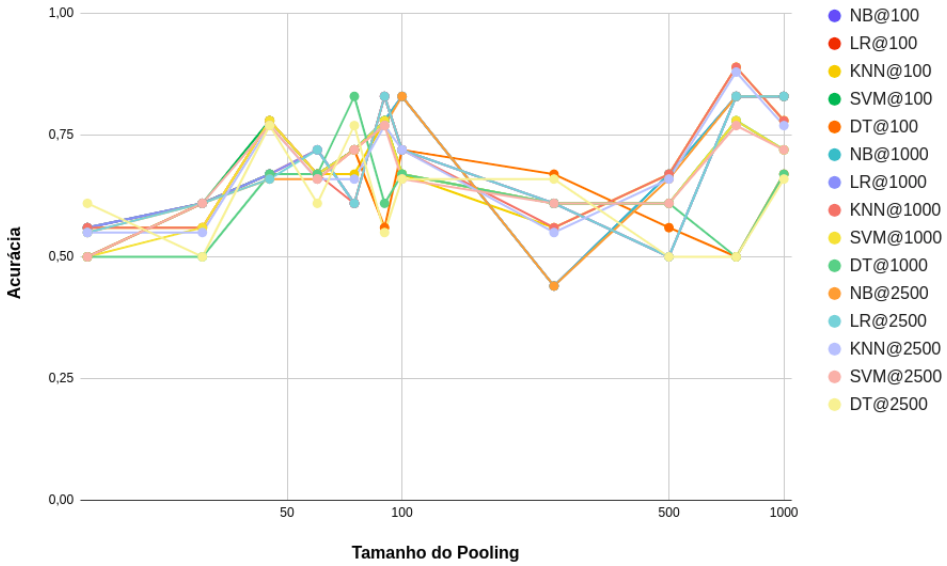


Figura 6.9 – Comportamento da *acc* dos classificadores quando usam as representações geradas pelo PTRAE de acordo com os parâmetros EP e TP

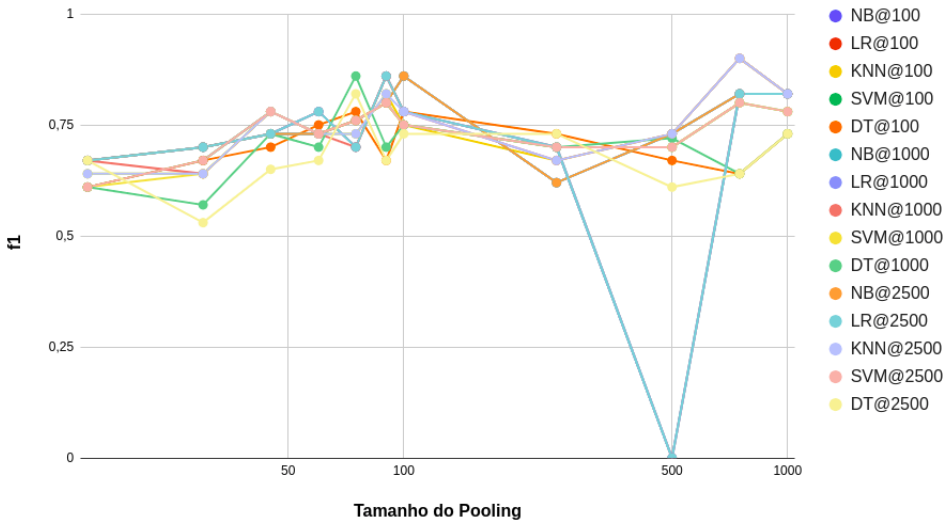


Figura 6.10 – Comportamento do $f1$ dos classificadores quando usam as representações geradas pelo PTRAE de acordo com os parâmetros EP e TP

6.4.2.2 Resultados do RSTRAE sem GF

As tabelas A.9, A.10, A.11 e A.11, que estão no apêndice A, contém todos os resultados obtidos pela abordagem RSTRAE sem GF variando o parâmetro época do RAE e variando o tamanho do *pooling*.

A figura 6.11 demonstra a influência exercida pelo RSTRAE sem GF , aliado à variação do parâmetro EP do RAE e a variação do TP do *Dynamic Pooling*, sobre os classificadores na detecção de caso plágio de paráfrase. Outra vez é notório que os classificadores tem o seu desempenho afetado pela alternância dos valores dos parâmetros do RAE e do *Dynamic Pooling*. Olhando apenas para o eixo vertical, onde ocorre a variação das épocas entre os tamanhos do *pooling*, pode encontrar variação mínima de 1% e variação máxima de 11%. Observando o eixo horizontal pode-se detectar variação mínima de 5% e variação máxima de 39%. Todas essas variações no valor da acurácia indicam que as representações geradas por RSTRAE sem GF podem influenciar o comportamento dos modelos na tarefa de detecção de plágio de paráfrase.

Novamente, a *Decision Tree* (DT) é o classificador mais afetado pela combinação dos parâmetros EP e TP, tendo variação entre 1% e 33% no valor da sua acurácia. Enquanto que os outros modelos mantém-se quase que constantes, ocorrendo uma alteração ou outra em sua acurácia. O modelo NB é o que detém a maior amplitude na variação da acurácia, atingindo 39%. A tabela 6.10 apresenta os melhores resultados obtidos pelo RSTRAE sem GF para cada combinação dos parâmetros EP e TP, dentre os desempenhos descritos, os modelos DT, KNN, SVM e NB obtiveram os melhores resultados com as seguintes configurações: DT com EP=[100, 1000, 2500] e TP=[30, 500]; SVM com EP=[100, 1000, 2500] e TP=500; KNN com EP=[100,

1000, 2500] e TP=30; NB com EP=1000 e TP=75. Com essas configurações os classificadores alcançaram 83% de *acc* e 86% de $f1$.

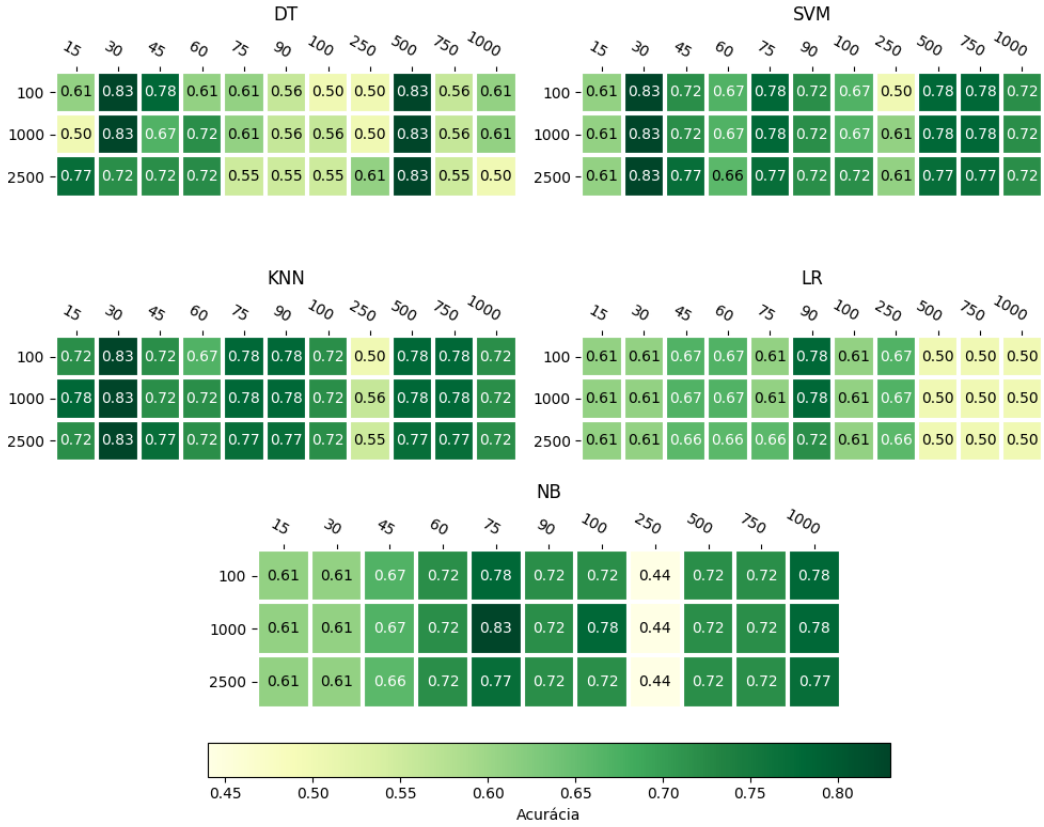


Figura 6.11 – Mapa de calor demonstrando a influência do RSTRAE sem GF e dos parâmetros EP (vertical) e TP (horizontal) no comportamento dos classificadores

As figuras 6.12 e 6.13 indicam o gráfico do comportamento da acurácia e do f1, respectivamente, dos classificadores utilizando as representações geradas pelo RSTRAE sem GF. No gráfico 6.12 vale destacar a baixa performance apresentada pelo NB com EP=[100, 1000, 2500] e TP=250, atingindo 44% de *acc* e 62% de $f1$, o motivo desse desempenho é o mesmo descrito na seção 6.4.2.1. O classificador LR, de forma semelhante ao descrito na seção 6.4.2.1, apresenta uma performance de 50% de *acc* e $f1$ igual à zero, para as configurações EP=[100, 1000, 2500] e TP=[500, 750, 1000]. O motivo desse resultado é devido ao mesmo motivo descrito na seção 6.4.2.1, mas de forma diferente, esse comportamento perdura além do TP igual 500 alcançando TP=[750, 1000].

EP	100			1000			2500		
TP	Clsfc	acc	f1	Clsfc	acc	f1	Clsfc	acc	f1
15	KNN	0,72	0,74	KNN	0,78	0,78	DT	0,77	0,78
30	**KNN**	**0,83**	**0,86**	**KNN**	**0,83**	**0,86**	**KNN**	**0,83**	**0,86**
45	DT	0,78	0,8	SVM*	0,72	0,78	KNN*	0,77	0,82
60	NB	0,72	0,76	KNN	0,72	0,78	KNN	0,72	0,78
75	NB	0,78	0,8	NB	0,83	0,84	NB	0,77	0,8
90	LR*	0,78	0,8	KNN*	0,78	0,8	KNN	0,77	0,8
100	KNN	0,72	0,76	NB	0,78	0,78	SVM*	0,72	0,76
250	LR	0,67	0,73	LR	0,67	0,73	LR	0,66	0,73
500	**DT**	**0,83**	**0,86**	**DT**	**0,83**	**0,86**	**DT**	**0,83**	**0,86**
750	KNN*	0,78	0,82	KNN*	0,78	0,82	KNN	0,77	0,82
1000	NB	0,78	0,78	NB	0,78	0,78	NB	0,77	0,78

Tabela 6.10 – Os melhores resultados obtidos durante os experimentos para a abordagem RSTRAE sem global features. O * indica que há mais de um modelo com os mesmos valores do qual está sendo apresentado na tabela.

