

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ALINHAMENTO TEXTO-IMAGEM EM SITES DE NOTÍCIAS

WELLINGTON C. VELTRONI

ORIENTADOR: PROFA. DRA. HELENA DE MEDEIROS CASELI

São Carlos – SP

Janeiro/2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ALINHAMENTO TEXTO-IMAGEM EM SITES DE NOTÍCIAS

WELLINGTON C. VELTRONI

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Inteligência artificial

Orientador: Profa. Dra. Helena de Medeiros Caseli

São Carlos – SP

Janeiro/2018



Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Wellington Cristiano Veltroni, realizada em 02/03/2018:

Helena de M. Caseli

Profa. Dra. Helena de Medeiros Caseli
UFSCar

Ricardo Cerri

Prof. Dr. Ricardo Cerri
UFSCar

Profa. Dra. Vladia Celia Monteiro Pinheiro
UNIFOR

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Vladia Celia Monteiro Pinheiro e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Helena de M. Caseli

“Dedico este trabalho a Deus e a Virgem Maria por me suportarem nos momentos de dor e angústia e pela convivência nos momentos bons.”

AGRADECIMENTOS

Gostaria de agradecer à minha família por sempre apoiar os meus estudos e confiar na minha escolha de abrir mão de um emprego, em busca do meu sonho de ser professor.

Agradecendo também a minha futura esposa Carla Rompa, que estava comigo lado a lado em todas as dificuldades e não deixou que eu desistisse. Outra pessoa importante foi o meu grande amigo Padre Bruno, Mestre em Letras e grande incentivador do meu mestrado envolvendo Linguística e Inteligência Artificial. Não menos importante no sucesso desse trabalho, está a minha família Filhos da Cruz. Sem o ombro amigo e as orações, o caminho seria muito mais difícil.

Gostaria de agradecer de forma especial a minha orientadora Helena Caseli, que fez um grandioso trabalho em auxiliar um aluno com vivência na área técnica em um aluno que começou a se aprimorar na área acadêmica. Agradeço pela ajuda, puxões de orelha e principalmente o conhecimento adquirido.

Por fim agradeço a CNPq, que forneceu uma bolsa de estudos para o desenvolvimento do meu mestrado, à FAPESP (Projeto MMeaning: 2016/13002-0) e à empresa Gryfo que me incentivou e me apoiou nessa trajetória.

RESUMO

O alinhamento texto-imagem é a tarefa de alinhar elementos presentes em um texto com elementos presentes na imagem que o acompanha. Neste trabalho, o alinhamento texto-imagem foi aplicado em sites de notícias. Muitas notícias não deixam clara para o leitor a correspondência entre elementos do texto e elementos contidos na imagem associada. Nesse cenário, o alinhamento texto-imagem surge com a intenção de orientar o leitor, trazendo clareza para a notícia e a imagem associada uma vez que explicita a correspondência direta entre regiões da imagem e palavras (ou entidades) no texto. O objetivo deste trabalho é combinar técnicas de Processamento de Linguagem Natural (PLN) e Visão Computacional (VC) para gerar um alinhador texto-imagem para notícias: o alinhador LinkPICS. O LinkPICS utiliza a rede convolucional (CNN) YOLO para detectar pessoas e objetos na imagem associada ao texto da notícia. Devido à limitação do número de objetos detectados pela YOLO (80 classes de objetos), optou-se também pela utilização de outras três CNNs para a geração de novos rótulos para objetos. Neste trabalho, o alinhamento texto-imagem foi dividido em dois processos distintos: (1) o alinhamento de pessoas e (2) o alinhamento de objetos. No alinhamento de pessoas, as entidades nomeadas são alinhadas com imagens de pessoas e na avaliação realizada no cópulo de notícias da Folha de São Paulo Internacional, em inglês, obteve-se uma precisão de 98%. No alinhamento de objetos, as palavras físicas são alinhadas com objetos (ou animais, frutas, etc.) contidos na imagem associada à notícia e na avaliação realizada no cópulo de notícias da BBC NEWS, também em inglês, obteve-se uma precisão de 72%. As principais contribuições deste trabalho são o alinhador LinkPICS e a estratégia proposta para sua implementação, que representam inovações para as áreas de PLN e VC. Além destas, outra contribuição deste trabalho é a possibilidade de geração de um dicionário visual (palavras associadas a imagens) contendo pessoas e objetos alinhados, que poderá ser utilizado em outras pesquisas e aplicações como, por exemplo, no auxílio ao aprendizado de outro idioma.

Palavras-chave: alinhamento, texto-imagem, imagem-texto, anotação de imagem, aprendizado visual, dicionário visual

ABSTRACT

Text-image alignment is the task of aligning elements in a text with elements in the image accompanying it. In this work the text-image alignment was applied in news sites. A lot of news do not make clear the correspondence between elements of a text and elements within the associated image. In this scenario, text-image alignment arises with the intention of guiding the reader, bringing clarity to the news and associated image since it explicitly explains the direct correspondence between regions of the image and words (or named entities) in the text. The goal of this work is to combine Natural Language Processing (NLP) and Computer Vision (CV) techniques to generate a text-image alignment for news: the LinkPICS aligner. LinkPICS uses the YOLO convolutional network (CNN) to detect people and objects in the image associated with the news text. Due to the limitation of the number of objects detected by YOLO (only 80 classes), we decided to use three other CNNs to generate new labels for detected objects. In this work, the text-image alignment was divided into two distinct processes: (1) people alignment and (2) objects alignment. In people alignment, the named entities identified in the text are aligned with images of people. In the evaluation performed with the Folha de São Paulo International news corpus, in English, LinkPICS obtained an accuracy of 98% precision. For the objects alignment, the physical words are aligned with objects (or animals, fruits, etc.) present in the image associated with the news. In the evaluation performed with the news corpus of BBC NEWS, also in English, LinkPICS achieved 72% precision. The main contributions of this work are the LinkPICS aligner and the proposed strategy for its implementation, which represent innovations for the NLP and CV areas. In addition to these, another contribution of this work is the possibility of generating a visual dictionary (words associated with images) containing people and objects aligned, which can be used in other researches and applications such as helping to learn a second language.

Keywords: alignment, text-image, image-text, image annotation, visual learning, visual dictionary

LISTA DE FIGURAS

2.1	Exemplo de texto, em inglês, antes e depois da lematização utilizando a ferramenta TreeTagger (SCHMID, 2013).	6
2.2	Diferentes funções da palavra “canto”.	6
2.3	Exemplo de texto, em inglês, antes e depois da etiquetagem morfosintática obtido usando o site http://parts-of-speech.info/	7
2.4	Exemplo de texto, em inglês, antes e depois do reconhecimento de entidades nomeadas reconhecidas pela ferramenta <i>Stanford Named Entity Recognizer</i> (FINKEL; GRENAGER; MANNING, 2005)	8
2.5	Pesquisa na WordNet, por duas palavras sinônimas entre si.	10
2.6	Pesquisa pela palavra “andar” no TeP2.0.	11
2.7	Estrutura de uma rede tradicional (à esquerda) e de uma rede convolucional (à direita).	14
2.8	Resultado apresentado em (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) para a classificação de imagens utilizando CNNs.	15
2.9	Exemplo de funcionamento da rede convolucional YOLO (REDMON et al., 2016).	16
2.10	Exemplo de detecção da rede convolucional YOLO (REDMON et al., 2016).	18
2.11	Exemplos de retângulos aplicados nas imagens para extração de características no método de detecção de faces Haar-like-cascade (VIOLA; JONES, 2001).	19
2.12	Exemplos de boas características para detecção de faces no Haar-like-cascade (VIOLA; JONES, 2001).	19
2.13	Exemplos de imagens com faces detectadas pelo algoritmo Haar-like-cascade (VIOLA; JONES, 2001).	20
2.14	Exemplo de um trecho da ImageNet.	21

2.15	Exemplo de imagens da ImageNet com <i>bounding boxes</i> (regiões delimitadas por retângulos amarelos).	21
2.16	Exemplo de classificação TOP-1 considerada correta. Note que a imagem contém um pinguim e o classificador atribuiu a palavra “ <i>penguin</i> ” como o rótulo mais provável.	24
2.17	Exemplo de classificação TOP-5 considerada correta. A imagem contém um submarino. Apesar da palavra “ <i>submarine</i> ” não ser o rótulo melhor classificado, ela está presente entre os cinco melhores rótulos significando que a classificação está correta.	24
3.1	Exemplo de alinhamento de texto manuscrito com sua respectiva transcrição.	26
3.2	Ferramenta para localização de texto. O usuário fornece o texto desejado e o sistema localiza a imagem do texto manuscrito correspondente ao texto pesquisado.	27
3.3	Etapas do alinhamento de textos manuscritos.	27
3.4	Exemplo de binarização. Os caracteres de texto são pintados de preto e o resto é pintado de branco, permanecendo na imagem apenas os objetos de interesse.	28
3.5	Segmentação das linhas.	29
3.6	Resultado da transformação Hough. Apesar das linhas não serem retilíneas, as linhas do texto manuscrito foram segmentadas corretamente	29
3.7	Alinhamento entre as palavras do texto manuscrito e palavras da transcrição.	30
3.8	Exemplo de alinhamento proposto. Note que não é necessário o uso de transcrições com o mesmo número de linhas do texto manuscrito.	31
3.9	Texto manuscrito digitalizado e texto manuscrito com linhas identificadas.	32
3.10	Sistema completo de geração de anotação proposto em (CHOI; KIM, 2012).	34
3.11	Processo de extração de elementos (substantivos e nomes próprios) significativos do texto no método de Choi e Kim (2012)	34
3.12	Anotações de imagem geradas pelo sistema proposto por Choi e Kim (2012).	35
3.13	Exemplo de imagens do banco de dados Corel segmentadas em regiões e suas respectivas anotações (palavras-chave).	36
3.14	Resultados bons gerados pelo sistema proposto por Duygulu et al. (2002)	38

3.15 Resultados não satisfatórios gerados pelo sistema proposto por Duygulu et al. (2002)	38
3.16 Exemplo de atuação do contexto considerada em (SOCHER; FEI-FEI, 2010).	39
3.17 A concatenação dos grupos mais similares gera as “palavras visuais”. Nesse exemplo, os mais similares formam a palavra visual 4-1-6-7.	40
3.18 Características e grafo utilizados no mapeamento entre as palavras visuais e textuais em (SOCHER; FEI-FEI, 2010).	41
3.19 Anotação de uma imagem a partir do sistema proposto em (SOCHER; FEI-FEI, 2010).	42
3.20 Resultados bons gerados pelo sistema de (SOCHER; FEI-FEI, 2010).	43
3.21 Resultados não satisfatórios gerados pelo sistema de (SOCHER; FEI-FEI, 2010).	44
3.22 Estrutura de árvore gerada pela RNN em (SOCHER et al., 2011).	45
3.23 Resultados da segmentação e anotação do sistema de Socher et al. (2011). Cada <i>pixel</i> foi pintado com a cor correspondente à sua classe.	47
3.24 Exemplo de um artigo (composto de imagem e texto) presente no córpuz utilizado em (TIRILLY et al., 2010).	48
3.25 Fórmulas utilizadas para o cálculo da pontuação das entidades nomeadas em (TIRILLY et al., 2010).	49
3.26 Arquitetura do sistema proposto por Noel e Peterson (2013).	50
3.27 Exemplo de uma imagem e as respectivas anotações geradas pelo sistema proposto em (NOEL; PETERSON, 2013) e pelo ALIPR (LI; WANG, 2008).	52
3.28 Exemplo de uma imagem acompanhada de um texto e todas as possibilidades de alinhamento texto-imagem de acordo com as três abordagens propostas em (PHAM; MOENS; TUYTELAARS, 2008).	54
3.29 Valores de precisão para as três abordagens propostas (para a primeira iteração do algoritmo EM e para o EM completo).	56
3.30 Arquitetura do sistema proposto para alinhamento texto-imagem em (TEGEN et al., 2014).	57
3.31 Alguns resultados de classificação do sistema proposto em (TEGEN et al., 2014) e o respectivo valor para o índice de Jaccard (1901)	59

4.1	Arquitetura do LinkPICS. O alinhador é dividido em etapas que lidam com texto e imagem, separadamente.	61
4.2	Elementos extraídos da notícia de jornal: (1) título da notícia, (2) texto da notícia, (3) imagem associada ao texto e (4) legenda da imagem.	62
4.3	Etapas de processamento da imagem do alinhador LinkPICS.	62
4.4	Exemplo de aplicação da YOLO na detecção de bicicletas e pessoas.	63
4.5	Exemplo de aplicação do YOLO. Nessa imagem o objeto “tocha olímpica” foi detectado com o rótulo “taco de baseball”.	64
4.6	Funcionamento da biblioteca DLIB para reconhecimento facial.	65
4.7	Etapas de processamento de texto do alinhador LinkPICS.	67
4.8	Exemplo de palavras físicas e entidades nomeadas extraídas da notícia da Figura 4.2 após o processamento do texto.	68
4.9	Estrutura do alinhador de pessoas.	68
4.10	Estrutura do alinhador de pessoas utilizando o banco de imagens do Google (à esquerda) e o banco LinkPICS (à direita).	69
4.11	Alinhador de objetos utilizando os rótulos obtidos do processamento da imagem e as palavras do processamento do texto.	70
4.12	Estrutura do alinhador de objetos utilizando a similaridade WUP (à esquerda) e as <i>word embeddings</i> do GloVe (à direita).	71
4.13	Saída do alinhador LinkPICS para a notícia de entrada apresentada na Figura 4.2. As palavras alinhadas da notícia são destacadas na mesma cor que a <i>bounding box</i> correspondente na imagem.	72
5.1	Exemplo de notícia extraída do cópulo BBC contendo: (1) título da notícia, (2) texto da notícia, (3) imagem associada ao texto e (4) legenda da imagem.	74
5.2	Trecho da lista de palavra de visuais gerada neste trabalho.	75
5.3	Interface web da plataforma construída para a avaliação dos alinhadores <i>baseline</i> e LinkPICS.	77

5.4	Exemplos de objetos que nunca poderiam ser alinhados por não serem mencionados nas notícias presente no córpus da Folha de São Paulo Internacional. Em (a) temos uma cadeira e uma mochila; em (b), o logotipo da empresa; em (c), uma taça de vidro e, em (d), trecho de um documento.	80
5.5	Exemplos de alinhamentos beneficiados pela estratégia de utilização das CNNs em conjunto com a YOLO. Esses objetos foram alinhados corretamente porque o rótulo atribuído pela YOLO (especificado em amarelo em cada imagem) foi expandido pelas outras redes.	81
6.1	Texto original e texto após a resolução de anáfora.	86
6.2	Exemplo de um dicionário visual contendo a imagem e termos associados. . . .	87

LISTA DE TABELAS

2.1	Resultados de um suposto algoritmo de anotação de imagens.	23
3.1	Comparação dos resultados de anotação de imagem do sistema de (SOCHER; FEI-FEI, 2010) e os sistemas Alipr (LI; WANG, 2003), Corr LDA (BLEI; JORDAN, 2003) e Total Scene (LI; SOCHER; FEI-FEI, 2009)	42
3.2	Comparação dos resultados de segmentação de imagem do sistema de (SOCHER; FEI-FEI, 2010) e os sistemas (CAO; FEI-FEI, 2007) e Total Scene (LI; SOCHER; FEI-FEI, 2009)	43
3.3	Comparação dos valores de precisão obtidos pelo sistema de RNN proposto em (SOCHER et al., 2011) e os sistemas do estado da arte (TIGHE; LAZEBNIK, 2010) (TL) e (GOULD; FULTON; KOLLER, 2009).	46
3.4	Resultado dos testes por categoria para o sistema de (NOEL; PETERSON, 2013).	52
5.1	Comparação dos resultados para o alinhamento de pessoas utilizando o alinhador <i>baseline</i> e o LinkPICS.	78
5.2	Média de entidades nomeadas e regiões contendo pessoas para o corpúsculo de notícias Folha Internacional.	79
5.3	Resultado da avaliação do alinhamento de objetos do alinhador <i>baseline</i> e do LinkPICS utilizando WUP e distância euclidiana calculada para <i>word embeddings</i> (WE).	80
5.4	Resultado da avaliação do alinhamento de objetos usando o alinhador LinkPICS com WUP e WE com limiares para o alinhamento.	81
5.5	Precisão do alinhador LinkPICS utilizando as palavras visuais e a precisão do alinhador sem o uso das palavras visuais.	82

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO	1
1.1 Objetivos	2
1.2 Motivação	3
1.3 Organização do texto	3
CAPÍTULO 2 – FUNDAMENTAÇÃO TEÓRICA	4
2.1 Processamento de Língua Natural	4
2.1.1 Lematização	5
2.1.2 Etiquetagem morfosintática	6
2.1.3 Reconhecimento de entidades nomeadas	7
2.1.4 Unificação de sinônimos	9
2.1.5 Cálculo de similaridade lexical	11
2.1.5.1 Similaridade WUP	11
2.1.5.2 Distância Euclidiana com <i>word embeddings</i>	12
2.2 Visão Computacional	13
2.2.1 Rede Neural Convolucional	13
2.2.1.1 Rede YOLO	15
2.2.2 Detecção de faces	17
2.2.3 ImageNet	20
2.3 Medidas de avaliação	22

2.3.1	Precisão, Cobertura e Medida-F	22
2.3.2	Classificação TOP-1 e TOP-5	23
CAPÍTULO 3 – TRABALHOS RELACIONADOS		25
3.1	Alinhamento de imagem e texto manuscrito	26
3.1.1	Stamatopoulos, Louloudis e Gatos (2010)	29
3.1.2	Schmidt (2014)	30
3.2	Anotação de imagem	32
3.2.1	Anotação de imagem usando apenas técnicas de PLN	33
3.2.1.1	Choi e Kim (2012)	33
3.2.1.2	Tiwari e Kamde (2015)	35
3.2.2	Anotação de imagem usando técnicas de PLN e de VC	36
3.2.2.1	Duygulu et al. (2002)	36
3.2.2.2	Socher e Fei-Fei (2010)	37
3.2.2.3	Socher et al. (2011)	44
3.2.2.4	Tirilly et al. (2010)	46
3.2.2.5	Noel e Peterson (2013)	49
3.3	Alinhamento texto-imagem	53
3.3.1	Pham, Moens e Tuytelaars (2008)	53
3.3.2	Tegen et al. (2014)	56
CAPÍTULO 4 – O ALINHADOR TEXTO-IMAGEM LINKPICS		60
4.1	Arquitetura do LinkPICS	60
4.1.1	Extração da notícia	60
4.1.2	Processamento de imagem	61
4.1.2.1	Detecção de pessoas e objetos usando a YOLO	61
4.1.2.2	Detecção de pessoas e Reconhecimento Facial	63

4.1.2.3	Detecção de objetos	64
4.1.3	Processamento de texto	66
4.1.4	Alinhador de pessoas	67
4.1.5	Alinhador de objetos	70
4.2	Saída do Alinhador	71
CAPÍTULO 5 – EXPERIMENTOS E AVALIAÇÃO		73
5.1	Recursos	73
5.1.1	Cópus	73
5.1.2	WordNet	74
5.1.3	Palavras Visuais	75
5.2	Alinhador <i>baseline</i>	75
5.3	Plataforma de Avaliação	77
5.4	Avaliação do alinhamento de pessoas	78
5.5	Avaliação do alinhamento de objetos	79
CAPÍTULO 6 – CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS		83
6.1	Trabalhos Futuros	85
6.1.1	Melhorias para o alinhador LinkPICS	85
6.1.2	Criação de um Dicionário Visual	86
6.1.3	Criação de uma ferramenta para coleta automática de imagens	87
6.2	Contribuições	87
REFERÊNCIAS		89

Capítulo 1

INTRODUÇÃO

O alinhamento de texto-imagem é a tarefa de alinhar elementos presentes em um texto com elementos presentes na imagem que o acompanha. Tradicionalmente, o alinhamento textual é aplicado em textos paralelos¹ para encontrar as correspondências entre as palavras (ou sentenças) e suas traduções, por exemplo o alinhador de palavras GIZA++ (OCH; NEY, 2003). No domínio visual, o alinhamento surgiu a partir da necessidade da preservação e do entendimento de textos manuscritos históricos, por meio da digitalização e transcrição dos manuscritos seguidas do alinhamento da imagem digitalizada com o texto transcrito (ZINGER; NERBONNE; SCHOMAKER, 2009; STAMATOPOULOS; LOULLOUDIS; GATOS, 2010; FISCHER et al., 2011; LEYDIER et al., 2014; YIN; WANG; LIU, 2013).

Os textos manuscritos são de grande importância histórica para a sociedade, pois armazenam informações valiosas a respeito de costumes e práticas do passado, que contribuem para o entendimento dos costumes e práticas atuais. Os textos manuscritos são, então, digitalizados com o intuito de preservação. Devido a muitos textos manuscritos serem de difícil interpretação, alguns especialistas criam transcrições correspondentes a esses textos. Essas transcrições são textos em formato ASCII e precisam ser alinhadas com o manuscrito digitalizado.

A maioria dos trabalhos propostos na literatura com o objetivo de alinhar os textos manuscritos com as transcrições aplicam técnicas probabilísticas e de Visão Computacional (VC) na tentativa de fazer o alinhamento. Alguns trabalhos conquistaram um desempenho superior a 90% (YIN; WANG; LIU, 2013; STAMATOPOULOS; LOULLOUDIS; GATOS, 2010; SCHMIDT, 2014; FISHER, 1936), uma taxa bastante significativa.

Outra aplicação bastante relacionada à de alinhamento texto-imagem é a de anotação de imagens. Neste caso, as imagens são anotadas com palavras-chave mas, diferentemente do que

¹Textos paralelos são textos escritos em um idioma acompanhados de suas traduções para outro idioma.

ocorre no alinhamento texto-imagem, na anotação não fica claro a qual região da imagem uma palavra utilizada para anotá-la se refere. Os trabalhos de anotação de imagens (DESCHACHT; MOENS et al., 2007; DUYGULU et al., 2002; SOCHER; FEI-FEI, 2010; RAMISA et al., 2016; NOEL; PETERSON, 2013; TEGEN et al., 2014; PHAM; MOENS; TUYTELAARS, 2008; CHOI; KIM, 2012; TIWARI; KAMDE, 2015; TIRILLY et al., 2010) utilizam técnicas de Processamento de Língua Natural (PLN), de VC e, às vezes, de Aprendizado de Máquina (AM), para encontrar o conjunto de palavras que melhor defina (anote) a imagem em questão. Alguns autores descobriram que a combinação de técnicas de PLN e VC traz benefícios para as aplicações pretendidas, reforçados pela teoria de que o texto contribui para o entendimento do conteúdo da imagem e vice-versa.

Neste trabalho, investiga-se o alinhamento texto-imagem em sites de notícias, nos quais a dificuldade de determinar a melhor correspondência entre uma palavra no texto e uma região da imagem é maior do que nas tarefas relacionadas de alinhamento de textos manuscritos e suas transcrições e de anotação de imagens. Neste caso, acredita-se que o alinhamento texto-imagem é útil para deixar clara a correspondência entre elementos do texto e elementos da imagem. Essa correspondência facilita o entendimento das entidades presentes no texto e na imagem, podendo auxiliar o leitor a compreender o assunto mencionado no texto, bem como a conhecer as pessoas e objetos presentes na imagem.

1.1 **Objetivos**

O objetivo deste trabalho é, portanto, combinar técnicas de PLN e de VC para a geração de um alinhador texto-imagem para notícias *online*.

Neste trabalho, o alinhamento texto-imagem foi dividido em dois processos distintos: (1) o alinhamento de pessoas e (2) o alinhamento de objetos. No alinhamento de pessoas, ocorre o alinhamento de entidades nomeadas com regiões da imagem contendo pessoas e, no alinhamento de objetos, acontece o alinhamento de palavras com objetos contidos na imagem associada à notícia.² Para os experimentos, foram utilizados dois *corpuses* no idioma inglês compostos por: (1) notícias do site Folha de São Paulo Internacional³ e (2) notícias do site BBC News⁴.

²Neste trabalho, considerou-se como objetos: animais, frutas, automóveis, etc.

³Disponível em: <http://www1.folha.uol.com.br/internacional/en/>. Acesso em: 06 fev. 2017.

⁴Disponível em: <http://www.bbc.com/news>. Acesso em: 01 dez. 2017.

1.2 Motivação

As notícias de jornal online são cada dia mais acessadas devido à praticidade da internet. Entretanto, há notícias que não deixam clara para o leitor a correspondência entre elementos do texto e elementos contidos na imagem associada. O entendimento da notícia fica ainda mais comprometido nos casos em que o leitor nunca viu uma imagem referente ao elemento citado no texto. O alinhamento texto-imagem surge com a intenção de orientar o leitor, trazendo clareza para a notícia e a imagem associada.

Como co-produto do alinhamento texto-imagem tem-se a criação de um dicionário visual. Um dicionário visual é composto de pares de imagem e termos associados. O dicionário visual pode auxiliar no aprendizado de outro idioma e pode ser mais útil que um dicionário tradicional contendo apenas palavras, tanto para aplicações manuais como automáticas. Por exemplo, quando traduzimos uma palavra estrangeira com o auxílio de um dicionário ou tradutor automático, nos é informada a tradução da palavra apenas em forma de texto, dificultando a associação e memorização da tradução. Entretanto, aprender uma nova palavra estrangeira visualizando-a em uma imagem é uma ótima alternativa para acelerar e facilitar o aprendizado.

1.3 Organização do texto

O restante deste documento está organizado conforme descrito a seguir.

O Capítulo 2 contém a fundamentação teórica que embasa este trabalho e os trabalhos relacionados. Nesse capítulo, são descritas as técnicas de PLN e de VC, bem como as medidas de avaliação mais utilizadas para medir o desempenho de alinhadores e anotadores de imagem.

O Capítulo 3 descreve alguns dos trabalhos relacionados envolvendo o alinhamento texto-imagem. O capítulo inicia-se contextualizando o alinhamento de textos manuscritos e relatando as principais etapas utilizadas pelos autores para efetuar o alinhamento. Em seguida, são apresentados alguns trabalhos referentes à anotação de imagens e, concluindo o capítulo, são apresentados os trabalhos referentes ao alinhamento texto-imagem que mais se aproximam das propostas contidas neste documento.

O Capítulo 4 descreve o alinhador texto-imagem implementado neste trabalho: o LinkPICS. O Capítulo 5 traz os resultados dos experimentos realizados para avaliar o alinhador proposto. Por fim, o Capítulo 6 fecha esta monografia com algumas considerações finais e propostas de trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentadas as principais técnicas, ferramentas, recursos e medidas utilizadas nos trabalhos relacionados (apresentados no capítulo 3), além de outras que se mostram interessantes para a proposta apresentada neste documento. Inicialmente, na seção 2.1, são descritas as técnicas, ferramentas e medidas de Processamento de Língua Natural utilizadas no processo de: lematização, etiquetagem morfossintática, reconhecimento de entidades nomeadas, unificação de sinônimos e cálculo de similaridade lexical. Em seguida, na seção 2.2, são apresentadas algumas técnicas e um recurso de Visão Computacional, iniciando com a apresentação das redes neurais convolucionais, uma técnica de aprendizado de máquina especializada no tratamento de imagens, seguida da aplicação de detecção de faces e da descrição do banco de imagens ImageNet. Por fim, a seção 2.3 traz as principais medidas de avaliação empregadas nos trabalhos relacionados e que também foram utilizadas neste trabalho.

2.1 Processamento de Língua Natural

A seguir são descritas algumas tarefas e ferramentas de PLN consideradas úteis para o alinhamento texto-imagem, segundo os trabalhos relacionados apresentados no capítulo 3: na seção 2.1.1 é explicado o processo de lematização utilizado no trabalho de Tiwari e Kamde (2015); na seção 2.1.2, a etiquetagem morfossintática aplicada em (CHOI; KIM, 2012; SOCHER; FEI-FEI, 2010); na seção 2.1.3, o reconhecimento de entidades nomeadas realizado por Pham, Moens e Tuytelaars (2008), Tirilly et al. (2010); e na seção 2.1.4, a unificação de sinônimos aplicada por Socher e Fei-Fei (2010), Noel e Peterson (2013). Dessas tarefas, vale comentar que o alinhador texto-imagem LinkPICS realiza a etiquetagem morfossintática e o reconhecimento de entidades nomeadas, como detalhado no Capítulo 4. Por fim, duas medidas de similaridade lexical são apresentadas na seção 2.1.5, as quais foram utilizadas no alinhamento de objetos do

LinkPICS.

2.1.1 Lematização

A lematização consiste em remover o sufixo das palavras transformando-as em sua forma canônica (não flexionada) e foi primeiramente apresentada por Lovins (1968). Com a lematização, palavras em inglês como “*stronger*” e “*strongest*” são convertidas para a forma lematizada “*strong*” (TIWARI; KAMDE, 2015). A lematização permite unificar palavras similares aumentando a abrangência de uma pesquisa. Por exemplo, supondo que o rótulo de imagem contenha a palavra “*strongest*”, uma pesquisa com a palavra “*strong*” não retornaria esse rótulo de uma imagem. Reduzir a palavra para a sua forma canônica permite incluir essa palavra em mais resultados de pesquisas.

As abordagens de lematização mais conhecidas são: simples (baseada em simples consulta a tabelas), de remoção de sufixo, estocásticos e baseados em regras específicas para um idioma (WIKIPEDIA, 2016)

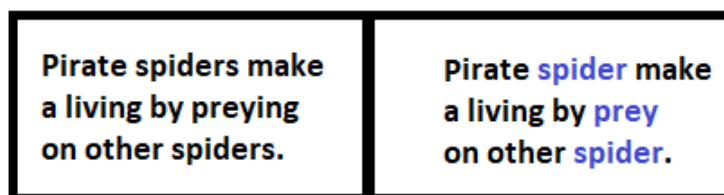
Os algoritmos simples recebem como entrada a palavra flexionada e buscam em uma tabela qual é a forma canônica correspondente. Vantagens desse algoritmo são sua simplicidade e rapidez, mas sua desvantagem é a baixa cobertura uma vez que ele não é capaz de lidar com palavras novas ou que não estejam presentes na tabela.

Os algoritmos de remoção de sufixo trabalham com uma base de regras que indica como será realizada a remoção do sufixo, transformando a palavra em sua forma canônica. A desvantagem desse algoritmo é não lidar com palavras que não entram nas regras gerais, como por exemplo os verbos irregulares do inglês (por exemplo, *ran-run*).

Os algoritmos estocásticos utilizam a probabilidade para reconhecer a forma canônica de uma palavra. O modelo é treinado baseado em exemplos de transformação de uma palavra para a forma canônica. Esse algoritmo possui mais complexidade de implementação e, como é puramente probabilístico, pode apresentar erros.

Por fim, estão os algoritmos que se baseiam em regras definidas para uma linguagem específica e tentam identificar primeiramente a morfologia das palavras. A partir do conhecimento morfológico de uma palavra, podem seguir as regras e definir a forma canônica mais adequada. Esse algoritmo pode cometer erros, principalmente no momento de atribuição morfológica a uma palavra. A Figura 2.1 traz um exemplo de um texto em inglês antes e depois da lematização.

Figura 2.1: Exemplo de texto, em inglês, antes e depois da lematização utilizando a ferramenta TreeTagger (SCHMID, 2013).



2.1.2 Etiquetação morfossintática

Um etiquetador morfossintático é responsável por identificar as categorias das palavras presentes nas sentenças de um texto levando em consideração, entre outros, as funções sintáticas de tais palavras nas sentenças do texto. Essas categorias podem ser, por exemplo: substantivo, adjetivo, verbo, advérbio e artigo.

A identificação da categoria de uma palavra no texto não é uma tarefa simples, uma vez que muitas palavras podem assumir funções variadas, como a palavra “canto” ilustrada na Figura 2.2. No exemplo apresentado nessa figura, a palavra “canto” possui diferentes funções sendo utilizada como verbo na primeira sentença e como substantivo na segunda. Para a resolução desse problema é necessário analisar a sentença inteira e não apenas as palavras individualmente.

Figura 2.2: Diferentes funções da palavra “canto”.

Sentença	Função da palavra
Eu <u>canto</u> na igreja	verbo
O <u>canto</u> direito do carro	substantivo

Assim, não é viável realizar a tarefa por meio de consultas a uma lista de palavras acompanhadas de suas respectivas funções. Os algoritmos mais utilizados para a etiquetação morfossintática são baseados em modelos ocultos de Markov (BAUM; PETRIE, 1966; BAUM; EAGON et al., 1967; BAUM; SELL, 1968; BAUM et al., 1970; BAUM, 1972), algoritmos de programação dinâmica (DEROSE, 1988; CHURCH, 1988) e algoritmos não-supervisionados (BRILL, 1995; GOLDWATER; GRIFFITHS, 2007; GAEL; VLACHOS; GHAHRAMANI, 2009; ZENNAKI; SEMMAR; BESACIER, 2015).

Os algoritmos baseados nos modelos ocultos de Markov utilizam a coocorrência entre categorias de palavras para a construção de uma tabela de probabilidades de sequências. Um exemplo pode ser uma sequência que diz que após um artigo há 70% de chance da palavra seguinte ser um substantivo, 10% de ser um verbo e 20% de ser um adjetivo. Há, também, algoritmos

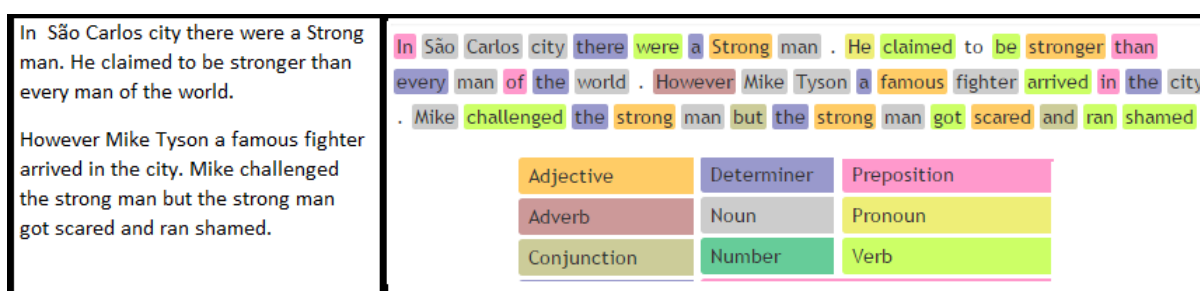
que não se baseiam apenas em pares, mas em triplas de palavras ou sequências maiores.

Os algoritmos de programação dinâmica costumam ser mais rápidos e trabalham com pares e triplas (CHURCH, 1988; DEROSE, 1988). A programação dinâmica consiste em dividir um problema em vários sub-problemas. Quando um sub-problema é resolvido, ele é memorizado e integrado à resolução do problema maior. Esse algoritmo é rápido porque o problema resolvido uma vez não necessita de uma nova resolução pois sua resolução já está memorizada pelo algoritmo.