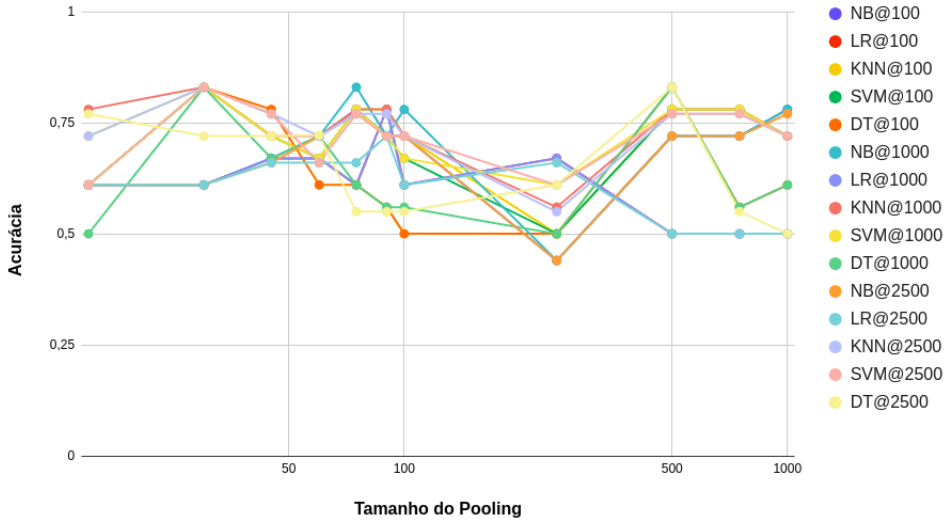


Figura 6.12 – Comportamento da *acc* dos classificadores quando usam as representações geradas pelo RSTRAE sem GF de acordo com os parâmetros EP e TP

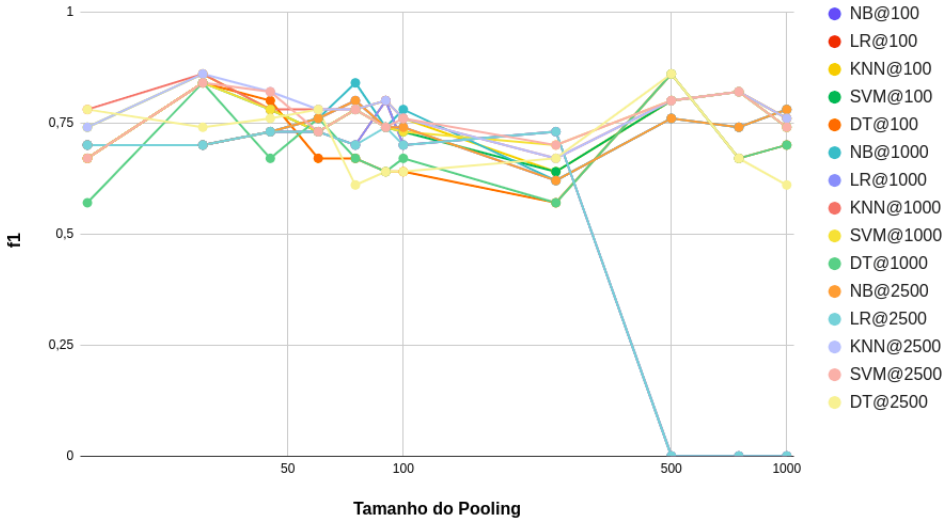


Figura 6.13 – Comportamento do $f1$ dos classificadores quando usam as representações geradas pelo RSTRAE sem GF de acordo com os parâmetros EP e TP

6.4.2.3 Resultados do RSTRAE com GF

As tabelas A.13, A.14, A.15 e A.16, que estão no apêndice A, contém todos os resultados obtidos pela abordagem RSTRAE com *global features* variando o parâmetro época do RAE e variando o tamanho do *pooling*.

A figura 6.14 exibe a influência exercida pelo RSTRAE com GF, aliado à variação do parâmetro EP do RAE e a variação do TP do *Dynamic Pooling*, sobre os classificadores na detecção de caso plágio de paráfrase. Como nas outra vezes, é notório que os classificadores tem o seu desempenho afetado pela alternância dos valores dos parâmetros do RAE e do *Dynamic Pooling*. Verificando o eixo vertical, onde ocorre a variação das épocas entre os tamanhos do *pooloing*, pode encontrar variação mínima de 1% e variação máxima de 22%. Observando o eixo horizontal pode-se detectar variação mínima de 6% e variação máxima de 39%. Todas essas variações no valor da acurácia indicam que as representações geradas por RSTRAE com GF podem alterar o comportamento dos modelos na tarefa de detecção de plágio de paráfrase.

Novamente, a *Decision Tree* (DT) é o classificador mais afetado pela combinação dos parâmetros EP e TP, tendo variação entre 1% e 33% no valor da sua acurácia, no entanto, dessa vez, os outros modelos apresentam uma maior variação, demonstrando que os modelos são mais sensíveis as representações geradas pelo RSTRAE com GF. Os modelos NB e DT detém a maior amplitude na variação da acurácia, atingindo 39%. A tabela 6.10 apresenta os melhores resultados obtidos pelo RSTRAE com GF para cada combinação dos parâmetros EP e TP, dentre os desempenhos descritos, o KNN foi o que alcançou melhor resultado com 89% de *acc* e 90% de $f1$, na configuração EP=1000 junto com TP=750.

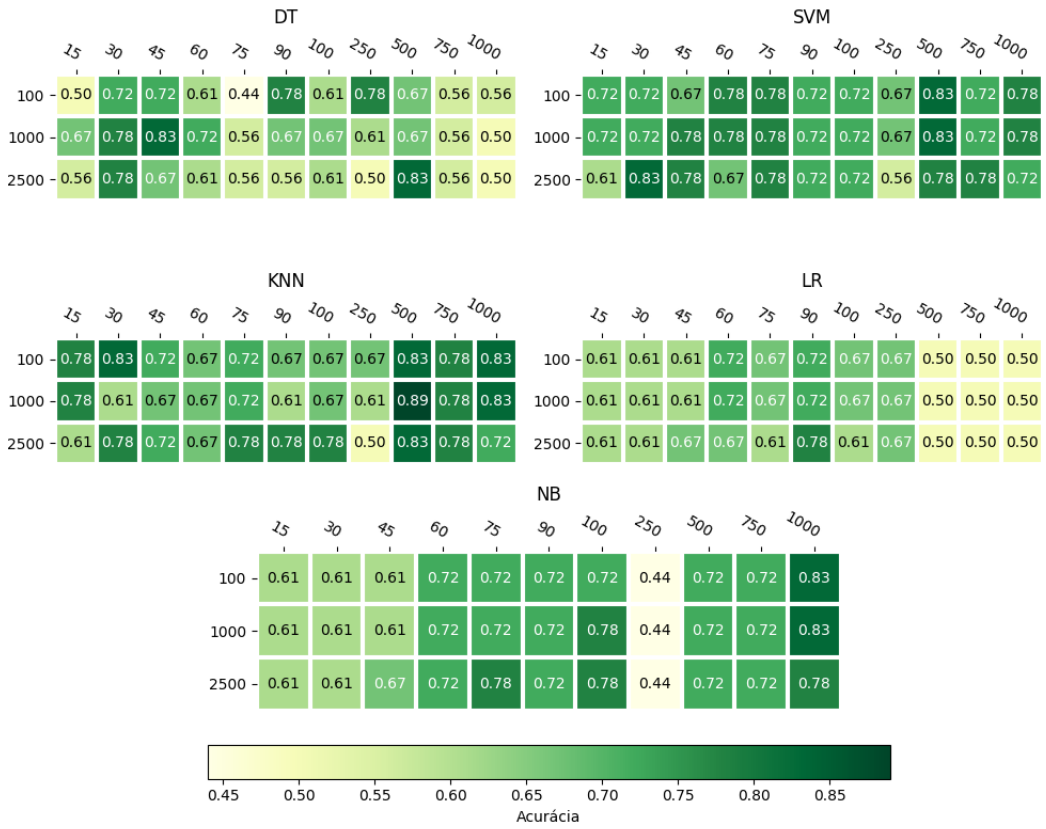


Figura 6.14 – Mapa de calor demonstrando a influência do RSTRAE com GF e dos parâmetros EP (vertical) e TP (horizontal) no comportamento dos classificadores

EP	100			1000			2500		
TP	Clsfc	acc	f1	Clsfc	acc	f1	Clsfc	acc	f1
15	KNN	0,78	0,8	KNN	0,78	0,8	LR	0,61	0,7
30	KNN	0,83	0,86	DT	0,78	0,82	SVM	0,83	0,84
45	KNN	0,72	0,76	DT	0,83	0,84	SVM	0,78	0,82
60	SVM	0,78	0,82	SVM	0,78	0,82	NB	0,72	0,76
75	SVM	0,78	0,8	SVM	0,78	0,78	NB	0,78	0,8
90	DT	0,78	0,8	LR	0,72	0,76	KNN	0,78	0,8
100	SVM	0,72	0,78	NB	0,78	0,78	KNN	0,78	0,8
250	DT	0,78	0,8	LR	0,67	0,73	LR	0,67	0,73
500	SVM	0,83	0,84	**KNN**	**0,89**	**0,9**	KNN	0,83	0,86
750	KNN	0,78	0,82	KNN	0,78	0,82	KNN	0,78	0,82
1000	KNN	0,83	0,86	KNN	0,83	0,86	NB	0,78	0,78

Tabela 6.11 – Os melhores resultados obtidos durante os experimentos para a abordagem RSTRAE com GF

As figuras 6.15 e 6.16 indicam o gráfico do comportamento da acurácia e do f1, respectivamente, dos classificadores utilizando as representações geradas pelo RSTRAE com GF. No gráfico 6.15 vale destacar a baixa performance apresentada pelo NB, novamente, com EP=[100, 1000, 2500] e TP=250, atingindo 44% de *acc* e

62% de $f1$, o motivo é mesmo descrito para as outras abordagens. O classificador LR comporta-se exatamente da mesma forma como descrita na seção 6.4.2.2.

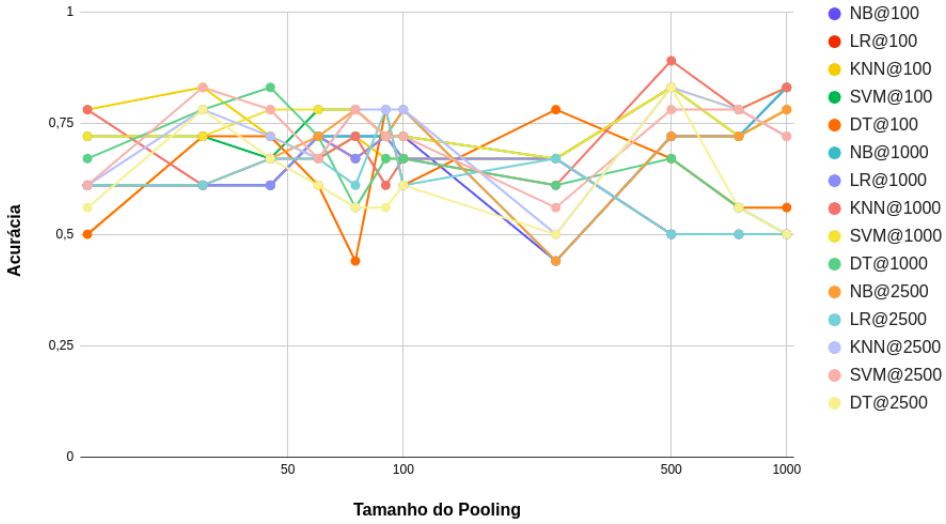


Figura 6.15 – Comportamento da *acc* dos classificadores quando usam as representações geradas pelo RSTRAE com GF de acordo com os parâmetros EP e TP

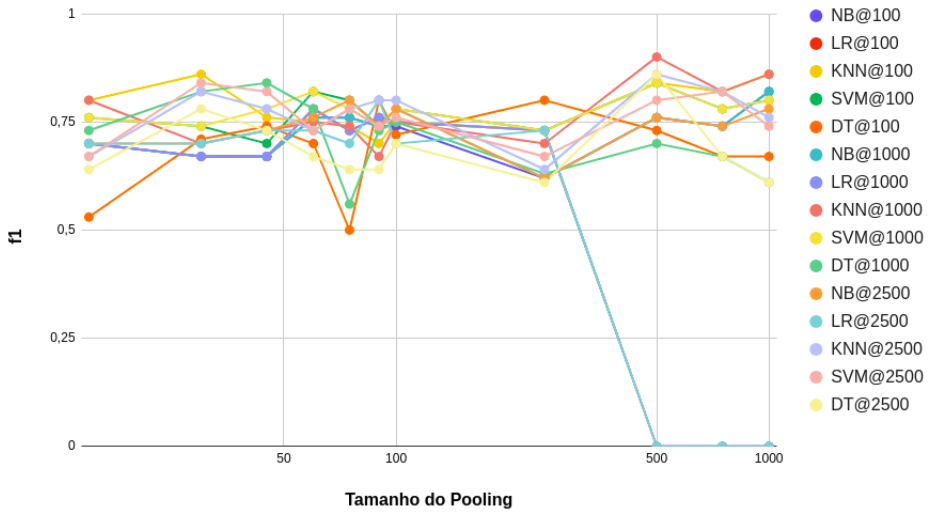


Figura 6.16 – Comportamento do $f1$ dos classificadores quando usam as representações geradas pelo RSTRAE com GF de acordo com os parâmetros EP e TP

6.4.2.4 Comparação com as outras abordagens

A tabela 6.12 demonstra os resultados obtidos por JI & EISENSTEIN (2013) e SANCHEZ-PEREZ *et al.* (2014) no *data set* P4P em comparação com os métodos propostos nesse trabalho.

A abordagem de SANCHEZ-PEREZ *et al.* (2014) obteve um bom resultado de 88% de acurácia e 88% de f1, como esperado, já que a sua técnica foi a vencedora na competição da PAN 2014 (POTTHAST *et al.*, 2014) e obteve resultado semelhante

ao apresentado. Isso deve-se ao fato do *data set* P4P ser um recorte sem modificações do *corpus* utilizado na competição, o que não afeta o comportamento da técnica de SANCHEZ-PEREZ *et al.* (2014).

Já o método de JI & EISENSTEIN (2013) apresenta um resultado muito abaixo do obtido no *data set* MSRPC, o qual foi 80,04% e 85,9% de acurácia e f1, respectivamente. Já no P4P obteve 44% de acurácia e 40% de f1. Esse mal desempenho é devido a um dos parâmetros de suas configurações, o número de componentes para serem gerado pelo SVD. Em seu artigo, JI & EISENSTEIN (2013) descreve o melhor número de componentes a serem gerados pelo SVD para o *data set* da MSRPC sendo $k = 400$. No entanto, em seu código, utiliza a implementação da biblioteca *scipy.sparse.linalg.svd*, qual limita o número de componentes pela menor dimensão da matriz, isto é, seja a matriz $A_{i,j}$, onde $[i, j] \in \mathbb{R}$, deve $k \leqslant min(i, j)$. Aliado a esse fato, o recorte feito ao P4P fez com que o número máximo de componente ficasse limitado quantidade de documentos, a qual é igual à 138. Logo, não foi possível avaliar de forma eficiente o comportamento da abordagem de JI & EISENSTEIN (2013).

Métodos	acc	f1
Sanches	0,88	0,88
Ji & Eisenstein	0,44	0,49
PTRAE	0,89	0.90
RSTRAE com GF	0,89	0,90
RSTRAE sem GF	0,83	0,86

Tabela 6.12 – Comparação entre as abordagens no *data set* P4P

Capítulo 7

Conclusões e Trabalhos Futuros

Neste trabalho foi abordado o problema de detecção de plágio de paráfrase em documentos. Visto que, como afirma ALZAHRANI *et al.* (2011), muitas abordagens falham em identificar plágio de paráfrase em documentos; por conta disso, o objetivo do estudo desenvolvido nesse dissertação focou em criar uma forma de representar os documentos para auxiliar na tarefa de análise detalhada do Sistema de Detecção de Plágio Externo a identificar casos de plágio de paráfrase.

Segundo ALZAHRANI *et al.* (2011) e CHOW & RAHMAN (2009), para detectar casos de plágio mais complexos (plágio inteligente), é necessária criar representações capazes de assimilar a característica estrutural do texto relacionando-a com as outras características como: léxica, sintática e semântica. As representações propostas nessa dissertação são aptas a unificar todas as características do texto em uma única estrutura em árvore binária, conseguindo representar desde uma palavra localizada em uma de suas folhas até um nó não terminal que detém a relação entre dois trechos do texto.

O PTRAE uni as árvores sintáticas das sentenças do documento por suas raízes, essas árvores geradas pelo POS-Tagging, após essa união, a árvore unificado passa pelo processo "binarização", atendendo as regras da Forma Normal de Chomsky. Em seguida, juntamento com as *word embeddings* das palavras, essa árvore "binarizada"é submetida ao RAE para gerar representação aos nós não terminais. E por fim, as árvores dos documentos que estão sendo analisados passam pelo processo do comparação nó a nó das árvores, onde é calculado distância entre eles dando origem a uma matriz de similaridade que por sua vez passa pelo processo do *Dynamic Pooling* para fixar o tamanho dessa matriz de modo que sevir de entrada para o classificador para indicar se há plágio de paráfrase entre eles.

Já o RSTRAE, cria a sua estrutura em árvore utilizando a teoria da estrutura retórica, onde o texto é dividido em EDUs e elas são unidas por meio de relações estabelecidas entre elas ao fim desse processo é gerado uma árvore de retórica "binarizada". Por conta de ainda não existir EDU *embeddings*, foi necessário aplicar

o POS-Tagging sobre as EDUs e trabalhar com *word embeddings*. Como o POS-Tagging não gera uma árvore "binarizada", a árvore passa pelo processo de "binarização". Então segue para RAE, depois a comparação nó a nó, *Dynamic pooling* e por fim o classificador para identificar se há plágio de paráfrase ou não.

Com os resultados obtidos durante os experimentos, pode-se concluir sobre o RSTRAE: no *data set* MSRPC teve baixo desempenho, atingindo 67% de acurácia e 80% de $f1$ tanto para sem GF como para com GF no classificador KNN, demonstrando uma certa ineficiência em ser aplicado apenas em sentenças; já no *data set* P4P *Sample*, a abordagem com GF alcançou 89% de acurácia e 90% de $f1$ e sem GF teve 83% de acurácia e 86% de $f1$, ambos com o classificador KNN. Esses resultados demonstram que o RSTRAE tem grande potencial em representar documentos ao ponto de permitir a detecção de plágio de paráfrase. Os resultados gerados utilizando a abordagem PTRAE no *data set* P4P *Sample* alcançaram 89% de acurácia e 90% de $f1$, novamente apresentando uma nova forma de detectar plágio de paráfrase.

No entanto, mesmo diante desses resultados, ainda existem parâmetros a serem explorados como os do RAE, por exemplo. Como trabalho futuro seria explorar os outros parâmetros do RAE que foram fixados como: *learning rate*, as outras dimensões para as *words embeddings*, quantidade de camadas escondidas, função de ativação e entre outras. Um outro trabalho futuro seria ajustar a abordagem RSTRAE para trabalhar diretamente com as EDUs geradas pelo RST e não com as árvores do POS-Tagging para utilizar as *words embeddings*, uma possível solução seria a criação de EDU *embedding*. Mais um trabalho após essa dissertação, seria executar o RSTRAE e o PTRAE em um outro *data set* de paráfrase maior o qual foi utilizado nos experimentos desse trabalho, no entanto até o fim desse trabalho não existia um corpus de plágio de paráfrase em documentos como o P4P. Uma solução para essa questão, seria criar um *data set* de paráfrase utilizando método semelhante ao descrito por DOLAN *et al.* (2004).

Um outro trabalho futuro seria utilizar o *Abstract Meaning Representation* (AMR), o qual consiste em representar uma sentença na estrutura em grafo. Essa abordagem não entrou nesse trabalho por conta de dificuldades encontradas em reproduzir o método descrito por ISSA *et al.* (2018). A descrição sobre o AMR e a proposta de utilizá-lo para auxiliar na tarefa de detecção de plágio de paráfrase estão descritos no apêndice B.

Referências Bibliográficas

ADHVARYU, N., BALANI, P., 2015, "Survey: Part-Of-Speech Tagging in NLP". In: *International Journal of Research in Advent Technology (E-ISSN: 2321-9637) SpecialIssue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015"*, March.

ALZAHRANI, S. M., SALIM, N., ABRAHAM, A., 2011, "Understanding plagiarism linguistic patterns, textual features, and detection methods", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 42, n. 2 (May), pp. 133–149.

BAKER, A., GHOSH, S., KUMAR, A., et al., 2008, "Apology [Plagiarism]", *IEEE Circuits and Systems Magazine*, v. 8, n. 3 (August), pp. 95–95.

BANARESCU, L., BONIAL, C., CAI, S., et al., 2013, "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, August.

BARRÓN-CEDEÑO, A., 2012, "On the mono-and cross-language detection of text reuse and plagiarism". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 914–914. ACM, March.

BARRÓN-CEDEÑO, A., VILA, M., MARTÍ, M. A., et al., 2013, "Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection", *Computational Linguistics*, v. 39, n. 4 (December), pp. 917–947.

BAUERLEIN, M., GAD-EL HAK, M., GRODY, W., et al., 2010, "We must stop the avalanche of low-quality research", *The Chronicle of Higher Education*, v. 13 (June).

BENGIO, Y., DUCHARME, R., VINCENT, P., et al., 2003, "A neural probabilistic language model", *Journal of machine learning research*, v. 3, n. Feb, pp. 1137–1155.

BERRY, C., DE LA FUENTE, J., MULLIN, M., et al., 2007, "Notice of violation of IEEE publication principles nuclear localization of HIV-1 Tat functionalized gold nanoparticles", *IEEE transactions on nanobioscience*, v. 6, n. 4 (December), pp. 262–269.

BHAGAT, R., HOVY, E., 2013, "What is a paraphrase?" *Computational Linguistics*, v. 39, n. 3 (August), pp. 463–472.

BORCHERS, T., HUNDLEY, H., 2018, *Rhetorical theory: An introduction*. Waveland Press.

BROWN, P. F., DESOUZA, P. V., MERCER, R. L., et al., 1992, "Class-based n-gram models of natural language", *Computational linguistics*, v. 18, n. 4 (December), pp. 467–479.

CARUANA, R., LAWRENCE, S., GILES, C. L., 2001, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping". In: *Advances in neural information processing systems*, pp. 402–408, December.

CASATI, F., GIUNCHIGLIA, F., MARCHESE, M., 2006, *Publish and perish: why the current publication and review model is killing research and wasting your money*. Relatório técnico, University of Trento, January.

CHENG, J., KARTSAKLIS, D., 2015, "Syntax-aware multi-sense word embeddings for deep compositional models of meaning", *arXiv preprint arXiv:1508.02354*, (August).

CHOW, T. W., RAHMAN, M., 2009, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection", *IEEE Transactions on Neural Networks*, v. 20, n. 9 (July), pp. 1385–1402.

CLOUGH, P., 2000, "Plagiarism in natural and programming languages: an overview of current tools and technologies", *Citeseer*, (June).

CLOUGH, P., OTHERS, 2003, "Old and new challenges in automatic plagiarism detection". In: *National Plagiarism Advisory Service, 2003; http://ir. shef. ac. uk/cloughie/index. html*. Citeseer, February.

COMAS, R., SUREDA, J., 2008, "Academic cyberplagiarism: tracing the causes to reach solutions", *Digithum*, v. 10, n. 10 (December).

DICIO, 2019a. "Dicionário Online de Portugês - significado de paráfrase". Septembera. `https://www.dicio.com.br/parafrase/` .