Os algoritmos não-supervisionados trabalham em cópulas não anotados e tentam aprender a classificar categorias de palavras ou frases por indução. Os aprendizados podem ser a partir de sistemas baseados em regras (BRILL, 1995), modelos bayesianos (GOLDWATER; GRIFFITHS, 2007), modelos estatísticos (GAEL; VLACHOS; GHAHRAMANI, 2009) ou redes neurais (ZENNAKI; SEMMAR; BESACIER, 2015).

A Figura 2.3 traz um exemplo de um texto antes e depois da etiquetagem morfossintática.

Figura 2.3: Exemplo de texto, em inglês, antes e depois da etiquetagem morfossintática obtido usando o site <http://parts-of-speech.info/>



A título de ilustração, etiquetadores morfossintáticos bastante utilizados como o Stanford Part-of-Speech Tagger (TOUTANOVA et al., 2003) e TreeTagger (SCHMID, 2013) apresentam, respectivamente, cerca de 97% e 96% de precisão para o idioma inglês.

2.1.3 Reconhecimento de entidades nomeadas

O reconhecimento de entidades nomeadas (REN) é uma tarefa de PLN que tem o objetivo de detectar e classificar nomes (entidades) em um texto. As entidades podem ser classificadas em categorias como: pessoas, instituições, locais e expressões de tempo (por exemplo: Junho, 2001) (SANG; MEULDER, 2003).

Para proceder com o reconhecimento de entidades nomeadas (ENs) é necessário, primeiro, detectar as candidatas. A etapa de detecção é considerada uma etapa de segmentação de texto

e apresenta algumas dificuldades. Considere, por exemplo, o seguinte trecho extraído de uma notícia de jornal¹:

“Relatório do Banco Central aponta que o uso da manobra cresceu a partir do governo do ex-presidente Luiz Inácio Lula da Silva e disparou no governo de Dilma Rousseff. No final do ano passado, o saldo acumulado com diferentes bancos federais era de cerca de R\$60 bilhões.”

Três entidades nomeadas ocorrem nesse texto: Banco Central, Luiz Inácio Lula da Silva e Dilma Rousseff. A segmentação dessas entidades deve ser realizada com cuidado para não separar, por exemplo, o nome “Luiz Inácio” do sobrenome “da Silva”.

Os algoritmos de REN são, geralmente, baseados em modelos estatísticos, os quais necessitam de uma grande quantidade de dados anotados manualmente para o treinamento. O *Conditional Random Field* (CRF) é o método mais utilizado para o reconhecimento de entidades nomeadas (FINKEL; GRENAGER; MANNING, 2005). Ele é um método probabilístico de predição que calcula a probabilidade de um termo ser uma EN e também sua classificação.

Para a avaliação das abordagens de REN geralmente é utilizada a medida-F, que é a média harmônica entre a precisão e a cobertura (veja seção 2.3). A título de ilustração, tem-se que 93,39% de medida-F foi obtido por um algoritmo (MARSH; PERZANOWSKI, 1998) treinado para a língua inglesa em um corpus de notícia do New York Times, enquanto anotadores humanos chegaram a 97%.

A Figura 2.4 traz um exemplo de texto antes e depois do reconhecimento de entidades nomeadas, as quais aparecem sublinhadas no texto.

Figura 2.4: Exemplo de texto, em inglês, antes e depois do reconhecimento de entidades nomeadas reconhecidas pela ferramenta *Stanford Named Entity Recognizer* (FINKEL; GRENAGER; MANNING, 2005)

In São Carlos city there were a Strong man. He claimed to be stronger than every man of the world. However Mike Tyson a famous fighter arrived in the city. Mike challenged the strong man but the strong man got scared and ran shamed.	In <u>São Carlos</u> city there were a <u>Strong man</u> . He claimed to be stronger than every man of the world. However <u>Mike Tyson</u> a famous fighter arrived in the city. <u>Mike</u> challenged the <u>strong man</u> but the <u>strong man</u> got scared and ran shamed.
---	--

¹Disponível em: <http://noticias.uol.com.br/politica/ultimas-noticias/2016/06/17/banco-do-brasil-diz-ao-senado-que-pedaladas-do-plano-safra-nao-tem-ato-de-dilma.htm>. Acesso em: 29 dez. 2017.

2.1.4 Unificação de sinônimos

A unificação de sinônimos consiste em agrupar e unificar palavras que possuem o mesmo significado. Para a realização dessa tarefa geralmente é utilizada a WordNet (FELLBAUM, 1998).

A WordNet é um banco de dados léxico disponível para diversos idiomas, entre eles o inglês, que agrupa verbos, adjetivos, advérbios e substantivos em conjuntos de sinônimos chamados de *synsets*. Atualmente, a WordNet do inglês possui 117.000 *synsets*² que são interligados com outros *synsets*, formando uma arquitetura em formato de árvore que obedece as relações de hiperonímia e hiponímia.

As relações de hiperonímia e hiponímia podem ser explicadas utilizando a teoria de conjuntos da matemática. Na teoria de conjuntos, pode-se dizer que o conjunto dos números inteiros contém o número 1 assim como pode-se afirmar que o número 1 está contido no conjunto dos números inteiros. De forma semelhante pode-se dizer que a palavra “fruta” é hiperonímia da palavra “banana” pois, fazendo uma analogia com a teoria dos conjuntos, ela contém a palavra banana em seu conjunto. De modo similar, pode-se dizer que a palavra “banana” é hiponímia da palavra “fruta” pois está contida no conjunto representado pela palavra “fruta”.³

Na WordNet, por exemplo, é possível fazer uma consulta fornecendo como parâmetro uma palavra e obtendo como retorno as palavras que possuem relação de sinonímia com a palavra pesquisada. A Figura 2.5 apresenta uma consulta à WordNet⁴ na qual é possível verificar a relação de sinonímia existente entre as palavras “*lady*” e “*dame*”.

A WordNet pode ser acessada para pesquisas através da internet ou integrada a sistemas computacionais a partir de seu download. Além da relação de sinônimos, a WordNet fornece o significado de cada palavra e também traz exemplos do uso da palavra em frases.

A unificação de sinônimos pode auxiliar no cálculo da importância de um termo em determinado texto, pois a unificação permite que todos os sinônimos de um termo sejam contados como o próprio termo.

Além do inglês, há versões de WordNets disponíveis para outros idiomas como: Word-

²Informação retirada do site oficial da Wordnet. Disponível em: <http://wordnet.princeton.edu/>. Acesso em: 29 dez. 2017.

³Disponível em: <http://mundoeducacao.bol.uol.com.br/gramatica/hiperonimia-hiponimia.htm>. Acesso em: 29 dez. 2017.

⁴Pesquisa realizada no site <http://wordnetweb.princeton.edu/perl/webwn>

Figura 2.5: Pesquisa na WordNet, por duas palavras sinônimas entre si.

WordNet Search - 3.1
[- WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
 Display options for sense: (gloss) "an example sentence"

Noun

- [S:](#) (n) **lady** (a polite name for any woman) *"a nice lady at the library helped me"*
- [S:](#) (n) [dame](#), [madam](#), [ma'am](#), [lady](#), [gentlewoman](#) (a woman of refinement) *"a chauffeur opened the door of the limousine for the grand lady"*
- [S:](#) (n) **Lady**, [noblewoman](#), [peeress](#) (a woman of the peerage in Britain)

WordNet Search - 3.1
[- WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
 Display options for sense: (gloss) "an example sentence"

Noun

- [S:](#) (n) **dame**, [doll](#), [wench](#), [skirt](#), [chick](#), [bird](#) (informal terms for a (young) woman)
- [S:](#) (n) **dame**, [madam](#), [ma'am](#), [lady](#), [gentlewoman](#) (a woman of refinement) *"a chauffeur opened the door of the limousine for the grand lady"*

Fonte: <http://wordnetweb.princeton.edu/perl/webwn>

Net:WOLF⁵ e WoNeF⁶ para o francês, TALP⁷ para o catalão e espanhol e TeP2.0⁸, WordNet.Br⁹ e OpenWordnet-PT¹⁰ para o português. A Figura 2.6 ilustra a busca pela palavra “andar” que retorna vários significados para a palavra e também os sinônimos de cada significado, no TeP2.0.

⁵Disponível em: <http://alpage.inria.fr/~sagot/wolf-en.html>. Acesso em: 29 dez. 2017.

⁶Disponível em: <http://wonef.fr/>. Acesso em: 29 dez. 2017.

⁷Disponível em: <http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/99-euro-wordnet>. Acesso em: 29 dez. 2017.

⁸Disponível em: <http://www.nilc.icmc.usp.br/tep2/ajuda.htm>. Acesso em: 28 dez. 2017.

⁹Disponível em: <http://www.nilc.icmc.usp.br/wordnetbr/>. Acesso em: 28 dez. 2017.

¹⁰Disponível em: <http://wnpt.brcloud.com/wn/>. Acesso em: 28 dez. 2017.

Figura 2.6: Pesquisa pela palavra “andar” no TeP2.0.



Fonte: <http://www.nilc.icmc.usp.br/tep2/index.htm>

2.1.5 Cálculo de similaridade lexical

A similaridade lexical indica o grau de similaridade semântica/conceitual entre duas palavras. Para exemplificar, considere as seguintes palavras: “onça”, “gato” e “bola”. As palavras “onça” e “gato” possuem um alto grau de semelhança por representarem animais, enquanto a palavra “bola” que representa um objeto, é distante das outras palavras. Neste trabalho, a similaridade lexical foi utilizada no alinhador de objetos para encontrar as melhores candidatas a alinhamento entre rótulo da imagem e palavras no texto.

A seguir, são descritas as duas medidas de similaridade utilizadas, neste trabalho, para o calcular a proximidade entre duas palavras.

2.1.5.1 Similaridade WUP

A similaridade WUP (WU; PALMER, 1994) se baseia na hierarquia da WordNet para dizer quão similares são duas palavras. Como já mencionado anteriormente, a WordNet é organizada em hierarquias de substantivos e verbos agrupados em *synsets*. Por exemplo, os substantivos “lion” e “tiger” pertencem ao *synset* “feline”, enquanto o substantivo “dog” pertence ao *synset* “canine”. Esses dois *synsets* pertencem ao *synset* “mammals”. Assim, a WordNet organiza as palavras que representam os conceitos de maneira análoga ao que temos no mundo real. A hierarquia da WordNet indica que “lion” e “tiger” são mais similares entre si do que quando comparadas com a palavra “dog”.

A WUP calcula a similaridade entre palavras baseada em uma organização conceitual como mostra a equação (2.1).

$$Sim_{WUP}(c1, c2) = \log \frac{2 * depth(lcs(c1, c2))}{depth(c1) + depth(c2)} \quad (2.1)$$

Dadas duas entidades ($c1$ e $c2$), a similaridade WUP é calculada baseada na profundidade do nó comum entre essas entidades ($depth(lcs(c1, c2))$) e na profundidade de cada entidade ($depth(c1)$ e $depth(c2)$) na WordNet. Os valores da WUP variam entre 0 e 1. O valor 1 significa que as entidades pertencem ao mesmo nó e são similares e valores baixos significam que as palavras são distantes entre si. É importante mencionar que o valor da similaridade nunca será 0 porque a profundidade de um nó da WordNet nunca será 0 (LIU; ZHAO; YU, 2006).

Um exemplo do cálculo da similaridade WUP para os conceitos “lion”, “tiger” e “dog” mencionados anteriormente pode ser visto nas equações 2.2 e 2.3:

$$Sim_{WUP}(lion, tiger) = \log \frac{2 * depth(lcs(lion, tiger))}{depth(lion) + depth(tiger)} = 0,93 \quad (2.2)$$

$$Sim_{WUP}(lion, dog) = \log \frac{2 * depth(lcs(lion, dog))}{depth(lion) + depth(dog)} = 0,82 \quad (2.3)$$

A similaridade WUP foi utilizada por Choi e Kim (2012) para a anotação de imagem, por Choi et al. (2012) na recuperação de imagem, por Krishnamoorthy et al. (2013) na geração de rótulos automáticos e por Choi et al. (2014) na análise de conteúdo terrorista em documentos. Contudo, a abordagem adotada neste trabalho difere desses trabalhos relacionados, pois aqui a similaridade é calculada considerando o rótulo atribuído a uma região da imagem, considerando mais do que apenas a informação no título da notícia e na legenda.

2.1.5.2 Distância Euclidiana com *word embeddings*

As *word embeddings* são representações de palavras em forma de vetores e são construídas baseadas no contexto de ocorrência das palavras nos textos (SHI; LIU, 2014). Recentemente, vários métodos foram propostos para gerar esses vetores, entre eles o Word2Vec (MIKOLOV et al., 2013) e o GloVe (PENNINGTON; SOCHER; MANNING, 2014). O método escolhido para este trabalho foi o GloVe (*Global Vectors for Word Representation*). O GloVe gera vetores baseado na razão da co-ocorrência entre palavras. A partir dessa razão é possível calcular a distância entre pares de *word embeddings* utilizando, por exemplo, a distância Euclidiana como demonstra a equação (2.4).

$$Sim_{WE}(v1, v2) = \sqrt{\sum_{i=1}^n (v1_i - v2_i)^2} \quad (2.4)$$

Dadas duas *word embeddings* ($v1$ e $v2$), a distância Euclidiana é calculada baseada na diferença entre os valores de cada dimensão i desses vetores, onde n é o número total de dimensões do vetor.

Por se tratar de um método novo, o GloVe não foi ainda muito explorado. Karpathy e Fei-Fei (2015) são um dos poucos trabalhos relacionados ao alinhamento texto-imagem que utilizaram o GloVe, e a ideia principal desse trabalho era aprender a correspondência entre imagem e texto, em uma abordagem multimodal.

2.2 Visão Computacional

Nas subseções a seguir são descritas algumas tarefas e ferramentas de Visão Computacional (VC) consideradas úteis para o alinhamento texto-imagem, segundo os trabalhos relacionados apresentados no capítulo 3, bem como este próprio trabalho: na seção 2.2.1, descreve-se uma rede neural convolucional, uma técnica de aprendizado de máquina especializada no tratamento de imagens, com ênfase especial para a rede YOLO (REDMON et al., 2016) e, na seção 2.2.2, apresenta-se o processo de detecção de faces empregado em (TIRILLY et al., 2010; PHAM; MOENS; TUYTELAARS, 2008; NOEL; PETERSON, 2013). Por fim, a seção 2.2.3 descreve o grande banco de imagens ImageNet utilizado neste e em diversos trabalhos relacionados.

2.2.1 Rede Neural Convolucional

As redes neurais convolucionais (CNNs), ao contrário das redes neurais tradicionais, foram criadas para trabalhar especificamente com imagens. Para entender o funcionamento de uma rede convolucional, considere primeiramente a breve descrição do funcionamento das redes neurais tradicionais (ROSENBLATT, 1958) e suas limitações para o tratamento de imagens apresentados a seguir.

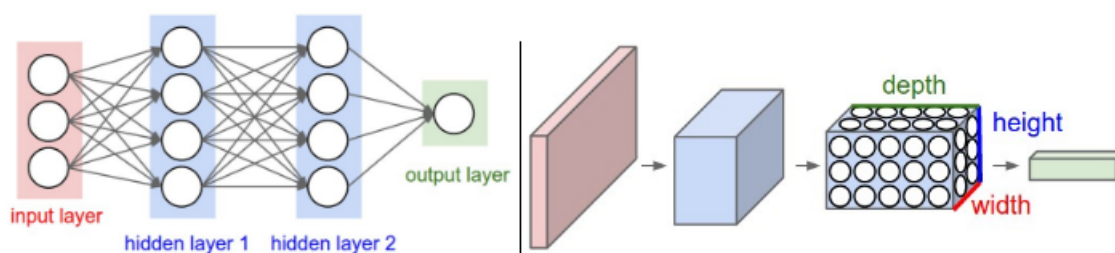
As redes neurais tradicionais recebem um dado como entrada e processam essa entrada por meio de camadas chamadas de camadas escondidas (*hidden layers*). Cada camada possui neurônios conectados a todos os neurônios da camada anterior, que lidam com os dados de entrada gerando pesos que serão propagados para as próximas camadas. Por exemplo, considerando-se uma imagem de entrada com dimensões 32×32 e com 3 canais da cor RGB,

tem-se o total de 3072 ($32 \times 32 \times 3$) pesos para cada neurônio da camada inicial, um número grande de pesos para uma imagem pequena. Uma imagem maior terá um número imenso de pesos, fazendo com que a rede confunda os dados.

Da maneira que é projetada, a rede tradicional recebe os dados de entrada de uma imagem mas é capaz de identificar apenas padrões simples nas imagens por dois motivos: (1) os neurônios de uma camada não compartilham informações com os neurônios da mesma camada e (2) cada neurônio recebe informações aprendidas da imagem inteira. Por exemplo, a rede neural consegue aprender sobre um objeto presente no canto direito de uma imagem mas, caso o objeto mude de posição, a identificação do objeto é prejudicada.

A função das CNNs é receber uma imagem de entrada e processá-la dentro da rede de forma que a última camada da rede gere uma classificação para cada imagem. As CNNs possuem os neurônios nas camadas posicionados em 3 dimensões: altura, largura e profundidade. Os neurônios são conectados apenas a pequenas regiões da imagem, o que facilita o aprendizado de padrões contidos nas regiões da imagem. Outra vantagem é o compartilhamento de informações aprendidas entre os neurônios da mesma camada, permitindo que objetos em diferentes posições na imagem sejam aprendidos pela rede convolucional. A Figura 2.7 ilustra a diferença entre uma rede tradicional e as CNNs.

Figura 2.7: Estrutura de uma rede tradicional (à esquerda) e de uma rede convolucional (à direita).