DICIO, 2019b. “Dicionário Online de Portugês - significado de plaǵio”. Augustb. https://www.dicio.com.br/plagio/ .

DICTIONARY, 2019a. “"plagiarism definition in dictionary.reference.com"”. Septembera. https://www.dictionary.com/browse/paraphrase?s=t .

DICTIONARY, 2019b. “"plagiarism definition in dictionary.reference.com"”. Septemberb. https://www.dictionary.com/browse/plagiarism?s=t .

DOLAN, B., QUIRK, C., BROCKETT, C., 2004, “Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources”. In: *Proceedings of the 20th international conference on Computational Linguistics*, p. 350. Association for Computational Linguistics, August.

DUARTE, F. R., 2017, *Identificando Plágio Externo com Lacality-Sensitive Hashing*. Ph.D. Thesis, Universidade Federal do Rio de Janeiro, Julho.

EISELT, M. P. B. S. A., ROSSO, A. B.-C. P., 2009, “Overview of the 1st international competition on plagiarism detection”. In: *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, p. 1, September.

ELHADI, M., AL-TOBI, A., 2009, “Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures”. In: *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, pp. 679–684. IEEE, November.

FENG, V. W., LIN, Z., HIRST, G., 2014, “The impact of deep hierarchical discourse structures in the evaluation of text coherence”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 940–949, August.

FENG, W. V., 2015, *RST-style discourse parsing and its applications in discourse analysis*. Ph.D. Thesis, University of Toronto (Canada), June.

FILICE, S., DA SAN MARTINO, G., MOSCHITTI, A., 2015, “Structural representations for learning relations between pairs of texts”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1003–1013, July.

GARG, N., GOYAL, V., PREET, S., 2012, "Rule based Hindi part of speech tagger". In: *Proceedings of COLING 2012: Demonstration Papers*, pp. 163–174, December.

GLOBERSON, A., CHECHIK, G., PEREIRA, F., et al., 2007, "Euclidean embedding of co-occurrence data", *Journal of Machine Learning Research*, v. 8, n. Oct (October), pp. 2265–2295.

HE, H., GIMPEL, K., LIN, J., 2015, "Multi-perspective sentence similarity modeling with convolutional neural networks". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1576–1586, September.

HINTON, G. E., OTHERS, 1986, "Learning distributed representations of concepts". In: *Proceedings of the eighth annual conference of the cognitive science society*, v. 1, p. 12. Amherst, MA, August.

HOPCROFT, J. E., 2006, *Introduction to automata theory, languages, and computation*. Pearson Education India.

INSTITUTE, J., 2012. "Josephson Institute's 2012 report card on the ethics of American youth". .

ISSA, F., DAMONTE, M., COHEN, S. B., et al., 2018, "Abstract meaning representation for paraphrase detection". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 442–452, June.

JI, Y., EISENSTEIN, J., 2013, "Discriminative improvements to distributional sentence similarity". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 891–896, October.

JOHNSON, M., 1998, "PCFG models of linguistic tree representations", *Computational Linguistics*, v. 24, n. 4 (December), pp. 613–632.

JURAFSKY, D., MARTIN, J. H., 2014. "Speech and language processing. Vol. 3". .

KASPER, R. T., 1989, "A flexible interface for linking applications to Penman's sentence generator". In: *Proceedings of the workshop on Speech and Natural Language*, pp. 153–158. Association for Computational Linguistics, February.

KINGSBURY, P., PALMER, M., 2002, "From TreeBank to PropBank." In: *LREC*, pp. 1989–1993. Citeseer, May.

KRÁL, P., 2014, *Lexical Information, Syntax and Semantics for Natural Language Processing*. Habilitation thesis, University of West Bohemia Faculty of Applied Sciences, Univerzitní 2732/8, 301 00 Plzeň 3, Czechia, December.

KUMAR, D., JOSAN, G. S., 2010, "Part of speech taggers for morphologically rich indian languages: a survey", *International Journal of Computer Applications*, v. 6, n. 5 (September), pp. 32–41.

LAI, S., LIU, K., HE, S., et al., 2016, "How to generate a good word embedding", *IEEE Intelligent Systems*, v. 31, n. 6 (May), pp. 5–14.

LAURA BANARESCU, CLAIRE BONIAL, S. C. M. G. K. G. U. H. K. K. P. K. M. P. N. S., 2014, *Abstract Meaning Representation (AMR) 1.2 specification*. USC - Information Sciences Institute.

LE, Q., MIKOLOV, T., 2014, "Distributed representations of sentences and documents". In: *International conference on machine learning*, pp. 1188–1196, January.

LEUNG, C.-H., CHAN, Y.-Y., 2007, "A natural language processing approach to automatic plagiarism detection". In: *Proceedings of the 8th ACM SIGITE conference on Information technology education*, pp. 213–218. ACM, October.

LI, Y., YANG, T., 2018, "Word embedding for understanding natural language: a survey". In: *Guide to Big Data Applications*, Springer, pp. 83–104, May.

LIDDY, E. D., 1998, "Enhanced text retrieval using natural language processing", *Bulletin of the American Society for Information Science and Technology*, v. 24, n. 4 (May), pp. 14–16.

LUND, K., BURGESS, C., 1996, "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behavior research methods, instruments, & computers*, v. 28, n. 2 (June), pp. 203–208.

MADNANI, N., TETREAULT, J., CHODOROW, M., 2012, "Re-examining machine translation metrics for paraphrase identification". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 182–190. Association for Computational Linguistics, June.

MANN, W. C., THOMPSON, S. A., 1983, *Relational Propositions in Discourse*. Technical Report ISI/RR-83-115, Information Sciences Institute.

MANN, W. C., THOMPSON, S. A., 1986, "Relational propositions in discourse", *Discourse processes*, v. 9, n. 1, pp. 57–90.

MANN, W. C., THOMPSON, S. A., 1987, *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

MANN, W. C., THOMPSON, S. A., 1988, "Rhetorical structure theory: Toward a functional theory of text organization", *Text-Interdisciplinary Journal for the Study of Discourse*, v. 8, n. 3, pp. 243–281.

MARCU, D., 1999, "Instructions for manually annotating the discourse structures of texts", *Unpublished manuscript, USC/ISI*.

MARTIN, B., 1994, "Plagiarism: a misplaced emphasis", *Journal of Information Ethics*, v. 3, n. 2 (November), pp. 36–47.

MAURER, H. A., KAPPE, F., ZAKA, B., 2006, "Plagiarism-A survey." *J. UCS*, v. 12, n. 8, pp. 1050–1084.

MERRIAM-WEBSTER, 2019. "signal definition in merriam-webster dictionary". August. `https://www.merriam-webster.com/dictionary/plagiarism#other-words` .

O'REILLY, C., PAUROBALLY, S., 2010, "Lassoing rhetoric with OWL and SWRL", *Unpublished MSc dissertation. Available: http://computationalrhetoricworkshop. uwaterloo. ca/wpcontent/uploads/2016/06/LassoingRhetoricWithOWLAndSWRL. pdf*.

OXFORD, 2019a. "The Cambridge Dictionary of Philosophy". a. `https://www.lexico.com/en/definition/paraphrase` .

OXFORD, 2019b. "The Cambridge Dictionary of Philosophy". b. `http://www.oxforddictionaries.com/us/definition/american_english/plagiarism` .

PALMER, M., GILDEA, D., KINGSBURY, P., 2005, "The proposition bank: An annotated corpus of semantic roles", *Computational linguistics*, v. 31, n. 1 (March), pp. 71–106.

POTTHAST, M., EISELT, A., BARRÓN CEDEÑO, L. A., et al., 2010, "Overview of the 2rd international competition on plagiarism detection". In: *CEUR workshop proceedings*. CEUR Workshop Proceedings, September.

POTTHAST, M., HAGEN, M., GOLLUB, T., et al., 2013, "Overview of the 5th international competition on plagiarism detection". In: *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. CELCT, September.

POTTHAST, M., HAGEN, M., GOLLUB, T., et al., 2014, "Overview of the 6th international competition on plagiarism detection". In: *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. CELCT, September.

PRECHELT, L., 1998, "Automatic early stopping using cross validation: quantifying the criteria", *Neural Networks*, v. 11, n. 4 (June), pp. 761–767.

RAFAEL T. ANCHIÊTA, M. A. S. C., PARDO, T. A. S., 2019, "SEMA: An Extended Semantic Evaluation Metric for AMR", *CoRR*, v. abs/1905.12069. `http://arxiv.org/abs/1905.12069` .

RÖSNER, D., MANFRED, S., 1986, "Zur Struktur von Texten", *Eine Einführung in die Rhetorical Structure Theory. Künstliche Intelligenz*, v. 2, pp. 14–21.

SALTON, G., WONG, A., YANG, C.-S., 1975, "A vector space model for automatic indexing", *Communications of the ACM*, v. 18, n. 11 (November), pp. 613–620.

SANCHEZ-PEREZ, M. A., SIDOROV, G., GELBUKH, A. F., 2014, "A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014." In: *CLEF (Working Notes)*, pp. 1004–1011. Citeseer, September.

SFU, 2019. "INTRO TO RST RHETORICAL STRUCTURE THEORY". `http://www.sfu.ca/rst/01intro/intro.html` .

SOCHER, R., HUANG, E. H., PENNIN, J., et al., 2011a, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection". In: *Advances in neural information processing systems*, pp. 801–809, Decembera.

SOCHER, R., PENNINGTON, J., HUANG, E. H., et al., 2011b, "Semi-supervised recursive autoencoders for predicting sentiment distributions". In: *Proceedings of the conference on empirical methods in natural language processing*, pp. 151–161. Association for Computational Linguistics, Julyb.

SOCHER, R., PERELYGIN, A., WU, J., et al., 2013, "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, October.

STEDE, M., TABOADA, M., DAS, D., 2017. "Annotation guidelines for rhetorical structure". .

SU, Z., AHN, B.-R., EOM, K.-Y., et al., 2008, "Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm". In: *2008 3rd International Conference on Innovative Computing Information and Control*, pp. 569–569. IEEE, June.

SUN, F., GUO, J., LAN, Y., et al., 2015, "Learning word representations by jointly modeling syntagmatic and paradigmatic relations". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 136–145, July.

TABOADA, M., MANN, W. C., 2006, "Rhetorical structure theory: Looking back and moving ahead", *Discourse studies*, v. 8, n. 3, pp. 423–459.

TOFILOSKI, M., BROOKE, J., TABOADA, M., 2009, "A syntactic and lexical-based discourse segmenter", *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 77–80.

TURIAN, J., RATINOV, L., BENGIO, Y., 2010, "Word representations: a simple and general method for semi-supervised learning". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394. Association for Computational Linguistics, July.

WAN, S., DRAS, M., DALE, R., et al., 2006, "Using dependency-based features to take the'para-farce'out of paraphrase". In: *Proceedings of the Australasian Language Technology Workshop 2006*, pp. 131–138, November.

WANG, Z., MI, H., ITTYCHERIAH, A., 2016, "Sentence similarity learning by lexical decomposition and composition", *arXiv preprint arXiv:1602.07019*, (February).

WEBER-WULFF, D., 2010, "Test Cases for Plagiarism Detection Software", *In Proceedings of the 4th International Plagiarism Conference, Newcastle upon Tyne, UK*, (June).

Apêndice A

Resultados Completos dos Experimentos

	TP	15		30		45	
EP	Clsfc	acc	f1	acc	f1	acc	f1
10	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,61	0,73	0,63	0,76	0,48	0,47
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,58	0,68	0,62	0,74	0,64	0,76
100	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,61	0,73	0,64	0,76	0,48	0,46
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,57	0,67	0,62	0,73	0,64	0,76
500	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,61	0,73	0,63	0,76	0,48	0,46
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,56	0,67	0,62	0,74	0,64	0,76

Tabela A.1 – Valores da acurácia e do f1 obtidos por RSTRAE sem GF na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [15, 30, 45] no *data set* MSRPC

	TP	60		75		90		100	
EP	Clsfc	acc	f1	acc	f1	acc	f1	acc	f1
10	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,66	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79
100	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,66	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79
500	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,66	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79

Tabela A.2 – Valores da acurácia e dp f1 obtidos por RSTRAE sem GF na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [60, 75, 90, 100] no *data set* MSRPC

	TP	60		75		90	
EP	Clsfc	acc	f1	acc	f1	acc	f1
10	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,6	0,72	0,64	0,76	0,48	0,46
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,57	0,68	0,62	0,74	0,64	0,76
100	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,61	0,73	0,63	0,76	0,48	0,47
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,57	0,68	0,62	0,74	0,64	0,77
500	NB	0,66	0,78	0,37	0,11	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,61	0,73	0,63	0,76	0,48	0,47
	SVM	0,66	0,79	0,67	0,79	0,66	0,8
	DT	0,57	0,67	0,62	0,74	0,64	0,76

Tabela A.3 – Valores da acurácia e do f1 obtidos por RSTRAE com GF na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [15, 30, 45] no *data set* MSRPC

	TP	60		75		90		100	
EP	Clsfc	acc	f1	acc	f1	acc	f1	acc	f1
10	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,65	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79
100	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,65	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79
500	NB	0,66	0,79	0,66	0,79	0,66	0,79	0,66	0,79
	LR	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	KNN	0,65	0,78	0,67	0,8	0,67	0,8	0,67	0,8
	SVM	0,66	0,8	0,66	0,8	0,66	0,8	0,66	0,8
	DT	0,65	0,78	0,66	0,79	0,66	0,79	0,66	0,79

Tabela A.4 – Valores da acurácia e do f1 obtidos por RSTRAE com GF na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [60, 75, 90, 100] no *data set* MSRPC

	TP	15		30		45	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,56	0,67	0,61	0,7	0,67	0,73
	LR	0,56	0,67	0,61	0,7	0,67	0,73
	KNN	0,56	0,64	0,56	0,64	0,78	0,78
	SVM	0,5	0,61	0,61	0,67	0,78	0,78
	DT	0,5	0,61	0,61	0,67	0,67	0,7
1000	NB	0,56	0,67	0,61	0,7	0,67	0,73
	LR	0,56	0,67	0,61	0,7	0,67	0,73
	KNN	0,56	0,67	0,56	0,64	0,78	0,78
	SVM	0,5	0,61	0,56	0,64	0,78	0,78
	DT	0,5	0,61	0,5	0,57	0,67	0,73
2500	NB	0,55	0,67	0,61	0,7	0,66	0,73
	LR	0,55	0,67	0,61	0,7	0,66	0,73
	KNN	0,55	0,64	0,55	0,64	0,77	0,78
	SVM	0,50	0,61	0,61	0,67	0,77	0,78
	DT	0,61	0,67	0,5	0,53	0,77	0,65

Tabela A.5 – Valores da acurácia e do f1 obtidos por PTRAE na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [15, 30, 45] no *data set* P4P

	TP	60		75		90	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,67	0,73	0,72	0,76	0,78	0,8
	LR	0,72	0,78	0,61	0,7	0,83	0,86
	KNN	0,67	0,73	0,67	0,73	0,78	0,82
	SVM	0,67	0,73	0,72	0,76	0,78	0,8
	DT	0,67	0,75	0,72	0,78	0,56	0,67
1000	NB	0,67	0,73	0,72	0,76	0,78	0,8
	LR	0,72	0,78	0,61	0,7	0,83	0,86
	KNN	0,67	0,73	0,61	0,7	0,83	0,86
	SVM	0,67	0,73	0,72	0,76	0,78	0,8
	DT	0,67	0,7	0,83	0,86	0,61	0,7
2500	NB	0,66	0,73	0,72	0,76	0,77	0,8
	LR	0,72	0,78	0,61	0,7	0,83	0,86
	KNN	0,66	0,73	0,66	0,73	0,77	0,82
	SVM	0,66	0,73	0,72	0,76	0,77	0,8
	DT	0,61	0,67	0,77	0,82	0,55	0,67

Tabela A.6 – Valores da acurácia e do f1 obtidos por PTRAE na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [60, 75, 90] no *data set* P4P

	TP	100		250		500	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,83	0,86	0,44	0,62	0,67	0,73
	LR	0,72	0,78	0,61	0,7	0,5	0
	KNN	0,67	0,75	0,56	0,67	0,67	0,73
	SVM	0,67	0,75	0,61	0,7	0,61	0,7
	DT	0,72	0,78	0,67	0,73	0,56	0,67
1000	NB	0,83	0,86	0,44	0,62	0,67	0,73
	LR	0,72	0,78	0,61	0,7	0,5	0
	KNN	0,72	0,78	0,56	0,67	0,67	0,73
	SVM	0,67	0,75	0,61	0,7	0,61	0,7
	DT	0,67	0,75	0,61	0,7	0,61	0,72
2500	NB	0,83	0,86	0,44	0,62	0,66	0,73
	LR	0,72	0,78	0,61	0,7	0,5	0
	KNN	0,72	0,78	0,55	0,67	0,66	0,73
	SVM	0,66	0,75	0,61	0,7	0,61	0,7
	DT	0,66	0,73	0,66	0,73	0,5	0,61

Tabela A.7 – Valores da acurácia e do f1 obtidos por PTRAE na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [100, 250, 500] no *data set* P4P

	TP	750		1000	
Ep	clsfc	acc	f1	acc	f1
100	NB	0,83	0,82	0,83	0,82
	LR	0,83	0,82	0,83	0,82
	KNN	0,89	0,9	0,78	0,82
	SVM	0,78	0,8	0,72	0,78
	DT	0,5	0,64	0,67	0,73
1000	NB	0,83	0,82	0,83	0,82
	LR	0,83	0,82	0,83	0,82
	KNN	0,89	0,9	0,78	0,82
	SVM	0,78	0,8	0,72	0,78
	DT	0,5	0,64	0,67	0,73
2500	NB	0,83	0,82	0,83	0,82
	LR	0,83	0,82	0,83	0,82
	KNN	0,88	0,9	0,77	0,82
	SVM	0,77	0,8	0,72	0,78
	DT	0,5	0,64	0,66	0,73

Tabela A.8 – Valores da acurácia e do f1 obtidos por PTRAE na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [750, 1000] no *data set* P4P

	TP	15		30		45	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,61	0,7	0,61	0,7	0,67	0,73
	LR	0,61	0,7	0,61	0,7	0,67	0,73
	KNN	0,72	0,74	0,83	0,86	0,72	0,78
	SVM	0,61	0,67	0,83	0,84	0,72	0,78
	DT	0,61	0,67	0,83	0,84	0,78	0,8
1000	NB	0,61	0,7	0,61	0,7	0,67	0,73
	LR	0,61	0,7	0,61	0,7	0,67	0,73
	KNN	0,78	0,78	0,83	0,86	0,72	0,78
	SVM	0,61	0,67	0,83	0,84	0,72	0,78
	DT	0,5	0,57	0,83	0,84	0,67	0,67
2500	NB	0,61	0,7	0,61	0,7	0,66	0,73
	LR	0,61	0,7	0,61	0,7	0,66	0,73
	KNN	0,72	0,74	0,83	0,86	0,77	0,82
	SVM	0,61	0,67	0,83	0,84	0,77	0,82
	DT	0,77	0,78	0,72	0,74	0,72	0,76

Tabela A.9 – Valores da acurácia e do f1 obtidos por RSTRAE sem *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [15, 30, 45] no *data set* P4P

	TP	60		75		90	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,72	0,76	0,78	0,8	0,72	0,74
	LR	0,67	0,73	0,61	0,7	0,78	0,8
	KNN	0,67	0,73	0,78	0,78	0,78	0,8
	SVM	0,67	0,73	0,78	0,78	0,72	0,74
	DT	0,61	0,67	0,61	0,67	0,56	0,64
1000	NB	0,72	0,76	0,83	0,84	0,72	0,74
	LR	0,67	0,73	0,61	0,7	0,78	0,8
	KNN	0,72	0,78	0,78	0,78	0,78	0,8
	SVM	0,67	0,73	0,78	0,78	0,72	0,74
	DT	0,72	0,76	0,61	0,67	0,56	0,64
2500	NB	0,72	0,76	0,77	0,8	0,72	0,74
	LR	0,66	0,73	0,66	0,7	0,72	0,74
	KNN	0,72	0,78	0,77	0,78	0,77	0,8
	SVM	0,66	0,73	0,77	0,78	0,72	0,74
	DT	0,722	0,78	0,55	0,61	0,55	0,64

Tabela A.10 – Valores da acurácia e do f1 obtidos por RSTRAE sem *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [60, 75, 90] no *data set* P4P

	TP	100		250		500	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,72	0,74	0,44	0,62	0,72	0,76
	LR	0,61	0,7	0,67	0,73	0,5	0
	KNN	0,72	0,76	0,5	0,64	0,78	0,8
	SVM	0,67	0,73	0,5	0,64	0,78	0,8
	DT	0,5	0,64	0,5	0,57	0,83	0,86
1000	NB	0,78	0,78	0,44	0,62	0,72	0,76
	LR	0,61	0,7	0,67	0,73	0,5	0
	KNN	0,72	0,76	0,56	0,67	0,78	0,8
	SVM	0,67	0,73	0,61	0,7	0,78	0,8
	DT	0,56	0,67	0,5	0,57	0,83	0,86
2500	NB	0,72	0,74	0,44	0,62	0,72	0,76
	LR	0,61	0,7	0,66	0,73	0,5	0
	KNN	0,72	0,76	0,55	0,67	0,77	0,8
	SVM	0,72	0,76	0,61	0,7	0,77	0,8
	DT	0,55	0,64	0,61	0,67	0,83	0,86

Tabela A.11 – Valores da acurácia e do f1 obtidos por RSTRAE sem *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [100, 250, 500] no *data set* P4P

	TP	15		30	
Ep	clsfc	acc	f1	acc	f1
100	NB	0,72	0,74	0,78	0,78
	LR	0,5	0	0,5	0
	KNN	0,78	0,82	0,72	0,76
	SVM	0,78	0,82	0,72	0,74
	DT	0,56	0,67	0,61	0,7
1000	NB	0,72	0,74	0,78	0,78
	LR	0,5	0	0,5	0
	KNN	0,78	0,82	0,72	0,76
	SVM	0,78	0,82	0,72	0,74
	DT	0,56	0,67	0,61	0,7
2500	NB	0,72	0,74	0,77	0,78
	LR	0,5	0	0,5	0
	KNN	0,77	0,82	0,72	0,76
	SVM	0,77	0,82	0,72	0,74
	DT	0,55	0,67	0,5	0,61

Tabela A.12 – Valores da acurácia e do f1 obtidos por RSTRAE sem *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [750, 1000] no *data set* P4P