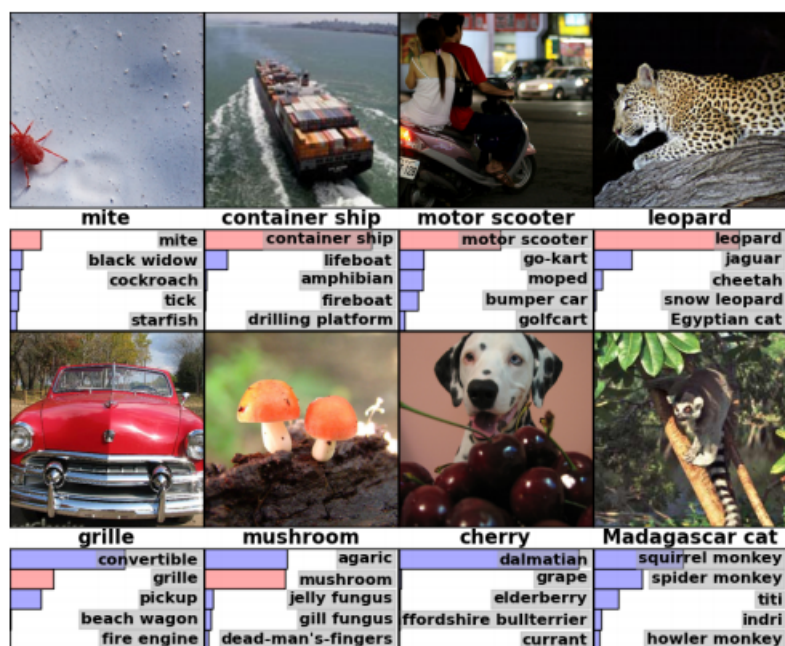
Fonte: <http://cs231n.github.io/convolutional-networks/>

As CNNs apresentam camadas que possuem funções específicas. O objetivo é receber o volume de dados (imagem) e reduzir esse volume até que ele alcance a classificação na última camada. Geralmente as CNNs possuem três camadas (LECUN et al., 1998): (1) camada convolucional, (2) *pooling*, (3) camada inteiramente-conectada. A camada convolucional computa a saída dos neurônios que estão conectados a pequenas regiões da imagem. A camada *pooling* reduz o tamanho da imagem. A camada inteiramente-conectada é responsável por gerar a classificação da imagem de entrada.

A Figura 2.8 traz um exemplo de resultado da aplicação de uma rede convolucional em algumas imagens realizado em (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Cada imagem de

teste possui cinco rótulos mais prováveis de acordo com a CNN. Embaixo das imagens está o rótulo correto e a probabilidade atribuída a esse rótulo é mostrada na barra vermelha quando o rótulo correto está entre os cinco rótulos mais prováveis.

Figura 2.8: Resultado apresentado em (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) para a classificação de imagens utilizando CNNs.



Fonte: (KRIZHEVSKY; SUTSKEVER; HINTON, 2012)

As CNNs vêm se tornando a principal ferramenta para a classificação de imagens. Na literatura, as CNNs alcançaram taxas de erro realmente baixas na classificação de imagens da ImageNet (DENG et al., 2009): 3,57% em (HE et al., 2015) e 7% em (SIMONYAN; ZISSERMAN, 2014).

2.2.1.1 Rede YOLO

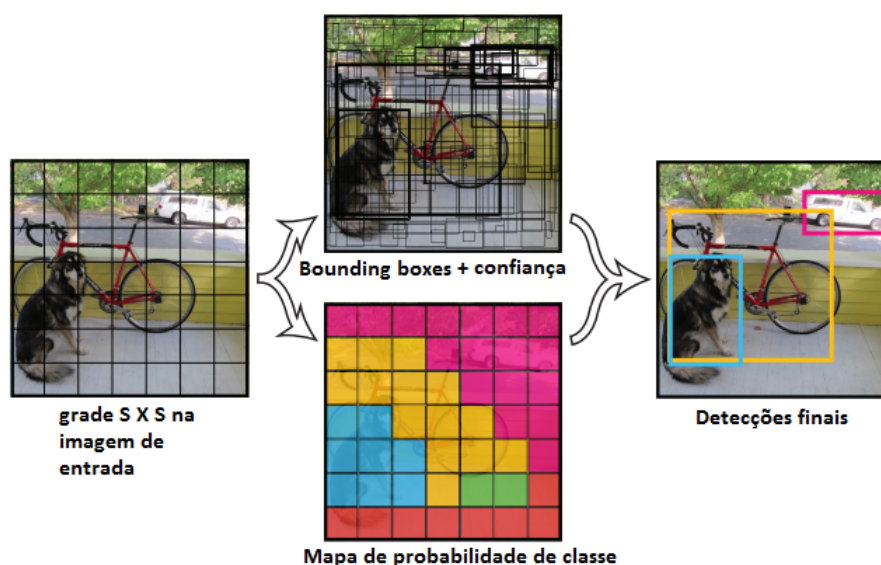
Na área de detecção e localização de objetos, o trabalho de Redmon et al. (2016) apresenta a rede convolucional YOLO que detecta objetos em tempo real. Além da detecção de quais objetos estão presentes em uma imagem, a YOLO também indica a localização de cada objeto através de *bounding boxes*. A YOLO é formada por uma rede neural convolucional inspirada na arquitetura da rede GoogleLeNet (SZEGEDY et al., 2015) e tem as seguintes funções, todas realizadas de forma concorrente, o que agiliza a detecção:

- Extração de características da imagem;
- Predição de *bounding box* para cada objeto;

- Classificação e rotulação de cada *bounding box* com a classe mais provável.

Primeiramente, a YOLO divide a imagem de entrada em uma grade de tamanho $S \times S$. Cada célula da grade prediz no máximo duas *bounding boxes* em conjunto com as probabilidades de cada classe para cada *bounding box* e também o grau de confiança.¹¹ Em seguida, é realizado um procedimento que unifica as *bounding boxes* encontradas de forma que não haja dupla detecção de um mesmo objeto. A Figura 2.9 ilustra o procedimento da YOLO em uma imagem de entrada.

Figura 2.9: Exemplo de funcionamento da rede convolucional YOLO (REDMON et al., 2016).



Fonte: Adaptado de (REDMON et al., 2016)

A YOLO detecta 80 objetos diferentes que são listados a seguir:

- Pessoas;
- Veículos: bicicleta, carro, motocicleta, avião, trem, caminhão, ônibus, barco e balão;
- Animais: pássaro, gato, cachorro, cavalo, ovelha, vaca, elefante, urso, zebra e girafa;
- Objetos relacionados a rua: semáforo, hidrante, parquímetro, sinal de pare, banco;
- Comida: banana, maçã, sanduíche, laranja, brócolis, cenoura, cachorro-quente, pizza, rosquinha, bolo;

¹¹O grau de confiança determina se uma célula possui ou não um objeto e qual a confiabilidade de ser o objeto predito.

- Eletrônicos: monitor de tv, notebook, mouse, teclado, celular, controle remoto, refrigerador, relógio, sanduicheira, forno, micro-ondas;
- Móveis da casa: cadeira, sofá, cama, mesa de jantar, banheiro e pia;
- Objetos esportivos: *frisbee*, bola, prancha de surf, *skate*, raquete de tênis, luva de basebol, taco de basebol, *skis* e *snowboard*;
- Objetos de cozinha: garrafa, taça de vinho, xícara, faca, colher, garfo e tigela;
- Outros objetos: mochila, guarda-chuva, bolsa, gravata, mala, livro, vaso, vaso de planta, tesoura, pasta de dente, urso de pelúcia, secador de cabelo.

Para o treinamento da YOLO, foi utilizado o conjunto de dados de competição da ImageNet (RUSSAKOVSKY et al., 2015) e para a execução do treinamento e inferência dos resultados foi utilizado o *framework* DarkNet (REDMON, 2013). A validação foi aplicada no conjunto de dados PASCAL VOC (EVERINGHAM et al., 2015) e foi comparada com trabalhos do estado da arte (SADEGHI; FORSYTH, 2014; YAN et al., 2014; REN et al., 2015) no quesito velocidade de detecção e também em precisão mAP (*mean average precision*). A precisão mAP consiste em calcular a média de precisão para cada classe que a rede prediz.

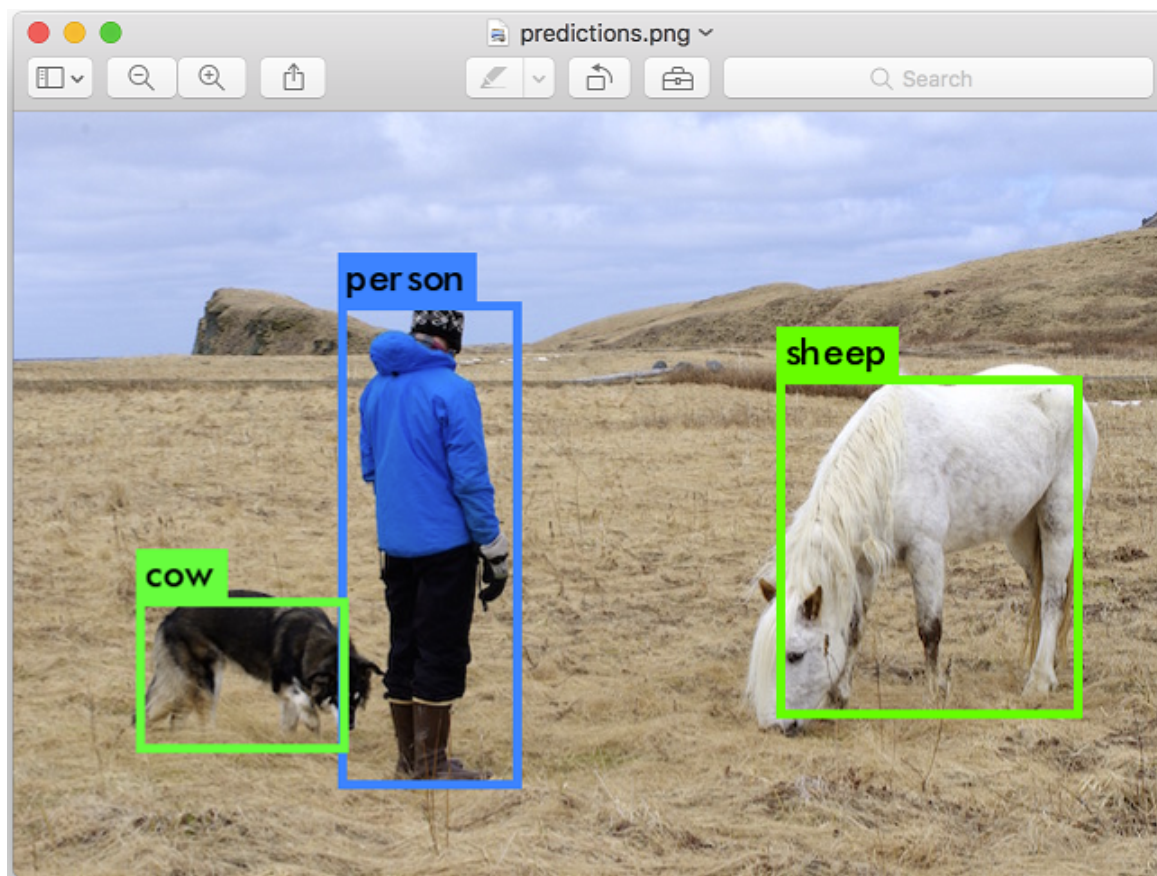
A Yolo superou todos os trabalhos no quesito velocidade de detecção e na precisão mAP alcançou 63,4% contra 73,2% da Faster R-CNN VGG-16 (REN et al., 2015). Apesar de ser superada na precisão mAP, a YOLO faz a detecção aproximadamente 4x mais rápido do que a R-CNN.

Uma limitação da YOLO na tarefa de classificação é que somente são permitidas duas *bounding boxes* por célula. Caso haja vários objetos próximos, e que pertençam à mesma célula, a YOLO não consegue detectar todos. A Figura 2.10 demonstra um exemplo de detecção realizada pela YOLO.

2.2.2 Detecção de faces

Uma das aplicações mais comuns de visão computacional é a detecção de faces. Os detectores de face são responsáveis por localizar diferentes faces em imagens, filtrando apenas as regiões de interesse (faces) e desconsiderando todo o conteúdo irrelevante da imagem. Vários trabalhos (PHAM; MOENS; TUYTELAARS, 2008; TIRILLY et al., 2010; NOEL; PETERSON, 2013) utilizam o algoritmo de detecção de faces Haar-like-cascade (VIOLA; JONES, 2001).

Figura 2.10: Exemplo de detecção da rede convolucional YOLO (REDMON et al., 2016).



Fonte: (REDMON et al., 2016)

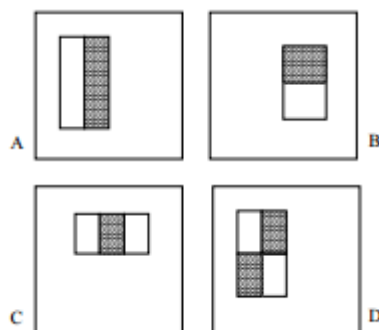
Para o treinamento desse algoritmo, é necessário fornecer um conjunto de imagens contendo faces e um conjunto que não possua faces. O algoritmo extrai as características dessas imagens e utiliza retângulos de diferentes tamanhos e combinações de duas cores (veja a Figura 2.11). Para compreender os retângulos, considere uma imagem de dimensões 24×24 . Os retângulos são pequenos fragmentos de *pixels* que podem assumir dimensões de 1×1 até 24×24 . Esses retângulos funcionam como máscaras de *pixels* e são sobrepostos em regiões da imagem. Essa sobreposição é utilizada para calcular as características. O cálculo é realizado conforme a equação a seguir:

$$caracteristica = \sum(P_{white}) - \sum(P_{black}) \quad (2.5)$$

onde P_{white} são os *pixels* dentro da região branca do retângulo, e P_{black} os *pixels* da região preta do retângulo.

Devido aos retângulos assumirem todos os tipos de tamanhos e diferentes combinações entre as cores preto e branco, o número de características gerado é enorme. Para uma imagem com 24×24 *pixels* são geradas mais de 160.000 características. Entretanto, nem todas essas

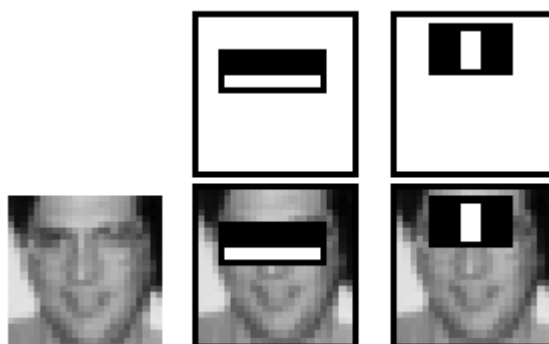
Figura 2.11: Exemplos de retângulos aplicados nas imagens para extração de características no método de detecção de faces Haar-like-cascade (VIOLA; JONES, 2001).



Fonte: (VIOLA; JONES, 2001)

características são relevantes. Por exemplo, considere a Figura 2.12 na qual são mostrados dois retângulos que geram ótimas características para a detecção de face. Com o cálculo dessas características pode-se chegar à conclusão de que a região dos olhos é mais escura do que a região da bochecha e também que os olhos são mais escuros do que a parte do começo do nariz. Entretanto essas características não podem ser aplicadas em outras regiões da imagem.

Figura 2.12: Exemplos de boas características para detecção de faces no Haar-like-cascade (VIOLA; JONES, 2001).



Fonte: (VIOLA; JONES, 2001)

Para a eliminação de características não relevantes, em (FREUND; SCHAPIRE, 1995) o treinamento é realizado com o algoritmo AdaBoost. Esse algoritmo treina as classificações que obtiveram erro até que alcance a taxa mínima de erro. Após o treinamento do AdaBoost, o número de características foi reduzido de 160.000 para 6.000. Com as características definidas, o processo de detecção de faces na imagem pode ser realizado.

A detecção de faces é aplicada em uma região de pixels de tamanho 24×24 na imagem. Para cada região é necessário que as 6.000 características indiquem a presença de uma face. Como o

processo de aplicação de todas as características em uma região consome um grande tempo, foi proposto o conceito de classificação em cascata. As características são organizadas em pacotes e um pacote por vez é aplicado na região da imagem. Caso alguma característica assinala que a região não possui uma face, a aplicação de características é interrompida. Como a maioria das regiões de uma imagem não possui uma face, essas regiões são descartadas rapidamente, aumentando o desempenho do algoritmo.

A Figura 2.13 apresenta alguns exemplos de resultados para a detecção de faces em imagens obtidos com a aplicação do Haar-like-cascade (VIOLA; JONES, 2001).

Figura 2.13: Exemplos de imagens com faces detectadas pelo algoritmo Haar-like-cascade (VIOLA; JONES, 2001).



Fonte: (VIOLA; JONES, 2001)

No trabalho de (VIOLA; JONES, 2001), o Haar-like-cascade apresentou 95% de precisão para imagens frontais.

2.2.3 ImageNet

Um dos recursos mais utilizados nos trabalhos de visão computacional é o banco de imagens ImageNet. O ImageNet está organizado de acordo com a hierarquia da WordNet, contendo em média 1000 imagens por *synset* da WordNet. Essas imagens são bastante confiáveis pois passam por testes de qualidade e são anotadas por humanos. A Figura 2.14 ilustra um exemplo de imagens da ImageNet para o *synset* “pássaro” (*bird*). Destacada em vermelho está a estrutura de *synsets* em formato de árvore e, na cor amarela, a mesma representação em árvore porém no formato visual.

Cada *synset* da imageNet possui imagens que podem ser obtidas por download. Alguns *synsets* possuem imagens com *bounding box* (veja Figura 2.15) que podem ser úteis para a extração de características do objeto de interesse, descartando o restante da imagem.

A ImageNet está disponível para download em: <http://image-net.org/>.

Figura 2.14: Exemplo de um trecho da ImageNet.

Bird
Warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings

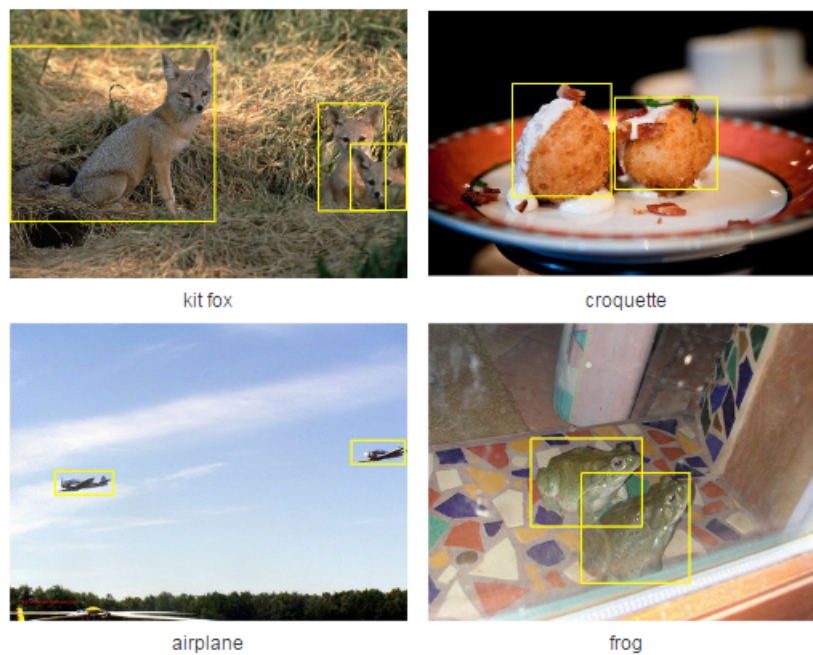
2126 pictures 92.85% Popularity Percentile Wordnet IDs

Treemap Visualization Images of the Synset Downloads

ImageNet 2011 Fall Release > > > Vertebrate, craniate > Bird

The screenshot shows the ImageNet website interface for the 'Bird' synset. On the left, a hierarchical tree of categories is displayed, with 'bird (871)' selected and highlighted by a red box. The tree includes sub-categories like 'gamecock, fighting cock', 'hen', 'nester', 'night bird', 'bird of prey', 'eagle', 'secretary bird', and 'parrot'. On the right, a grid of bird images is shown, with the grid area highlighted by a yellow box. The grid is organized into sub-categories such as 'Aquatic', 'Gallinaceous', 'Passerine', 'Archaeopteryx', 'Night', 'Carinate', 'Dickeybird', 'Caprimulgiform', 'Apodiform', 'Hen', 'Cuculiform', 'Twitterer', 'Nester', 'Cock', 'Nonpasseriform', 'Trogon', 'Ratite', 'Parrot', 'Coraciiform', and 'Piciform'.

Fonte: <http://image-net.org/>

Figura 2.15: Exemplo de imagens da ImageNet com *bounding boxes* (regiões delimitadas por retângulos amarelos).

Fonte: <http://image-net.org/>

2.3 Medidas de avaliação

As medidas de avaliação são utilizadas para mensurar o desempenho de determinado algoritmo e também para a realização de comparações entre algoritmos. As medidas mais utilizadas na área de recuperação de imagens, anotação de imagens e alinhamento texto-imagem são: precisão, cobertura (também chamada de revocação em alguns trabalhos) e medida-F (KENT et al., 1955) (veja seção 2.3.1). Na classificação de imagens também utiliza-se bastante a avaliação considerando-se apenas uma (TOP-1) ou as 5 (TOP-5) melhores classificações (veja seção 2.3.2).

2.3.1 Precisão, Cobertura e Medida-F

Essas medidas de avaliação são calculadas com base em 4 quantidades:

- Verdadeiro Positivo (VP) – Quantidade de casos que deveriam ser anotados e foram anotados com sucesso;
- Verdadeiro Negativo (VN) – Quantidade de casos que não deveriam ser anotados e efetivamente não foram;
- Falso Positivo (FP) – Quantidade de casos que foram anotados quando não deveriam ter sido;
- Falso Negativo (FN) – Quantidade de casos que deveriam ter sido anotados e não foram.

A partir dessas quantidades, os valores de precisão (*precision*), cobertura (*recall*) e medida-F (*F-measure*) são calculados como segue:

$$Precisao = \frac{VP}{VP + FP} \quad (2.6)$$

$$Cobertura = \frac{VP}{VP + FN} \quad (2.7)$$

$$Medida - F = 2 * \frac{Precisao * Cobertura}{Precisao + Cobertura} \quad (2.8)$$

A medida de precisão revela a qualidade do algoritmo ao anotar uma imagem. A medida da cobertura revela a capacidade do algoritmo de detectar áreas que necessitam ser anotadas

e também leva em consideração as anotações corretas. A medida-F, por sua vez, é a média harmônica entre a cobertura e a precisão.

Para um melhor entendimento de como essas medidas funcionam, suponha o seguinte cenário: um conjunto de dados possui 500 objetos distribuídos em diversas imagens. Esses objetos serão anotados por um algoritmo que anota as imagens automaticamente. Após a execução do algoritmo, são obtidos os resultados apresentados na Tabela 2.1.

Tabela 2.1: Resultados de um suposto algoritmo de anotação de imagens.

Resultados				
Anotações realizadas	Corretas	Incorretas	Não anotadas	Total
430	330	100	70	500

Calculando os valores das medidas para o exemplo citado anteriormente tem-se:

$$Precisao = \frac{VP}{VP + FP} = \frac{330}{330 + 100} = 0,76 = 76\% \quad (2.9)$$

$$Cobertura = \frac{VP}{VP + FN} = \frac{330}{330 + 70} = 0,825 = 82,5\% \quad (2.10)$$

$$Medida - F = 2 * \frac{0,76 * 0,825}{0,76 + 0,825} = 2 * \frac{0,627}{1,585} = 0,791 = 79,1\% \quad (2.11)$$

Como exemplos de valores encontrados na literatura para essas medidas tem-se que em (SOCHER; FEI-FEI, 2010) obteve-se 35% de precisão, 71% de cobertura e 47% de medida-F na tarefa de anotação de regiões da imagem; em (TEGEN et al., 2014) alcançou-se 96,6% de precisão, 61,8% de cobertura e 75,3% de Medida-F na tarefa de anotação de regiões da imagem; e em (SOCHER et al., 2011) chegou-se a 78,1% de precisão na tarefa de segmentação e anotação da imagem.

2.3.2 Classificação TOP-1 e TOP-5

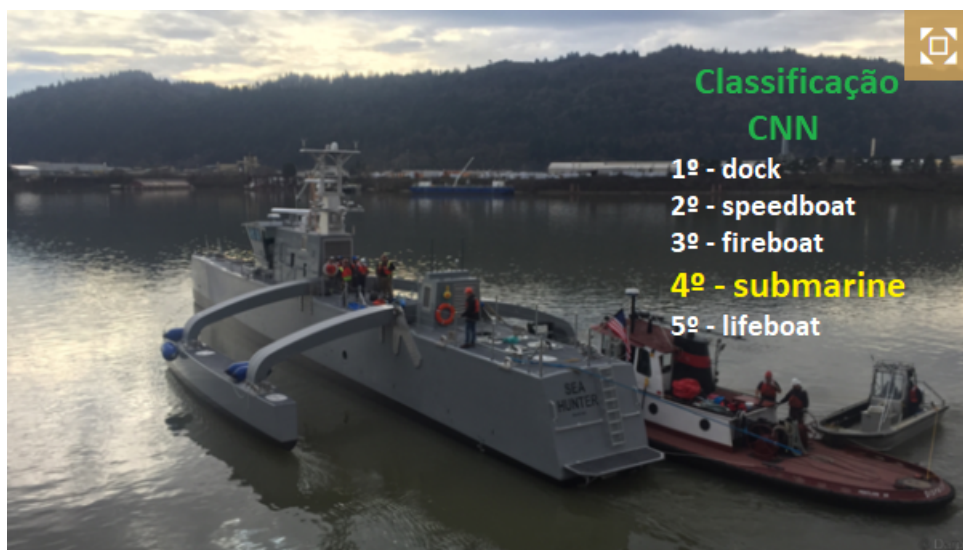
Na classificação de imagens, geralmente utilizam-se as classificações TOP-1 e TOP-5. A classificação TOP-1 considera como correta a classificação se o rótulo com maior probabilidade fornecido pela saída do classificador coincidir com o objeto presente na imagem. A classificação TOP-5, por sua vez, considera como correta a classificação quando o rótulo correto para o objeto presente na imagem está entre os cinco rótulos melhor classificados. A Figura 2.16 ilustra um exemplo de classificação TOP-1, enquanto a Figura 2.17 um exemplo de classificação TOP-5. As duas classificações foram geradas pela rede convolucional DenseNet (HUANG et al., 2016).

Figura 2.16: Exemplo de classificação TOP-1 considerada correta. Note que a imagem contém um pinguim e o classificador atribuiu a palavra “*penguin*” como o rótulo mais provável.



Fonte: <http://www.bbc.com/earth/story/20141117-why-seals-have-sex-with-penguins>

Figura 2.17: Exemplo de classificação TOP-5 considerada correta. A imagem contém um submarino. Apesar da palavra “*submarine*” não ser o rótulo melhor classificado, ela está presente entre os cinco melhores rótulos significando que a classificação está correta.



Fonte: <http://www.bbc.com/autos/story/20160409-meet-darpas-long-range-autonomous-submarine-hunter>

A classificação TOP-5 é uma escolha interessante para o trabalho de alinhamento, pois com ela é possível utilizar as palavras contidas no texto associado como auxílio para filtrar as palavras do TOP-5. Em outras palavras, com a TOP-5 aumenta-se a possibilidade de encontrar uma boa palavra para o alinhamento.

Capítulo 3

TRABALHOS RELACIONADOS

O alinhamento de texto e imagem, referenciado neste documento como alinhamento texto-imagem ou imagem-texto, consiste na associação de elementos presentes no texto com elementos presentes na imagem. Na literatura, o alinhamento de texto-imagem possui um grande foco em textos manuscritos digitalizados (ZINGER; NERBONNE; SCHOMAKER, 2009; STAMATOPOULOS; LOULLOUDIS; GATOS, 2010; FISCHER et al., 2011; LEYDIER et al., 2014; YIN; WANG; LIU, 2013), onde a imagem é o texto digitalizado e o alinhamento ocorre entre as regiões da imagem que representam palavras e suas transcrições. Esse problema também está relacionado a trabalhos sobre anotação de imagem (DESCHACHT; MOENS et al., 2007; DUYGULU et al., 2002; SOCHER; FEI-FEI, 2010; RAMISA et al., 2016), nos quais rótulos são atribuídos automaticamente a imagens completas e não a áreas específicas da imagem.

O alinhamento de regiões da imagem com palavras que ocorrem no texto que a acompanha, como é apresentado neste documento, o alinhamento texto-imagem propriamente dito, é encontrado em poucos trabalhos da literatura (PHAM; MOENS; TUYTELAARS, 2008; TEGEN et al., 2014).

Embora o alinhamento de textos manuscritos seja bastante distante da proposta apresentada neste documento, ele é comentado brevemente na seção 3.1 devido a sua importância histórica. Maior ênfase é dada aos trabalhos de anotação de imagem (descritos na seção 3.2) e de alinhamento texto-imagem (descritos na seção 3.3), mais relacionados à proposta apresentada neste documento.

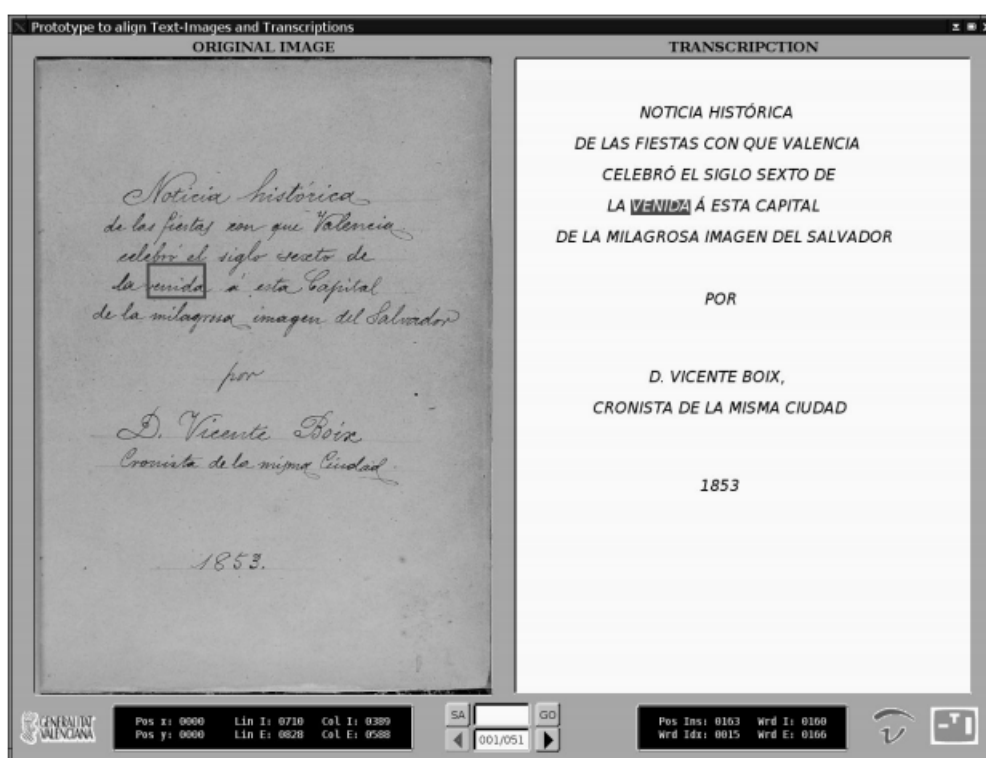
3.1 Alinhamento de imagem e texto manuscrito

Os textos manuscritos possuem uma grande importância para a nossa sociedade. Eles relatam assuntos como a história da política, costumes antigos e práticas de religião. A partir de estudos desses manuscritos pode-se entender como os costumes antigos se refletem no mundo atual. Apesar de serem uma fonte rica de conhecimento, os textos manuscritos são frágeis podendo ser facilmente danificados (LEYDIER et al., 2014). Para preservar esses documentos geralmente realiza-se a digitalização, que transforma o texto manuscrito em imagem.

Após a digitalização, muitos trechos do texto manuscrito são de difícil interpretação sendo necessária a transcrição do documento por um especialista. A existência de uma transcrição correspondente ao texto manuscrito digitalizado abre a oportunidade do alinhamento entre o texto original (imagem) e sua transcrição.

A Figura 3.1 traz um exemplo de alinhamento entre o texto manuscrito (à esquerda) e a transcrição correspondente (à direita).

Figura 3.1: Exemplo de alinhamento de texto manuscrito com sua respectiva transcrição.

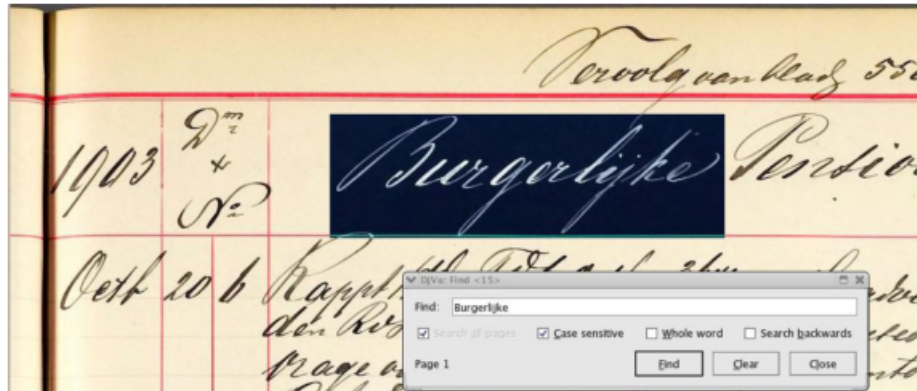


Fonte: (TOSELLI; ROMERO; VIDAL, 2007)

O alinhamento texto-imagem facilita a leitura do manuscrito e também pode ser utilizado como sistema de recuperação de informação textual. A Figura 3.2 ilustra uma ferramenta de pesquisa de texto. Utilizando essa ferramenta é possível localizar a imagem no texto manuscrito

correspondente ao texto pesquisado. O alinhamento texto-imagem também pode ser usado na criação de uma base de dados útil para o treinamento de algoritmos de aprendizado de máquina (ZINGER; NERBONNE; SCHOMAKER, 2009) capazes de auxiliar no reconhecimento de palavras e letras.

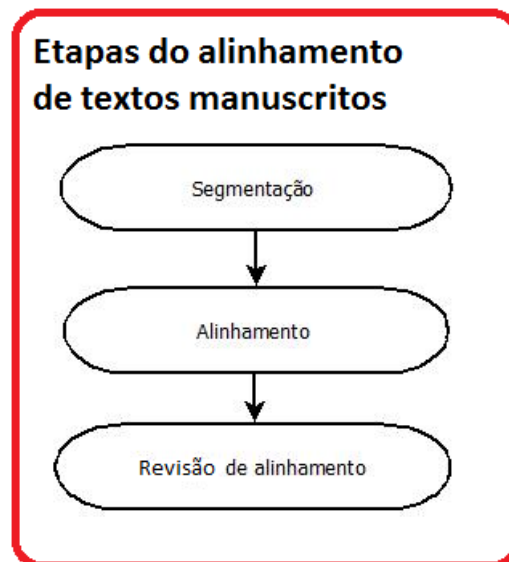
Figura 3.2: Ferramenta para localização de texto. O usuário fornece o texto desejado e o sistema localiza a imagem do texto manuscrito correspondente ao texto pesquisado.