	TP	15		30		45	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,61	0,7	0,61	0,67	0,61	0,67
	LR	0,61	0,7	0,61	0,67	0,61	0,67
	KNN	0,78	0,8	0,83	0,86	0,72	0,76
	SVM	0,72	0,76	0,72	0,74	0,67	0,7
	DT	0,5	0,53	0,72	0,71	0,72	0,74
1000	NB	0,61	0,7	0,61	0,67	0,61	0,67
	LR	0,61	0,7	0,61	0,67	0,61	0,67
	KNN	0,78	0,8	0,61	0,7	0,67	0,73
	SVM	0,72	0,76	0,72	0,74	0,78	0,78
	DT	0,67	0,73	0,78	0,82	0,83	0,84
2500	NB	0,61	0,7	0,61	0,7	0,67	0,73
	LR	0,61	0,7	0,61	0,7	0,67	0,73
	KNN	0,61	0,67	0,78	0,82	0,72	0,78
	SVM	0,61	0,67	0,83	0,84	0,78	0,82
	DT	0,56	0,64	0,78	0,78	0,67	0,73

Tabela A.13 – Valores da acurácia e do f1 obtidos por RSTRAE com *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [15, 30, 45] no *data set* P4P

	TP	60		75		90	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,72	0,76	0,72	0,76	0,72	0,74
	LR	0,72	0,78	0,67	0,73	0,72	0,76
	KNN	0,67	0,75	0,72	0,74	0,67	0,7
	SVM	0,78	0,82	0,78	0,8	0,72	0,74
	DT	0,61	0,7	0,44	0,5	0,78	0,8
1000	NB	0,72	0,76	0,72	0,76	0,72	0,74
	LR	0,72	0,78	0,67	0,73	0,72	0,76
	KNN	0,67	0,75	0,72	0,74	0,61	0,67
	SVM	0,78	0,82	0,78	0,78	0,72	0,74
	DT	0,72	0,78	0,56	0,56	0,67	0,73
2500	NB	0,72	0,76	0,78	0,8	0,72	0,74
	LR	0,67	0,73	0,61	0,7	0,78	0,8
	KNN	0,67	0,73	0,78	0,78	0,78	0,8
	SVM	0,67	0,73	0,78	0,78	0,72	0,74
	DT	0,61	0,67	0,56	0,64	0,56	0,64

Tabela A.14 – Valores da acurácia e do f1 obtidos por RSTRAE com *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [60, 75, 90] no *data set* P4P

	TP	100		250		500	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,72	0,74	0,44	0,62	0,72	0,76
	LR	0,67	0,75	0,67	0,73	0,5	0
	KNN	0,67	0,75	0,67	0,73	0,83	0,84
	SVM	0,72	0,78	0,67	0,73	0,83	0,84
	DT	0,61	0,72	0,78	0,8	0,67	0,73
1000	NB	0,78	0,78	0,44	0,62	0,72	0,76
	LR	0,67	0,75	0,67	0,73	0,5	0
	KNN	0,67	0,75	0,61	0,7	0,89	0,9
	SVM	0,72	0,78	0,67	0,73	0,83	0,84
	DT	0,67	0,75	0,61	0,63	0,67	0,7
2500	NB	0,78	0,78	0,44	0,62	0,72	0,76
	LR	0,61	0,7	0,67	0,73	0,5	0
	KNN	0,78	0,8	0,5	0,64	0,83	0,86
	SVM	0,72	0,76	0,56	0,67	0,78	0,8
	DT	0,61	0,7	0,5	0,61	0,83	0,86

Tabela A.15 – Valores da acurácia e do f1 obtidos por RSTRAE com *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [100, 250, 500] no *data set* P4P

	TP	15		30		45	
Ep	clsfc	acc	f1	acc	f1	acc	f1
100	NB	0,72	0,74	0,83	0,82	0,72	0,76
	LR	0,5	0	0,5	0	0,5	0
	KNN	0,78	0,82	0,83	0,86	0,83	0,84
	SVM	0,72	0,78	0,78	0,8	0,83	0,84
	DT	0,56	0,67	0,56	0,67	0,67	0,73
1000	NB	0,72	0,74	0,83	0,82	0,72	0,76
	LR	0,5	0	0,5	0	0,5	0
	KNN	0,78	0,82	0,83	0,86	0,89	0,9
	SVM	0,72	0,78	0,78	0,8	0,83	0,84
	DT	0,56	0,67	0,5	0,61	0,67	0,7
2500	NB	0,72	0,74	0,78	0,78	0,72	0,76
	LR	0,5	0	0,5	0	0,5	0
	KNN	0,78	0,82	0,72	0,76	0,83	0,86
	SVM	0,78	0,82	0,72	0,74	0,78	0,8
	DT	0,56	0,67	0,5	0,61	0,83	0,86

Tabela A.16 – Valores da acurácia e do f1 obtidos por RSTRAE com *global features* na variação do parâmetro época do RAE e na variação do tamanho do *pooling* [750, 1000] no *data set* P4P

Apêndice B

Abstract Meaning Representation

A abordagem de ISSA *et al.* (2018) não está no corpo principal do texto porque não foi possível reproduzir o resultado obtido pela técnica descrita em seu artigo. Durante divulgação do seu trabalho em seu artigo, ISSA *et al.* (2018) não deixa muito claro alguns pontos que o levaram para atingir o desempenho obtido por sua abordagem. Por exemplo, não informa qual o tipo de pré-processamento que fez, se o fez. Outro ponto confuso é forma como preencheu a matriz termo-documento, não deixa se usou *tf-idf*, ou uma nova forma que apresentou durante o artigo, ou se preencheu com o *PageRank*. Não há um fluxo contínuo que permita reproduzir os experimento sem ter dúvida se está usando a técnica certo no momento correto.

A forma como foi interpretada e reproduzida a abordagem nesse trabalho, não corresponde ao resultado obtido por ISSA *et al.* (2018). No entanto, ainda assim, for descrito o conceito sobre *Abstract Meaning Representation* (seção B.1), foi apresentado o entendimento obtido pela leitura do artigo (seção B.2), assim como também, foi elaborada uma proposta de adaptação da técnica para ser aplicada em documento (seção B.3).

B.1 *Abstract Meaning Representation*

Abstract Meaning Representation (AMR) é uma estrutura de representação linguística desenvolvida para capturar a semântica contida em uma sentença (RAFAEL T. ANCHIÊTA & PARDO, 2019). Segundo LAURA BANARESCU (2014), dada uma sentença, o AMR é capaz de distinguir "quem está fazendo o quê a quem".

LAURA BANARESCU (2014) afirma que o AMR funciona como uma *parse tree*[1] fornecendo uma única estrutura de fácil acesso à suas partes, que leva em consideração todas as palavras de uma sentença, além de não criar uma camada de anotações desconexa.

[1] parse tree é uma estrutura com raiz e ordenada que representa a organização sintática de um texto (JOHNSON, 1998)

Mesmo sendo semelhante a uma *parse tree*, o AMR comporta-se de forma diferente (LAURA BANARESCU, 2014). Ele é abstrato podendo representar inúmeras sentenças em linguagem natural (LAURA BANARESCU, 2014). Diz BANARESCU *et al.* (2013), que um dos principais objetivos do AMR é abstrair as peculiaridades sintáticas das sentenças observadas. Com essa abstração, BANARESCU *et al.* (2013) afirma: as sentenças com o mesmo significado básico podem ter o mesmo AMR.

Segundo BANARESCU *et al.* (2013), a motivação da criação do AMR foi a necessidade de prover um grande *sembanking*[2], o qual mapeia as sentenças, as suas estruturas e seus significados lógicos, para uma comunidade de pesquisa. Por ser uma estrutura fácil de lidar, o AMR facilitou o processo de anotação permitindo o surgimento de *datasets* anotados em diversos idiomas (RAFAEL T. ANCHIÊTA & PARDO, 2019).

A concepção do AMR está extremamente ligada a língua inglesa, isso o torna uma representação enviesada para o idioma inglês (BANARESCU *et al.*, 2013). Por esse fato, BANARESCU *et al.* (2013) diz que o AMR não é interlingual, ou seja, não é possível trabalhar com dois idiomas em uma única sentença. Uma outra limitação do AMR, afirma BANARESCU *et al.* (2013), ele não consegue distinguir entre eventos reais e hipotéticos, além de omitir os artigos das sentenças quando aplicado.

B.1.1 Estrutura e Notação do AMR

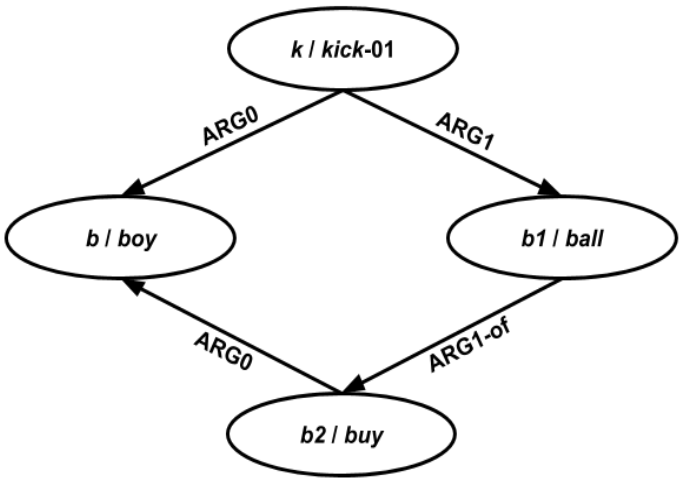
Tem a seguinte sentença: *The boy kicked the ball that he bought.* A partir dessa sentença, será apresentada a estrutura de dados utilizada pelo AMR como também a notação utilizada para representar essa estrutura. Veja a figura B.1.

Como demonstra a figura B.1a, o AMR é um grafo direcionado acíclico, com arestas rotuladas, nós rotulados e com variáveis, e um nó atuando como raiz (BANARESCU *et al.*, 2013). Sobre as variáveis, um exemplo, tem um nó com *b / boy*, *b* representa uma instância do tipo *boy* (BANARESCU *et al.*, 2013). Os nós representam os eventos e/ou os conceitos da sentença, enquanto que as arestas apresentam os relacionamentos entre os eventos e conceitos (ISSA *et al.*, 2018). Sobre eventos e conceitos, esses dois itens serão revistos e detalhados mais a frente, na seção B.1.2.2.

Já na figura B.1b, é apresentado a notação Penman[3]. Essa notação segue um estilo semelhante ao *Backus–Naur Form* (KASPER, 1989). Os termos entre parênteses, (*variáveis/ tipo da instância*), são os nós do grafo por conseguinte representam os conceitos ou os eventos da sentença (BANARESCU *et al.*, 2013). Enquanto que os

[2]semantic bank

[3]https://www.isi.edu/natural-language/penman/penman.html



(a) AMR notação grafo

```
(k / kick-01
      :ARG0 (b / boy)
      :ARG1 (b1 / ball)
            :ARG1-of(b2 / buy-01
                        :ARG0 b))
```

(b) AMR notação Penman

Figura B.1 – Estrutura AMR para a sentença: *The boy kicked the ball that he bought*

"ARG0", "ARG1" e "ARG0-of" são as arestas do grafo que por suas vez caracterizam os relacionamentos entre as entidades envolvidas (KASPER, 1989).

B.1.2 AMR e suas Definições

B.1.2.1 Relação do AMR com o PropBank

RAFAEL T. ANCHIÊTA & PARDO (2019) declara que *Proposition Bank* (PropBank) é um projeto que adota uma abordagem prática para representação semântica, adicionando uma camada de informação argumento-predicado, ou rótulos de papéis semânticos para estruturas sintáticas do *Penn TreeBank*.

RAFAEL T. ANCHIÊTA & PARDO (2019) diz que PropBank utiliza *framesets*, que são essencialmente verbos ligados a uma lista de possíveis argumentos e seus papéis semânticos. RAFAEL T. ANCHIÊTA & PARDO (2019) afirma que os *framesets* são acompanhadas por, em média, cinco argumentos. Esses argumentos são conhecidos, por convenção, como ARG0 à ARG4. Comumente, o ARG0 é agente, ARG1 é paciente, e outros (KINGSBURY & PALMER, 2002).