Fonte: (ZINGER; NERBONNE; SCHOMAKER, 2009)

Os sistemas de alinhamento de textos manuscritos frequentemente realizam três etapas: segmentação, alinhamento e revisão do alinhamento (Figura 3.3).

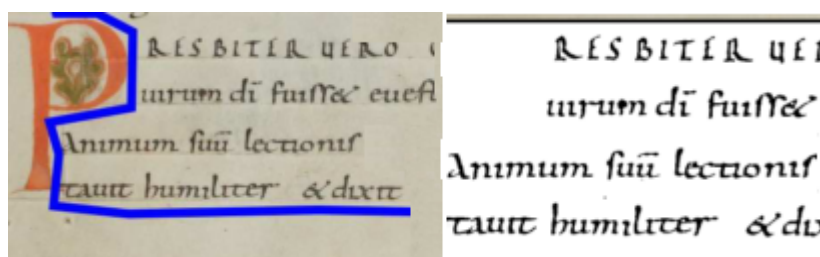
Figura 3.3: Etapas do alinhamento de textos manuscritos.



A segmentação é o processo de separação entre o texto manuscrito digitalizado e o fundo da imagem. Isso é necessário para que as técnicas aplicadas no alinhamento trabalhem apenas com as localizações da imagem que possuam caracteres de texto e excluam todas as regiões que não são de interesse (fundo). O algoritmo geralmente utilizado na segmentação é a binarização (*thresholding*), cuja função é transformar os pixels da imagem em duas cores (preto e branco).

Um valor de limiar (*threshold*) é definido baseado em testes e o valor de cada pixel é comparado com esse limiar. Pixels com valores menores que o limiar são transformados em uma das cores (preto ou branco) e pixels com valores maiores são transformados na cor inversa (SAUVOLA; PIETIKÄINEN, 2000). A Figura 3.4 ilustra esse processo com uma imagem contendo texto manuscrito (à esquerda) e sua respectiva binarização (à direita).

Figura 3.4: Exemplo de binarização. Os caracteres de texto são pintados de preto e o resto é pintado de branco, permanecendo na imagem apenas os objetos de interesse.



Fonte: Adaptado de (FISCHER et al., 2011)

Com os resultados da segmentação, inicia-se o processo de alinhamento, cujo objetivo é alinhar palavras do texto manuscrito com as palavras da transcrição. O alinhamento pode ser realizado aplicando-se diversas técnicas como os modelos ocultos de Markov (ZIMMERMANN; BUNKE, 2002; ROTHFEDER; MANMATHA; RATH, 2006), *Dynamic time warping* (KORNFIELD; MANMATHA; ALLAN, 2004; LORIGO; GOVINDARAJU, 2007) e técnicas simples envolvendo apenas cálculos de distâncias (ZINGER; NERBONNE; SCHOMAKER, 2009; STAMATOPOULOS; LOULLOUDIS; GATOS, 2010; SCHMIDT, 2014).

A etapa final consiste na correção de erros do alinhamento automático e, como tal, é realizada manualmente pelo usuário do sistema.

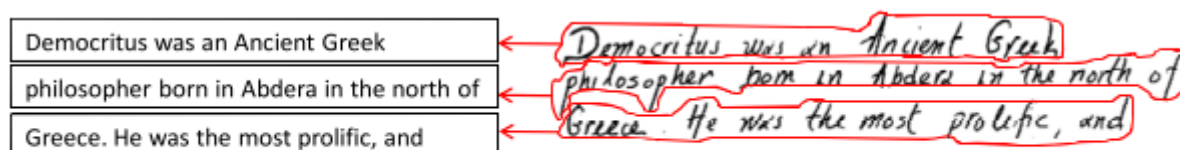
Muitos trabalhos partem do princípio de que as transcrições estão perfeitas, ou seja, que todas as palavras e caracteres do manuscrito estão presentes na transcrição, que o manuscrito e sua transcrição possuem o mesmo número de linhas com as palavras na mesma posição, e que as primeiras e últimas páginas do manuscrito são as mesmas da transcrição. Entretanto, esse cenário perfeito nem sempre ocorre e, em alguns trabalhos como (SCHMIDT, 2014), é preciso lidar com transcrições contendo abreviações e que não obedecem o número de linhas presentes no manuscrito, significando que as palavras não aparecem na mesma ordem.

A seguir são descritos dois trabalhos de alinhamento de imagem e texto manuscrito, são eles: Stamatopoulos, Louloudis e Gatos (2010) e Schmidt (2014).

3.1.1 Stamatopoulos, Louloudis e Gatos (2010)

Considerando transcrições perfeitas, Stamatopoulos, Louloudis e Gatos (2010) propõem um sistema capaz de segmentar a imagem e fazer o alinhamento das palavras. Para cada página da transcrição, são extraídos: o número de linhas e o número de palavras presentes em cada linha. Essas informações auxiliam as técnicas de alinhamento.

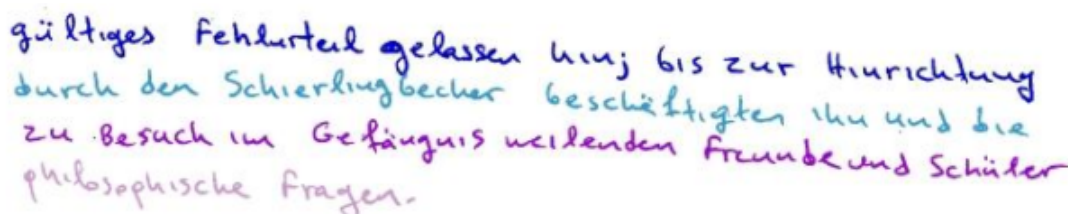
Figura 3.5: Segmentação das linhas.



Fonte: (STAMATOPOULOS; LOULLOUDIS; GATOS, 2010)

A Figura 3.5 ilustra a segmentação de linhas. Para a segmentação das linhas do manuscrito foi aplicada uma técnica de (CHANG; CHEN; LU, 2004) que extrai da imagem apenas os *pixels* que são caracteres de texto, tratando-os como componentes conectados (CCs). Com os CCs obtidos é calculada a média dos tamanhos dos caracteres de texto, com base na média dos CCs. Após o cálculo, é aplicada a transformação Hough (LOULLOUDIS et al., 2008) que transforma pontos pertencentes aos CCs em linhas, vetorizando-as. Com os vetores de linhas e o tamanho médio dos caracteres (calculado pelo tamanho médio dos CCs), a transformação Hough procura a melhor combinação dos vetores, em busca das melhores linhas que segmentem a imagem. A transformação termina quando o número de linhas de texto, for igual ao número de linhas já conhecidas, extraídas da transcrição. A Figura 3.6 mostra segmentação das linhas após a transformação Hough.

Figura 3.6: Resultado da transformação Hough. Apesar das linhas não serem retilíneas, as linhas do texto manuscrito foram segmentadas corretamente

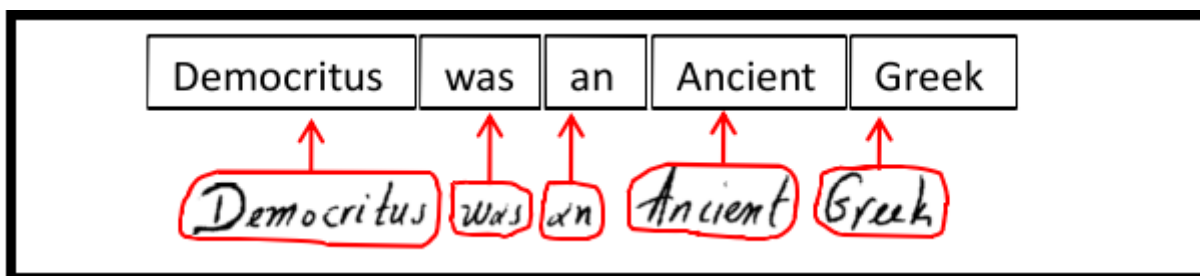


Fonte: (STAMATOPOULOS; LOULLOUDIS; GATOS, 2010)

Após a segmentação das linhas, o trabalho de alinhamento a nível de palavra é realizado em duas etapas. A primeira etapa lida com o cálculo das distâncias euclidianas entre os componentes adjacentes nas linhas (veja Figura 3.6) geradas no processo anterior. A segunda etapa é

aplicada em cada linha de texto: as medidas de distância obtidas são organizadas em um vetor de ordem decrescente sendo, então, realizada uma segmentação na linha. A primeira distância do vetor é aplicada na segmentação e a quantidade de palavras é encontrada. Se a quantidade de palavras encontradas após a segmentação for maior ou igual à quantidade de palavras da correspondente linha da transcrição, a distância é fixada para essa linha e o sistema passa para a próxima linha. Caso a distância escolhida não tenha segmentado corretamente as palavras, o processo continua obtendo o próximo valor de distância, até que as palavras estejam corretamente separadas. A Figura 3.7 traz um exemplo de alinhamento entre as palavras do texto manuscrito e palavras da transcrição.

Figura 3.7: Alinhamento entre as palavras do texto manuscrito e palavras da transcrição.



Fonte: (STAMATOPOULOS; LOULODIS; GATOS, 2010)

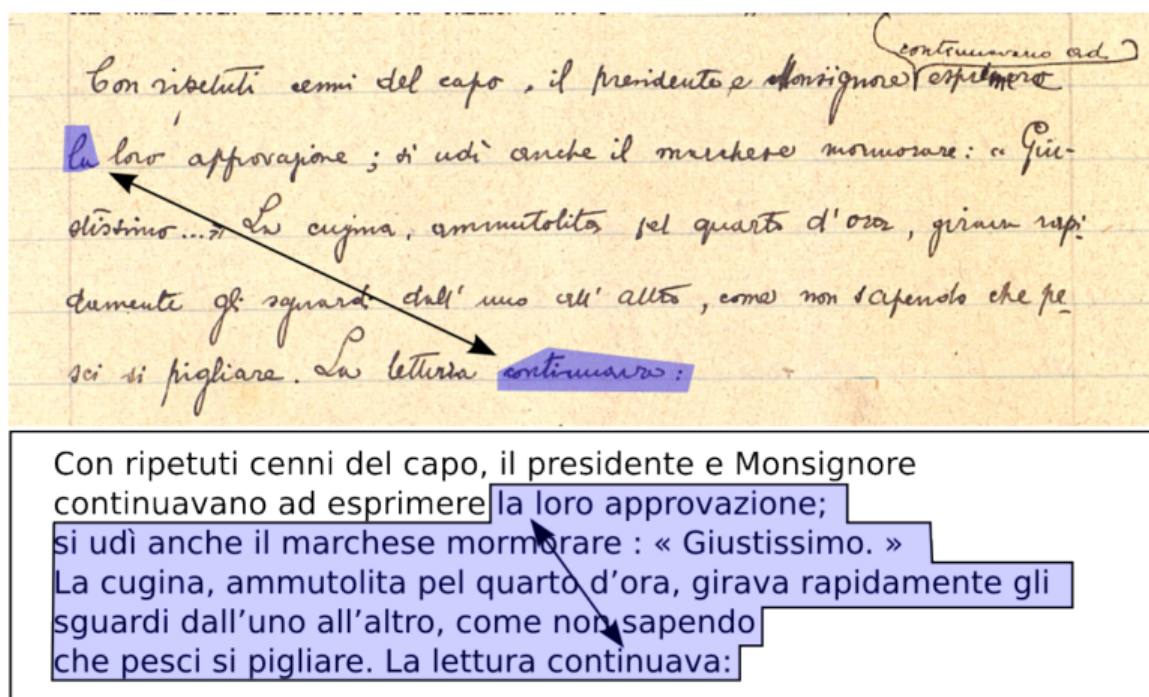
O sistema de alinhamento foi testado no conjunto de testes utilizado por Gatos, Stamatopoulos e Louloudis (2011) formado por 200 imagens de textos manuscritos, contendo 4034 linhas de textos e 29717 palavras. A precisão foi de 95% para o alinhamento de linhas e de 90% para alinhamento de palavras. O tempo de processamento também foi medido, resultando em 90% de redução de custo em relação ao trabalho manual e 12% em relação ao melhor sistema de alinhamento do estado da arte (GATOS; STAMATOPOULOS; LOULODIS, 2011).

3.1.2 Schmidt (2014)

Como já mencionado anteriormente, nem todas as transcrições são perfeitas (contêm o mesmo número de linhas do manuscrito com as palavras na mesma posição). Em (SCHMIDT, 2014), diferentemente de (STAMATOPOULOS; LOULODIS; GATOS, 2010), é possível realizar o alinhamento mesmo que o número de linhas e as posições das palavras da transcrição não sejam equivalentes àsquelas do texto manuscrito (veja Figura 3.8).

Para tanto, Schmidt (2014) inicia o seu trabalho aplicando uma segmentação na imagem do texto manuscrito convertendo-a para tons da escala de cinza e, em seguida, aplica a binarização transformando as cores da imagem em preto e branco. Na sequência, é aplicado um algoritmo utilizando a altura dos caracteres para encontrar as linhas da imagem.

Figura 3.8: Exemplo de alinhamento proposto. Note que não é necessário o uso de transcrições com o mesmo número de linhas do texto manuscrito.



Fonte: (SCHMIDT, 2014)

A Figura 3.9 traz um exemplo de texto manuscrito no lado esquerdo e no lado direito o mesmo texto com as linhas reconhecidas.

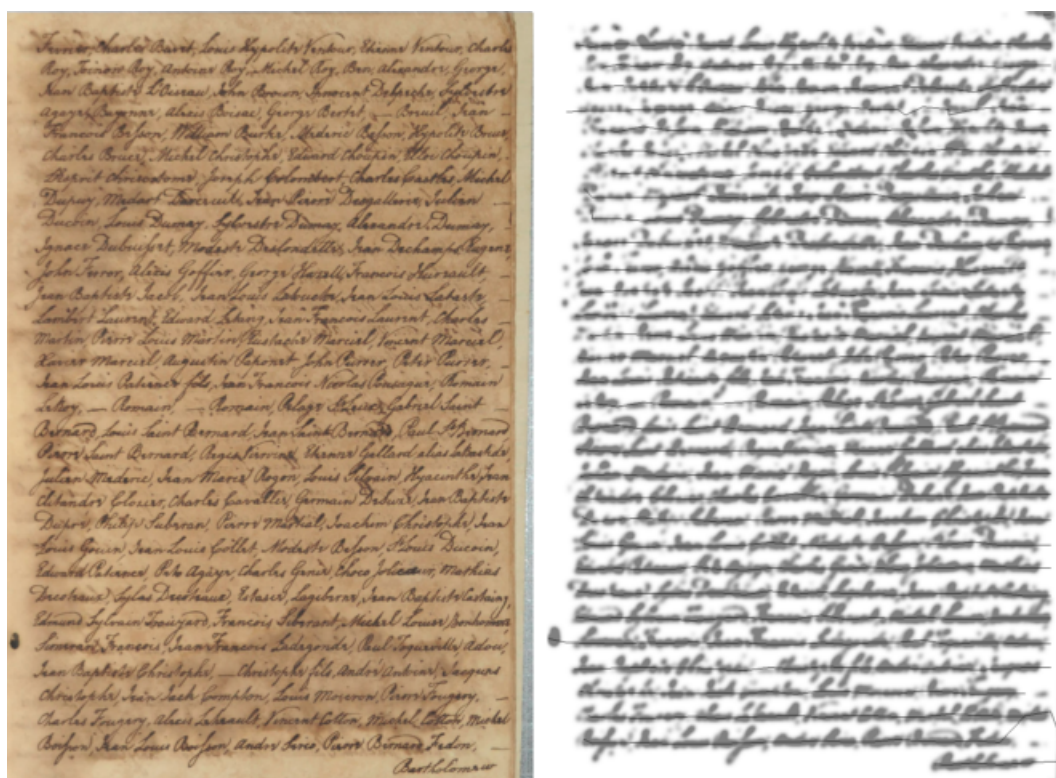
Após a identificação das linhas, segue a etapa de busca pelas palavras. Para encontrar cada palavra, foi aplicada uma técnica que identifica os objetos conectados de modo semelhante à técnica utilizada em (STAMATOPOULOS; LOULLOUDIS; GATOS, 2010). Cada objeto identificado é contornado por um polígono. Em seguida, cada espaçamento entre os polígonos é medido. Um número de espaçamentos é definido pela equação (3.1) e é utilizado para separar os polígonos. Polígonos não separados pelos espaçamentos são conectados formando um novo polígono nomeado pelo autor de *word-shapes*.

$$Num_{Espaçamentos} = (N_{palavras} - 1) - (N_{linhas} - 1) \quad (3.1)$$

onde $N_{palavras}$ o número de palavras do texto e N_{linhas} o número de linhas do texto manuscrito.

Conhecendo o número de letras do texto a partir da transcrição, é possível calcular o tamanho aproximado, em *pixels*, de cada letra dividindo o tamanho em *pixels* das *word-shapes* somadas pelo total de letras. Com o tamanho de cada letra é estimado o tamanho de cada palavra. O alinhamento é realizado com o auxílio dessas informações de medidas sendo possível alinhar uma *word-shape* com várias palavras ou várias palavras com apenas uma *word-shape*.

Figura 3.9: Texto manuscrito digitalizado e texto manuscrito com linhas identificadas.



Fonte: (SCHMIDT, 2014)

O sistema proposto por Schmidt (2014) alcançou entre 98% a 100% de precisão para livros impressos, entretanto, a porcentagem de acerto diminui para textos manuscritos que possuem muitas variações de espaçamentos entre as palavras, ficando subentendido que o alinhamento em textos manuscritos sem muitas variações de espaço é realizado com uma alta precisão pelo sistema proposto.

Diferentemente dos trabalhos citados nesta seção, o trabalho descrito neste documento visa alinhar texto e imagens em geral, ou seja, não envolverá necessariamente imagens que contenham texto, como é o caso dos textos manuscritos digitalizados. Assim, técnicas como as usadas para identificação de caracteres nas imagens não são as mais adequadas para o propósito do trabalho aqui apresentado e, por isso, não foram utilizadas na abordagem proposta.

3.2 Anotação de imagem

As ferramentas de anotação de imagem possuem a função de descrever o conteúdo de uma imagem. A descrição do conteúdo da imagem não precisa, necessariamente, ser uma sentença com sentido completo na língua natural. Por exemplo, suponha que em uma imagem exista uma menina, um cachorro e uma bola. A ferramenta de anotação de imagem tem como obje-

tivo prever que esses elementos (menina, cachorro e bola) estão presentes na imagem sem, necessariamente, gerar uma descrição do tipo “A menina está brincando com a bola e com seu cachorro”.

Nas subseções a seguir são apresentadas duas abordagens diferentes para a anotação de imagens: (seção 3.2.1) a abordagem utilizando apenas processamento de linguagem natural (PLN) e (seção 3.2.2) a abordagem utilizando PLN e visão computacional (VC).

3.2.1 Anotação de imagem usando apenas técnicas de PLN

A seguir são apresentados dois trabalhos de anotação de imagens que utilizam apenas técnicas de PLN: Choi e Kim (2012) e Tiwari e Kamde (2015).

3.2.1.1 Choi e Kim (2012)

Em (CHOI; KIM, 2012), propõe-se encontrar anotações para imagens provenientes de notícias do jornal CNN¹. Como as imagens necessitam de tempo para serem rotuladas manualmente, foi proposto um sistema automático de anotação de imagens que leva em consideração o título e o texto da notícia correspondente à imagem. De acordo com os autores, essas anotações podem ser utilizadas para detectar informações sobre terrorismo ou informações de contexto. O sistema proposto para geração de anotação pode ser visualizado na Figura 3.10.

Primeiramente, o sistema de Choi e Kim (2012) extrai os substantivos e nomes próprios presentes no título e no texto da notícia. Nesse processo são realizadas também etapas de remoção de caracteres especiais e *stopwords*², como ilustra a Figura 3.11. Esses elementos (substantivos e nomes próprios) geralmente indicam objetos de importância no texto e, por isso, são extraídos para a geração da anotação.

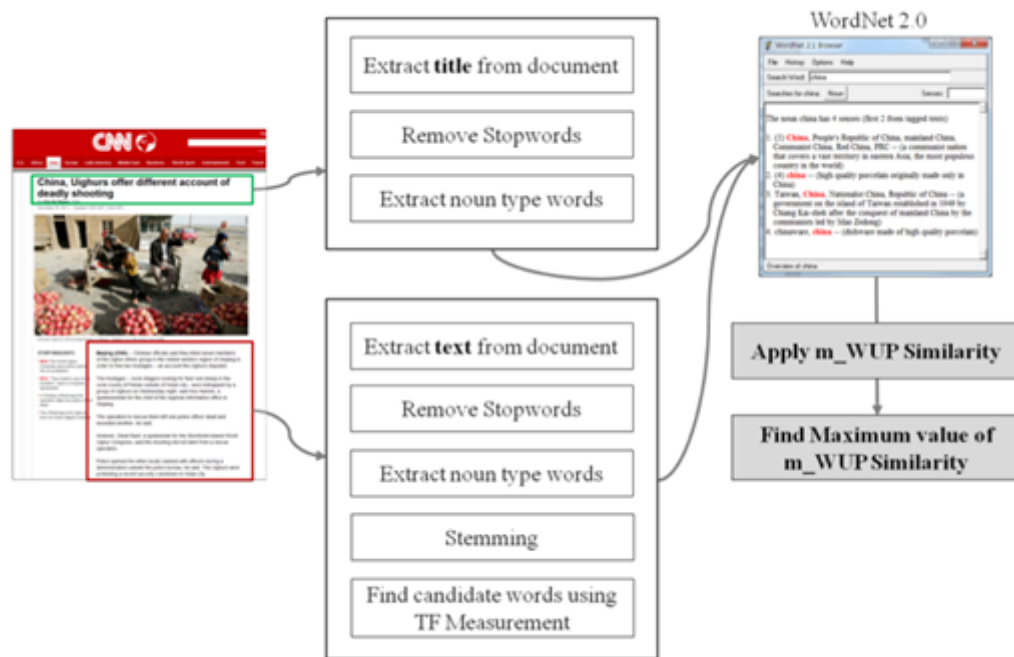
A partir dos elementos textuais extraídos, o sistema calcula a frequência de termo (TF) e a similaridade WUP (WU; PALMER, 1994) com base na WordNet (FELLBAUM, 1998).

Choi e Kim (2012) constataram que quanto maior a frequência de uma palavra, maior é sua relevância no texto, porém foi constatado que algumas palavras com baixa frequência (por exemplo, frequência = 1) possuíam grande relevância. Para tentar contornar esse problema de existência de palavras relevantes com baixa frequência, para cada termo do texto, o sistema multiplica o valor da similaridade WUP calculada com os termos do título e o valor de sua

¹Disponível em: <http://edition.cnn.com/>. Acesso em: 29 dez. 2017.

²As *stopwords* são palavras que aparecem geralmente com alta frequência em um texto, mas que não agregam significado ao seu conteúdo. Geralmente os pronomes, artigos e advérbios são considerados *stopwords*.

Figura 3.10: Sistema completo de geração de anotação proposto em (CHOI; KIM, 2012).



Fonte: (CHOI; KIM, 2012)

Figura 3.11: Processo de extração de elementos (substantivos e nomes próprios) significativos do texto no método de Choi e Kim (2012)

Step	Result
Surrounding Text	Beijing (CNN) -- Chinese officials said they killed seven members of the Uighur ethnic group in the restive western region of Xinjiang in order to free two hostages - - an account the Uighurs disputed.
Remove Special Characters	Beijing CNN Chinese officials killed seven Uighur ethnic restive western region Xinjiang free hostages account Uighurs disputed ...
Remove Stopwords	Beijing CNN Chinese officials killed seven Uighur ethnic restive western region Xinjiang free hostages account Uighurs disputed ...
Extract noun type words	Beijing CNN Chinese officials Uighur western region Xinjiang hostages Uighurs ...

Fonte: (CHOI; KIM, 2012)

frequência no texto. Dessa forma, um termo do texto que possui baixa frequência tem a chance de ser integrado à anotação se sua similaridade com os termos do título da notícia for alta. A Figura 3.12 traz exemplos de imagens de notícias anotadas com o sistema proposto por Choi e Kim (2012).

Os autores citam que o tempo de processamento do sistema proposto é muito menor que as

Figura 3.12: Anotações de imagem geradas pelo sistema proposto por Choi e Kim (2012).

Image		
Annotation	Uighurs, security, Xinjiang, China, terrorism, Asian, crackdown, police, Beijing, Pakistan	Philippines, storm, donation, Asia, Children, China, rain, flood, Australia, Europe

Fonte: (CHOI; KIM, 2012)

abordagens que utilizam técnicas de visão computacional, entretanto não apresentam a avaliação da abordagem proposta.

3.2.1.2 Tiwari e Kamde (2015)

Em (TIWARI; KAMDE, 2015), o objetivo é anotar imagens da web para, posteriormente, utilizá-las em um sistema de recuperação de imagens. Para tanto, os autores propõem o uso de informações contextuais presentes no texto próximo à imagem e também informações contidas em etiquetas (*tags*) da página html. Inicialmente, são considerados os seguintes elementos contextuais:

- Nome da imagem presente no atributo SRC da etiqueta IMG;
- Texto alternativo para a imagem presente no atributo ALT da etiqueta IMG;
- Título da página presente como conteúdo da etiqueta TITLE;
- O *link* de texto presente como conteúdo da etiqueta A;
- Textos que aparecem antes e depois da imagem.

Após a extração desses elementos, são removidas as *stopwords* e o texto restante é lematizado usando o algoritmo de (PORTER, 1980). A lematização permite que um maior número de dados seja recuperado em uma pesquisa (mais detalhes sobre a lematização na seção 2.1.1). Em seguida, calcula-se tf-idf (frequência de termo X frequência inversa do documento) (SALTON; MCGILL, 1983) para cada palavra do texto. O tf-idf é o peso atribuído a cada palavra que indica a relevância da palavra no texto. As cinco palavras com maiores pesos são selecionadas para a anotação da imagem.

As imagens anotadas são, então, armazenadas em um banco de dados para serem recuperadas a partir de um buscador.

Os autores concluem que os textos próximos à imagem e as informações contidas nas etiquetas HTML são corretas e relevantes para a anotação da imagem, porém eles não apresentam a avaliação das anotações das imagens.

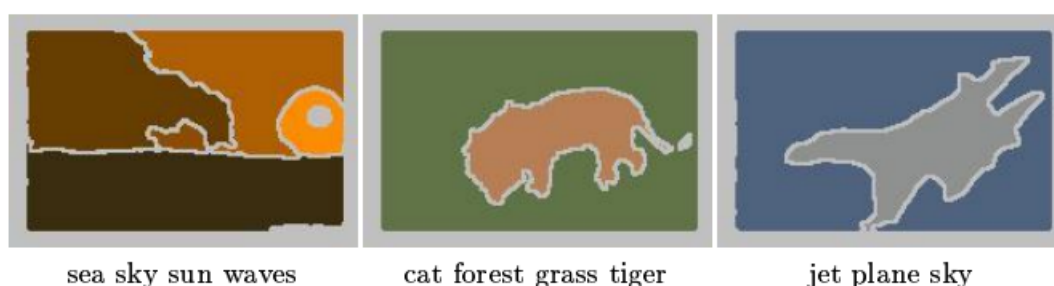
3.2.2 Anotação de imagem usando técnicas de PLN e de VC

Embora as técnicas de PLN (como as apresentadas na seção anterior) sejam essenciais para um bom resultado na anotação de imagens, as técnicas de VC tornam-se necessárias para a manipulação de regiões de uma imagem. Nesta seção são apresentados trabalhos que combinam PLN para manipulação do texto e VC para tratamento da imagem. Em alguns casos há também o uso de técnicas de aprendizado de máquina que fazem uso dos resultados obtidos pela combinação de PLN e VC.

3.2.2.1 Duygulu et al. (2002)

O sistema proposto em (DUYGULU et al., 2002) segmenta a imagem em regiões e encontra os rótulos mais adequados para elas, os quais são utilizados para a anotação da imagem como um todo. A Figura 3.13 traz exemplos de algumas imagens do Corel, segmentadas em regiões acompanhadas de suas palavras-chave.

Figura 3.13: Exemplo de imagens do banco de dados Corel segmentadas em regiões e suas respectivas anotações (palavras-chave).



Fonte: (DUYGULU et al., 2002)

As regiões da imagem foram obtidas aplicando a segmentação de cortes normalizados (SHI; MALIK, 2000) seguida da aplicação do algoritmo de agrupamento *k-means* (MACQUEEN et al., 1967). O algoritmo de *k-means* foi utilizado para discretizar as regiões da imagem, uma necessidade já que o alinhamento é feito entre elementos textuais discretos (parágrafos, sentenças, palavras), porém esses elementos discretos não estão naturalmente presentes em uma imagem.

Esse agrupamento gera um rótulo para cada região, chamado pelo autor de “*blob*” (*Binary Large Object*).

Para cada região da imagem, foram extraídas 33 características (cor, localização, desvio padrão, etc.) que servem como dados de entrada para o algoritmo *k-means*.

Com a aplicação do *k-means*, as regiões foram agrupadas em 500 *blobs* diferentes, significando que cada região de imagem presente no conjunto de treinamento é associada a um desses *blobs*. Para o treinamento do sistema, foram utilizadas 4500 imagens provenientes do banco de dados Corel³, sendo que cada imagem está associada a 4-5 palavras-chave previamente anotadas por humanos. A entrada no formato texto são as palavras-chaves que acompanham as imagens, totalizando 371 palavras. Com os 500 *blobs* obtidos no *k-means* e as 371 palavras-chaves, é criada uma tabela com as probabilidades das correspondências entre os *blobs* e as palavras.

Para anotar uma imagem de teste, o sistema treinado faz buscas na tabela de probabilidades. Primeiro, para cada região da nova imagem, o sistema busca o *blob* mais similar (de maior probabilidade). Em seguida, ele encontra a palavra com maior probabilidade de correspondência ao *blob* encontrado no passo anterior.

Como estratégia de teste, os autores optaram por priorizar a precisão em detrimento da cobertura. Para tanto, palavras com baixa probabilidade e suas regiões correspondentes são desconsideradas. Com essa decisão, algumas áreas não são cobertas pelo sistema treinado, porém, as áreas cobertas possuem uma probabilidade maior de serem anotadas corretamente. O valor mínimo de probabilidade de uma palavra é definido empiricamente.

Os autores relatam um teste de performance no qual 500 imagens retiradas do conjunto de treinamento foram anotadas com apenas 80 das 371 palavras-chave disponíveis. Em seguida, partiram para a avaliação manual da anotação retornada pelo sistema usando essas 80 palavras-chave e 100 imagens de teste (não presentes no conjunto de treinamento). Os autores relataram que algumas palavras tiveram uma taxa de acerto de 70%, significando que em 70% das vezes que elas foram atribuídas a uma região essa atribuição estava correta. As Figuras 3.14 e 3.15 mostram algumas imagens anotadas com bons resultados e com resultados não satisfatórios, respectivamente.

3.2.2.2 Socher e Fei-Fei (2010)

Diferentemente de Duygulu et al. (2002), que realizaram a anotação de imagens com um conjunto de palavras-chave, o trabalho de Socher e Fei-Fei (2010) visa identificar e anotar as

³Disponível em: <http://corel.digitalriver.com/>

Figura 3.14: Resultados bons gerados pelo sistema proposto por Duygulu et al. (2002)

Fonte: (DUYGULU et al., 2002)

Figura 3.15: Resultados não satisfatórios gerados pelo sistema proposto por Duygulu et al. (2002)

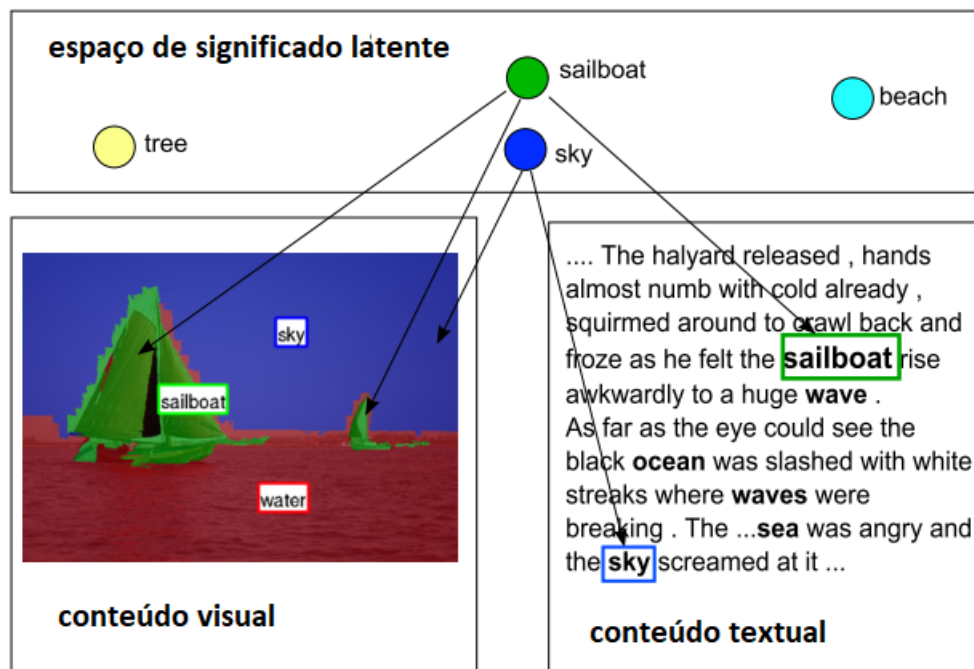
Fonte: (DUYGULU et al., 2002)

regiões de uma imagem com o auxílio de textos de notícias de jornal. Em seus experimentos, foram utilizadas 800 imagens referentes a oito categorias de esportes, com cinco imagens rotuladas para cada categoria (LI; SOCHER; FEI-FEI, 2009) e todos os textos de notícias do jornal New Work Times (SANDHAUS, 2008) que falam sobre essas oito categorias. Apesar das notícias de jornal tratarem de assuntos relacionados às imagens, a ligação entre texto e imagem não é tão direta quanto as anotações de imagens utilizadas em (DUYGULU et al., 2002).