Com a sentença citada anteriormente, *The boy kicked the ball that he bought*, tomando como exemplo o verbo *kick*. Para esse verbo, o PropBank têm os seguintes argumentos: ARG0: *kicker*; ARG1: *thing kicked*; ARG2: *Instrument*. Logo, aplicando esse *frameset* a sentença:

ARG0 (*kicker*)	*boy*
ARG1 (*thing kicked*)	*ball*
ARG2 (*Instrument*)	como não há nenhuma indicação com o quê chutou a bola, então conclui-se que foi com o pé (*foot*)

Tabela B.1 – *Frameset* para o conceito *kick*

BANARESCU *et al.* (2013) afirma que o AMR faz um uso extremo dos *framesets* do PropBank para abstrair as sintaxes do idioma inglês.

B.1.2.2 Fundamentos do AMR

O AMR tem o seu principal foco na estrutura argumento-predicado como definido no PropBank (RAFAEL T. ANCHIÊTA & PARDO, 2019).

Segundo RAFAEL T. ANCHIÊTA & PARDO (2019) os conceitos do AMR são palavras em sua forma léxica, ou PropBank *framesets*, ou palavras chaves como *date-entity, distance-entity*, ou outros.

As relações tendem a acompanhar os conceitos. Por exemplo, caso um conceito seja de um tipo que expresse quantidade, então o AMR poderia adotar as seguintes relações: :*quant*, :*unit* ou :*scale* (BANARESCU *et al.*, 2013). Outro exemplo seria um conceito do tipo data, esse poderia estar ligado à relações como: :*day*; :*month*; :*year*; :*week*; entre outros (BANARESCU *et al.*, 2013).

BANARESCU *et al.* (2013) diz que é essencial a escolha da raiz de um AMR que representa uma sentença. Essa escolha determinará o foco principal da sentença, que por sua vez terá a construção do seu AMR baseado em relações semânticas oriundas dessa raiz (BANARESCU *et al.*, 2013).

Segundo BANARESCU *et al.* (2013) o AMR é capaz de capturar e representar verbos, substantivos, adjetivos, preposições, entidades nomeadas, questões, polaridades (positivo ou negativo), co-referências, reificações e entre outros.

B.1.3 Exemplo de uma análise usando AMR

Para fazer essa análise será usado a sentença dita anteriormente, *The boy kicked the ball that he bought.*

A fim de iniciar a construção, é necessário escolher o foco dessa sentença para definir qual o conceito será a raiz do AMR. Nesse exemplo dois conceitos podem ser escolhidos como raiz, que são: *kick* e *buy*. Sem um contexto previamente definido para essa sentença, torna-se difícil decidir qual seria foco. Para essa análise será escolhido o conceito *kick*. Sendo esse o conceito escolhido já tem a sua análise na tabela B.1.

Fazendo análise para o conceito *buy* tem o seguinte *frameset*:

ARG0 (*buyer*)	*boy*
ARG1 (*thing bought*)	*ball*
ARG2 (*seller*)	não houve vendedor mencionado na sentença
ARG3 (*price paid*)	não houve preço pago mencionado na sentença
ARG4 (*benefactive*)	não houve benfeitor mencionado na sentença

Tabela B.2 – *Frameset* para o conceito *buy*

A partir essa análise inicial, pode construir o grafo demonstrado na figura B.2:

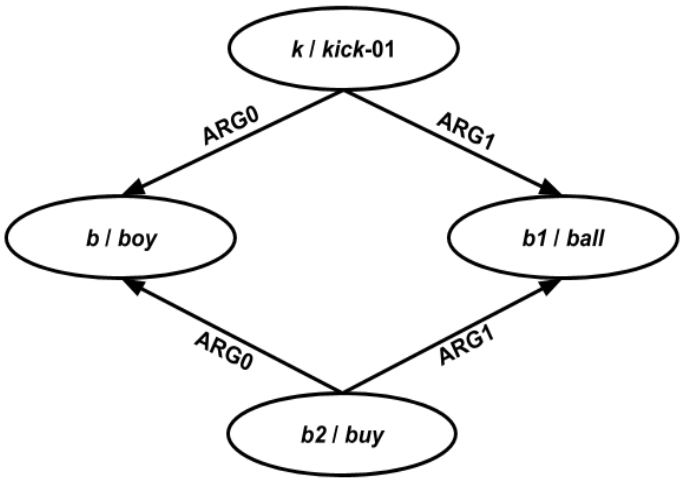


Figura B.2 – Grafo AMR resultante da análise inicial da sentença

Olhando para o grafo na figura B.2, é notável um ciclo formado, e isso vai contra ao princípio do AMR de gerar grafos acíclicos, como também ter apenas uma raiz.

Existe um artifício importante no AMR que permite a construção de relações inversas. Segundo BANARESCU *et al.* (2013) esse método tem por objetivo auxiliar que o grafo tenha apenas uma raiz e consequentemente evitar a criação de ciclos. A relação inversa funciona da seguinte maneira: *X ARG0-of Y = Y ARG0 X* (LAURA BANARESCU, 2014).

Revisando a tabela B.2, faz uma alteração em dos argumentos para desfazer o ciclo criado na análise, como mostra a tabela a seguir.

ARG0 (*buyer*)	*boy*
ARG1-of (*thing bought*)	*ball*
ARG2 (*seller*)	não houve vendedor mencionado na sentença
ARG3 (*price paid*)	não houve preço pago mencionado na sentença
ARG4 (*benefactive*)	não houve benfeitor mencionado na sentença

Tabela B.3 – *Frameset* para o conceito *buy* com alteração ARG1 para ARG1-of

Após a inversão do argumento ARG1 do conceito para ARG1-of, o grafo gerado fica igual ao da figura B.1.

B.2 AMR com fatorização de matriz

Uma outra abordagem semelhante a descrita na seção 4.3 é a utilizada por ISSA *et al.* (2018), na qual consiste na conciliação do AMR com o *PageRank*, com a análise semântica latente obtida através da fatorização de matriz e com o SVM. Como descrito na seção B.1, o AMR é uma estrutura de representação linguística desenvolvida para capturar a semântica contida em uma sentença, ISSA *et al.* (2018) utiliza essa estrutura para auxiliar na tarefa de detecção de paráfrase no *data set Microsoft Research Paraphrase Corpus* (MSRPC) (DOLAN *et al.*, 2004) transformando as sentenças em grafos direcionados acíclicos. Com esses grafos criados, é gerado o *bag of concepts*, o qual abrange à todos os conceitos distintos criados pelo AMR.

Com o *bag of concepts* pronto, torna-se possível criar a matriz sentença-conceito A sendo preenchida, inicialmente, com a frequência do conceito na sentença que é expressa na equação B.1:

$$A_{kl} = freq(l, k), \tag{B.1}$$

onde A_{kl} é a frequência do conceito l na *k-ésima* sentença. O *PageRank* é utilizado para ajustar os valores da matriz A. Esse ajuste é feito da seguinte forma: dada duas sentenças s_1 e s_2, são recuperados os seus respectivos grafos AMR, G_{s_1} e G_{s_2}, que são unidos por meio da equação B.2:

$$G_{s_1 s_2} = u(G_{s_1}, G_{s_2}), \tag{B.2}$$

onde $u(G_{s_1}, G_{s_2})$ é a união dos conceitos em comum, ou seja, todos os vértices dos grafos que apresentam o mesmo conceito servirão como ponto de junção entre eles gerando um novo grafo unificado $G_{s_1 s_2}$, como demonstra a figura B.3; os vértices $conceito - 02$ e $conceito - 03$ estão presentes no dois grafos AMR s_1 e s_2, logo eles serão pontos de união dos grafos. O grafo unificado $G_{s_1 s_2}$ é o resultado da aplicação da função $u(G_{s_1}, G_{s_2})$ que precisa atender a seguinte condição:

$$G_{s_1 s_2} = \begin{cases} u(G_{s_1}, G_{s_2}),\ \exists\ c\ [c \in g_i \leftrightarrow c \in g_j] \\ \emptyset,\ \exists!\ c\ [c \in g_i \leftrightarrow c \in g_j] \end{cases}, \tag{B.3}$$

onde c é o conceito em comum entre os dois grafos, $g_i \in G_{s_1}$ e $g_j \in G_{s_2}$.

Após a junção, o grafo unificado é submetido ao algoritmo *PageRank* que associará uma pontuação a cada vértice, que representa seu correspondente conceito

na matriz A, e essa pontuação será multiplicada pelos valores iniciais da matriz sentença-conceito como apresenta a equação B.4.

$$A_{kl} = freq(l, k) \times PR(l, k), \tag{B.4}$$

onde $PR(l, k)$ é o *PageRank* da palavra l na *k-ésima* sentença como é demonstrado a seguir:

$$PR(n) = \sum_{m \in I(n)} \frac{PR(m)}{l(m)}, \tag{B.5}$$

onde $I(n)$ são os vértices que apontam para vértice n e $l(m)$ são os arestas que saem do vértice m.

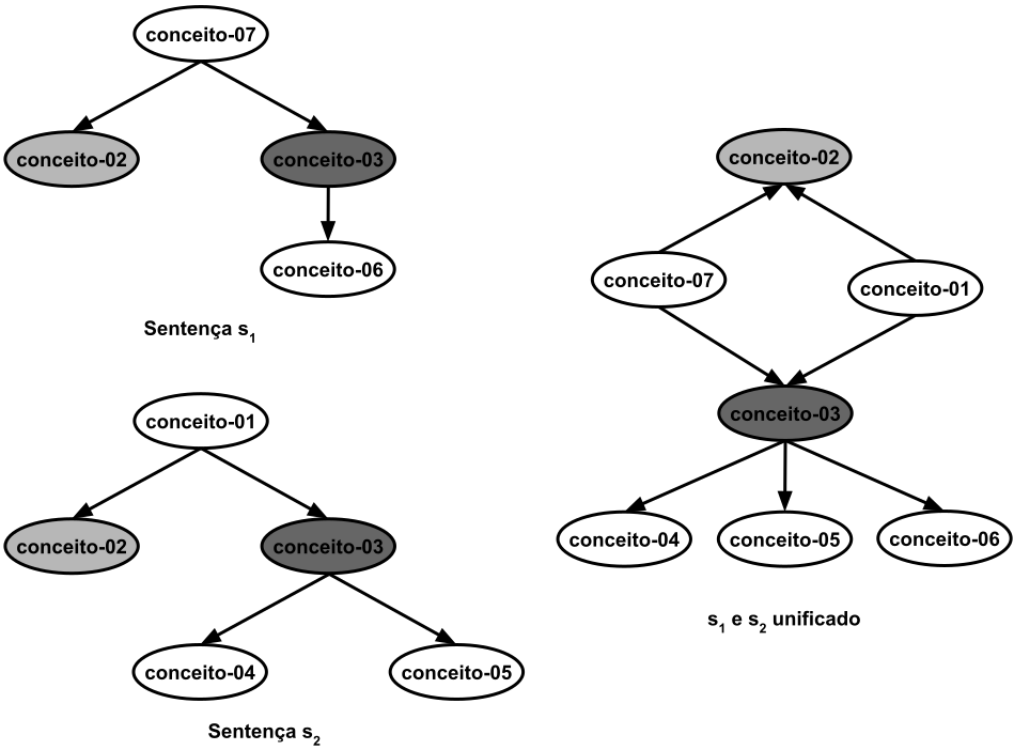


Figura B.3 – Grafos AMR das sentenças s_1 e s_2, e o grafo unificado

Finalizado as multiplicações do valores inicias na matriz sentença-conceito A através do *PageRank*, a fatorização é executada sobre a ela por meio do Decomposição de Valores Singulares gerando a representação das sentenças no espaço latente. Em posse dos vetores no espaço latente, a função descrita em 4.1 é aplicada sobre eles antes de serem submetidos ao classificador SVM para detectar se há paráfrase ou não entre o par de sentenças.

Mais uma vez, como na abordagem descrita na seção 4.3, o método proposto por ISSA *et al.* (2018) limita-se a sentenças não enxergando o documento por inteiro e as relações estabelecidas entre as suas partes, essa restrição é oriunda por conta

da RAS lidar apenas com sentenças. Esse fato impede de capturar a característica estrutural do documento.

B.3 Representação utilizando AMR

Já apresentada na seção B.2, a abordagem de ISSA *et al.* (2018) é a atual estado da arte para detecção de paráfrase no *Microsoft Research Paraphrase Corpus* (DOLAN *et al.*, 2004) e utiliza a estrutura em grafos para representar as sentenças a fim de capturar características léxica, sintática e semântica. Uma das propostas desse trabalho é adaptar o método utilizado por ISSA *et al.* (2018) e avaliar como comporta-se quando aplicado a detecção de plágio de paráfrase em documentos.

Deseja-se comparar os documentos d_1 e d_s utilizando a abordagem de ISSA *et al.* (2018), a qual faz uso do AMR, do *PageRank* e da fatorização de matrizes (SVD) para detectar plágio de paráfrase no espaço semântico de análise (LSA). Documentos comumente são estruturados em níveis acima do nível de sentença, como combinações de sentenças (linhas), parágrafos e o texto por inteiro. Como o AMR lida apenas com sentenças, torna-se necessário dividir os documentos neste nível de detalhamento.

Dados os documentos d_1 e d_s que são compostos por texto bruto, para criar as suas representações com a estrutura em grafo através do AMR é necessário "granularizar"o texto em sentenças. Toma-se como exemplo o trecho de texto "*Meet the Announcers*"(MANN & THOMPSON, 1988), utilizado na seção 3.2.5.2:

> *P. M. has been with KUSC longer than any other staff member. While attending Occidental College, where he majored in philosophy, he volunteered to work at the station as a classical music announcer. That was in 1970.* (B.6)

Ao dividir esse trecho em sentenças, é criado um conjunto S com três elementos como a seguir:

- s_1 = *P. M. has been with KUSC longerthan any other staff member.*
- s_2 = *While attending Occidental College, where he majoredin philosophy, he volunteered to work at the station as a classical music announcer.*
- s_3 = *That was in 1970.*

Os documentos d_1 e d_s ao passarem pelo processo de divisão dão origem aos seguintes conjuntos: $S_{d_1} = [s_1, s_2, ..., s_n]$, onde $S_{d_1} \subseteq d_1$; $S_{d_s} = [s_1, s_2, ..., s_m]$, onde $S_{d_s} \subseteq d_s$. Com os documentos "granulazidos"em sentenças, eles estão aptos para

serem submetidos ao AMR *parser*. Ao aplicar o *parser* sobre o conjunto de sentenças de cada documento são produzidos conjunto de grafos $G_{d_1} = [g_1, g_2, ..., g_n]$, e $G_{d_s} = [g_1, g_2, ..., g_m]$, onde G_{d_1} e G_{d_s} detém os grafos referentes as sentenças contidas em S_{d_1} e S_{d_s}, respectivamente.

Como o AMR gera grafos no qual os seus nós são conceitos, o conjunto de grafos dos documentos d_1 e d_s servirão como insumo para a construção do vocabulário, o *bag of concept* (saco de conceitos), $C = [c_1, c_2, ..., c_w]$. Construído o vocabulário e com os conjuntos de grafos das sentenças dos documentos, a matriz documento-conceito A pode ser estruturada e preenchida.

O preenchimento da matriz A é feita por meio da pontuação que cada conceito obtém durante o processo de *PageRank* aplicado ao grafo do qual pertence. Caso algum conceito esteja presente em mais de um grafo, o mesmo terá os seus valores somados dentro da matriz A. A figura B.4 demonstra como a matriz documento-conceito é preenchida assim como a equação B.7.

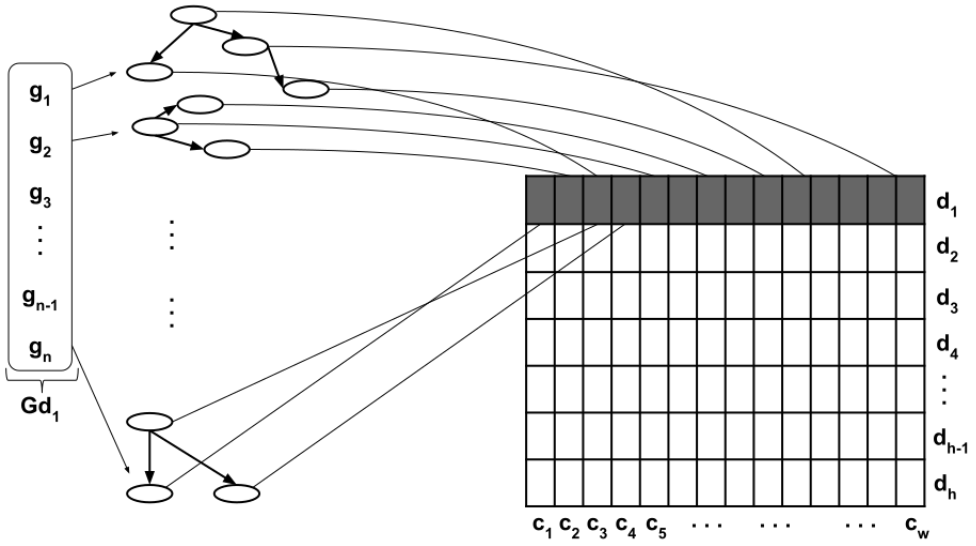


Figura B.4 – Matriz documento-conceito sendo preenchida pela pontuação dos nós dos grafos do documento d_1, pontuação essa obtida através do *PageRank*

$$A_{k,c} = \sum_{g_i \in G_d} PR(c), \tag{B.7}$$

onde: $A_{k,c}$ é a pontuação do conceito c no *k-ésimo* documento; G_d é conjunto de grafos das sentenças pertencentes ao documento d; e $PR(c)$ é o valor da pontuação do conceito c obtido durante o *PageRank*. O valor de $PR(c)$ é condicionado pela

seguinte expressão:

$$PR(c) \ = \begin{cases} PR(c), \ c \in g_i \\ 0, \ c \notin g_i \end{cases} \tag{B.8}$$

Construída a matriz A com as pontuações do *PageRank* de cada conceito, o próximo passo é a união dos grafos que tenham um ou mais conceitos em comum, analogamente ao explicado na seção B.2, no entanto com a diferença, que torna necessário executar a operação de junção para todo o conjunto de grafos de G_{d1} e G_{ds}, ou seja, produto cartesiano condicionado pelo grafos que compartilham o mesmo conceito. Ao fim dessa operação de junção entre os grafos é gerado o conjunto $M_{d_1 d_s} = [g_{u_1}, g_{u_2}, ..., g_{u_p}]$ que contém todos os grafos unificados de G_{d1} e G_{ds}, como demonstra a figura B.5. A quantidade de grafos unificados em $|M_{d_1 d_s}| = p$, onde $p \leqslant |G_{d1}| \times |G_{ds}|$.

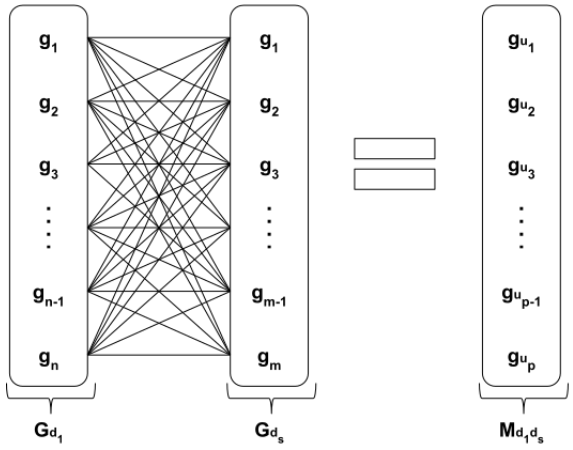


Figura B.5 – Operação de junção cartesiana entre G_{d1} e G_{ds}

O cada grafo g_{u_i} pertencente à $M_{d_1 d_s}$ é submetido ao *PageRank* para, com a pontuação recebida para cada conceito no grafo, reajustar os valores da matriz documento-conceito A como apresenta a equação B.9.

$$A_{k,c} = \sum_{g_{u_i} \in M_{d_1 d_s}} PR(c), \tag{B.9}$$

onde $c \in g_{u_i}$ e conceito no *k-ésimo* documento.

Após os reajustes dos valores na matriz A, por meio do grafos unificados, a matriz documento-conceito é fatorizada para obter as representação dos documentos no espaço latente semântico. Com os vetores de características dos documentos, antes de serem submetidos para o classificador, a função $\vec{f}(\vec{v}_{d_1}, \vec{v}_{d_s}) = [\vec{v}_{d_1} + \vec{v}_{d_s}; |\vec{v}_{d_1} - \vec{v}_{d_s}|]$, onde $\vec{f}(\vec{v}_{d_1}, \vec{v}_{d_s})$ é a concatenação da soma de $\vec{v}_{d_1} + \vec{v}_{d_s}$ e o valor absoluto de $|\vec{v}_{d_1} - \vec{v}_{d_s}|$, é aplicada sobre os vetores dos documentos d_1 e d_s dando origem ao insumo para o modelo SVM que fará a classificação de plágio de paráfrase.

Observando a abordagem proposta nesse seção, vale ressaltar, que, embora o documento esteja representado pelo conjunto da estrutura em grafo das sentenças, a hierarquia entre as partes do texto não é assimilada por conta da disjunção entre os grafos do mesmo documento, isso ocorre por não haver nós de junção entre os grafos do mesmo conjunto. Essa disjunção entre os grafos do mesmo documento, pode acarretar na perda de informações semânticas mais abrangente do documento quando mapeado para o espaço latente semântico, consequentemente influenciando na sua eficácia para a tarefa de detecção de plágio de paráfrase.

Apêndice C

Configurações do parâmetros dos Classificadores

Naive Bayer	
var_smoothing	1,00E-09
Logistic Regression	
penalty	l2
dual	FALSE
tol	1,00E-04
C	1
fit_intercept	TRUE
intercept_scaling	1
class_weight	None
random_state	None
solver	liblinear
max_iter	100
multi_class	ovr
verbose	0
warm_start	FALSE
n_jobs	None
l1_ratio	None

Tabela C.1 – Configurações dos parâmetros do *Naive Bayes* e do *Logistic Regression*

K-Nearst Neighbors	
n_neighbors	5
weights	uniform
algorithm	-
leaf_size	30
p	2
metric	minkowski
metric_params	None
n_jobs	None

Tabela C.2 – Configurações dos parâmetros do *K-Nearst Neighbors*

SVM	
C	1
kernel	rbf
degree	3
gamma	auto
coef0	0
shrinking	TRUE
probability	FALSE
tol	1,00E-03
cache_size	-
class_weight	-
verbose	FALSE
max_iter	-1
decision_function_shape	ovr
random_state	None

Tabela C.3 – Configurações dos parâmetros do SVM

Decision Tree	
criterion	gini
splitter	best
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0
max_features	None
random_state	None
max_leaf_nodes	None
min_impurity_decrease	0
min_impurity_split	1,00E-07
class_weight	None
presort	FALSE

Tabela C.4 – Configurações dos parâmetros do *Decision Tree*