O sistema proposto em (SOCHER; FEI-FEI, 2010) trabalha com o conceito de contexto para relacionar objetos. Para ilustrar essa ideia, por exemplo, considere que o contexto no qual “barco a vela” ocorre em textos geralmente traz palavras como “vento”, “água”, “céu” e “mar”. Desse modo, em uma imagem contendo um barco a vela pode-se supor que, além do barco, outras regiões da imagem estarão relacionadas a céu e água, por exemplo. A Figura 3.16 exemplifica esse caso. Nesse exemplo, pode-se observar a ocorrência da palavra “sailboat” (veleiro) com a palavra “sky” (céu) no texto. Na imagem, há a ocorrência de segmentos contendo o veleiro, a água e o céu. Todas essas informações de contexto são coletadas e utilizadas para o mapeamento visual-textual.

O primeiro passo do sistema proposto em (SOCHER; FEI-FEI, 2010) é extrair o conteúdo textual (notícias de jornal) e visual (imagem) e seus respectivos elementos contextuais. O conteúdo

Figura 3.16: Exemplo de atuação do contexto considerada em (SOCHER; FEI-FEI, 2010).



Fonte: (SOCHER; FEI-FEI, 2010)

textual é composto de substantivos e adjetivos extraídos dos artigos de jornal. Essas classes de palavras foram obtidas seguindo os passos:

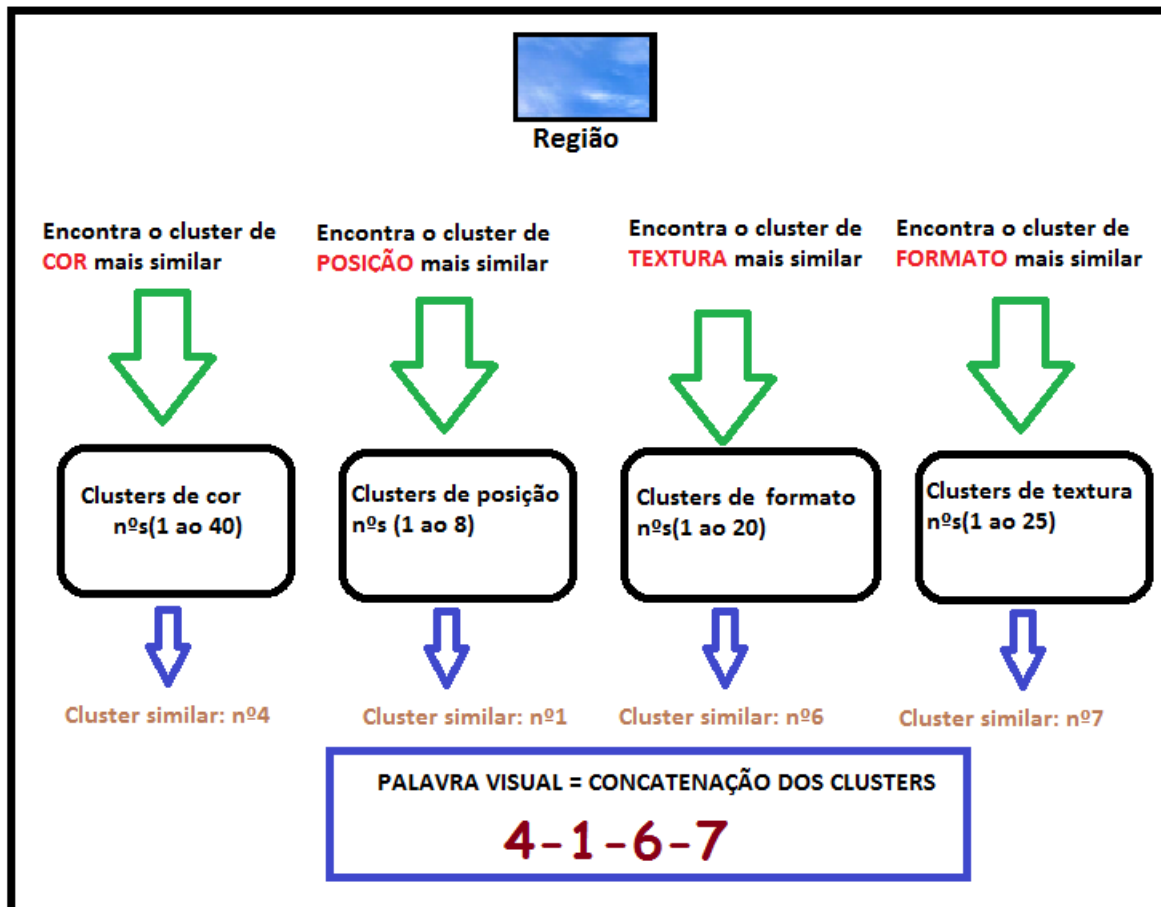
1. Extração de substantivos e adjetivos usando a ferramenta TreeTagger (SCHMID, 2013);
2. Mapeamento de nomes próprios para a categoria “humano” usando o Lexique (NEW, 2006);
3. Unificação de sinônimos (mapeando as diversas ocorrências para uma única palavra) usando a WordNet;
4. Eliminação das palavras consideradas não físicas (isto é, palavras que não podem ser mapeadas como um objeto em uma imagem) usando a WordNet;
5. Obtenção dos elementos contextuais do texto a partir da contagem de co-ocorrência das palavras.

O conteúdo visual, por sua vez, é obtido de regiões da imagem seguindo os passos:

1. Segmentação das imagens seguindo a abordagem de (FELZENSZWALB; HUTTENLOCHER, 2004), que faz a separação da imagem em regiões semelhantes;

2. Extração de características para cada segmento: cor, posição, formato e textura (MALISIEWICZ; EFROS, 2008);
3. Agrupamento de cada característica (cor, posição, formato e textura) usando o *k-means*. Diferindo do sistema de Duygulu et al. (2002), que concatenava as características de cada região e as agrupava, o sistema de Socher e Fei-Fei (2010) agrupa as características separadamente. Como resultado, este sistema gera grupos diferentes para cor, posição, formato e textura sendo as quantidades de cada grupo, estabelecidas empiricamente, como: 40 para cor, 8 para posição, 20 para formato e 25 para textura. A partir da concatenação dos grupos de características, chega-se a um novo tipo de objeto, nomeado pelos autores de “palavra visual”. A Figura 3.17 ilustra o processo de formação da palavra visual;
4. Obtenção dos elementos contextuais visuais a partir da aplicação de uma abordagem de similaridade que realiza a contagem de co-ocorrências dos segmentos de imagem vizinhos à palavra visual.

Figura 3.17: A concatenação dos grupos mais similares gera as “palavras visuais”. Nesse exemplo, os mais similares formam a palavra visual 4-1-6-7.

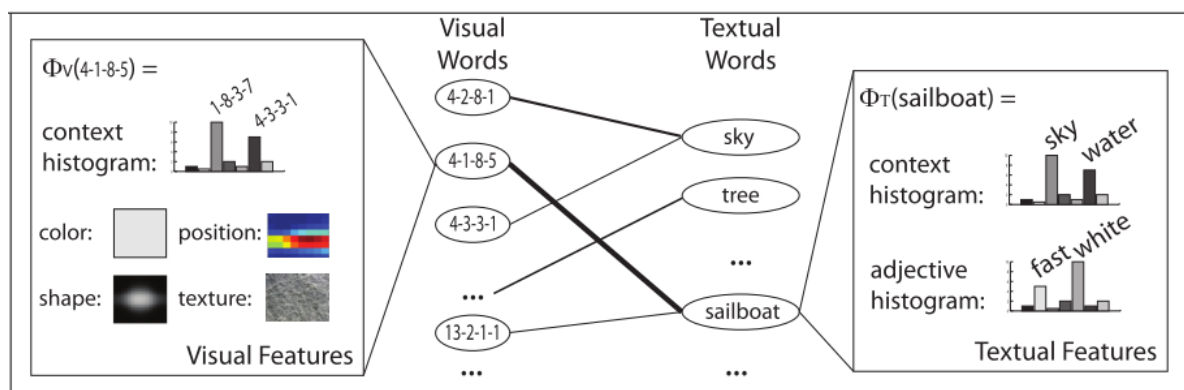


Após a extração dos conteúdos visual e textual, o próximo passo é mapear esses conteúdos em um espaço de significado latente. Para tanto, o sistema cria um grafo bipartido que liga o domínio visual ao textual, gerando um mapeamento entre as entidades. O mapeamento é gerado usando a técnica “análise de correlação canônica” (do inglês, *kernelized canonical correlation analysis*) que, a partir de cálculos de probabilidade condicional, encontra os relacionamentos entre palavras visuais e textuais.

Um detalhe importante do mapeamento é o fato de ser um mapeamento n-para-1 significando que podem existir várias palavras visuais que estão ligadas a uma palavra de texto, mas o contrário não é permitido. Para exemplificar, considere a palavra visual '4-1-6-7' atribuída no exemplo da Figura 3.17 e as palavras visuais: '4-2-6-6' e '4-2-7-7'. Note que os números dos grupos dessas três palavras são próximos, o que pode caracterizar que todas essas palavras visuais estão relacionadas a “céu”.

A Figura 3.18 exibe o grafo bipartido com um exemplo de mapeamento entre os domínios visual e textual da palavra “veleiro” (*sailboat*). Nesse exemplo, a palavra “veleiro” (*sailboat*) foi relacionada às palavras visuais '4-1-8-5' e '13-2-1-1'. Inicialmente, é possível atribuir palavras visuais para a palavra “veleiro” porque as imagens de treinamento possuem rótulos previamente anotados para cada região da imagem. Nessa figura é possível, também, visualizar que o histograma de contexto visual presente nas características visuais traz algumas palavras visuais como a palavra '4-3-3-1'. Essa palavra visual correspondente à palavra “céu” (*sky*) no grafo, significando que a palavra de texto “céu” pertence ao contexto da palavra “veleiro”. Com relação às características textuais, pode-se notar que a palavra “veleiro” também tem a palavra “céu” em seu histograma de contexto. A partir dessas informações de contexto, pode-se concluir que uma imagem contendo um “veleiro” provavelmente terá uma região da imagem contendo um “céu”.

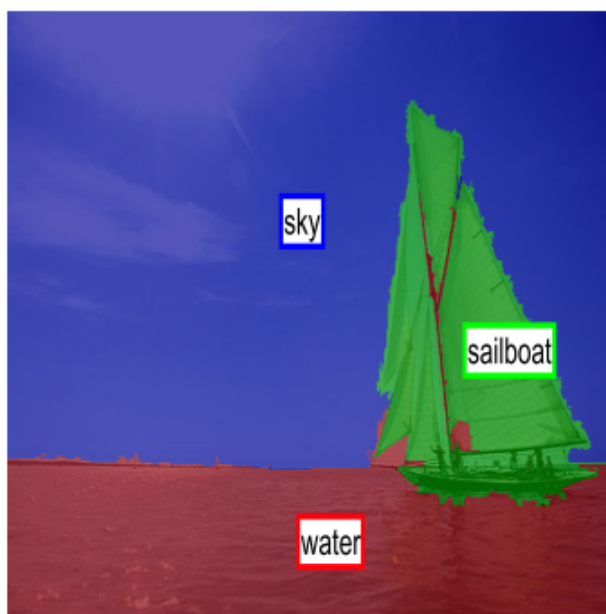
Figura 3.18: Características e grafo utilizados no mapeamento entre as palavras visuais e textuais em (SOCHER; FEI-FEI, 2010).



Fonte: (SOCHER; FEI-FEI, 2010)

Se a palavra textual e a palavra visual estiverem próximas no mapeamento, provavelmente pertencem ao mesmo conceito semântico. Com o mapeamento aprendido, é possível anotar uma imagem rotulando todas as regiões segmentadas. A Figura 3.19 exibe uma imagem anotada a partir do sistema proposto em (SOCHER; FEI-FEI, 2010).

Figura 3.19: Anotação de uma imagem a partir do sistema proposto em (SOCHER; FEI-FEI, 2010).



Fonte: (SOCHER; FEI-FEI, 2010)

O sistema proposto em (SOCHER; FEI-FEI, 2010) foi treinado e comparado com outros sistemas em relação à segmentação (LI; SOCHER; FEI-FEI, 2009; CAO; FEI-FEI, 2007) e à anotação (LI; WANG, 2003; BLEI; JORDAN, 2003; LI; SOCHER; FEI-FEI, 2009). As Tabelas 3.1 e 3.2 apresentam os resultados da comparação entre esses sistemas de acordo com as medidas de precisão (P), cobertura (R) e medida-F (F).

Tabela 3.1: Comparação dos resultados de anotação de imagem do sistema de (SOCHER; FEI-FEI, 2010) e os sistemas Alipr (LI; WANG, 2003), Corr LDA (BLEI; JORDAN, 2003) e Total Scene (LI; SOCHER; FEI-FEI, 2009)

Anotação	Alipr			Corr LDA			Total Scene			(SOCHER; FEI-FEI, 2010)		
	P	R	F	P	R	F	P	R	F	P	R	F
Média	17%	25%	20%	17%	37%	23%	29%	76%	42%	35%	71%	47%

Fonte: Adaptado de (SOCHER; FEI-FEI, 2010)

Os resultados da anotação apresentaram uma melhora de 5 pontos percentuais, em termos de medida-F, em relação ao estado da arte (LI; SOCHER; FEI-FEI, 2009). Os resultados da segmentação foram piores do que os dos sistemas usados na comparação. Um dos problemas relatados pelos autores foi o fato de algumas palavras serem confundidas, como os pares [céu

Tabela 3.2: Comparação dos resultados de segmentação de imagem do sistema de (SOCHER; FEI-FEI, 2010) e os sistemas (CAO; FEI-FEI, 2007) e Total Scene (LI; SOCHER; FEI-FEI, 2009)

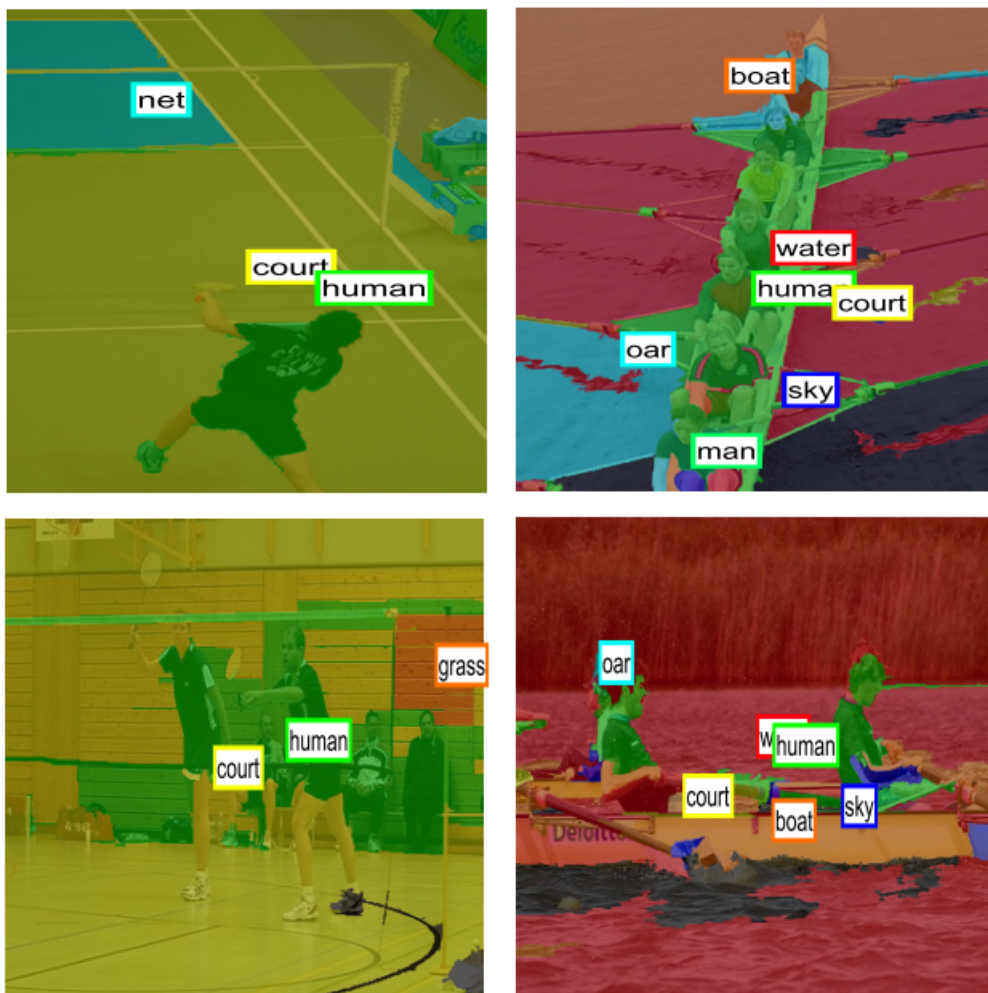
Segmentação	(CAO; FEI-FEI, 2007)			<i>Total Scene</i>			(SOCHER; FEI-FEI, 2010)		
	P	R	F	P	R	F	P	R	F
Média	35%	32%	33%	45%	43%	44%	30%	24%	27%

Fonte: Adaptado de (SOCHER; FEI-FEI, 2010)

e neve] e [nuvens e céu], o que ocorre devido à semelhança entre as características visuais e contextuais desses elementos e ao pequeno tamanho do conjunto de treinamento.

As Figuras 3.20 e 3.21 trazem exemplos de boas anotações e de anotações não muito satisfatórias, respectivamente, geradas pelo sistema de (SOCHER; FEI-FEI, 2010).

Figura 3.20: Resultados bons gerados pelo sistema de (SOCHER; FEI-FEI, 2010).



Fonte: (SOCHER; FEI-FEI, 2010)

Figura 3.21: Resultados não satisfatórios gerados pelo sistema de (SOCHER; FEI-FEI, 2010).

Fonte: (SOCHER; FEI-FEI, 2010)

3.2.2.3 Socher et al. (2011)

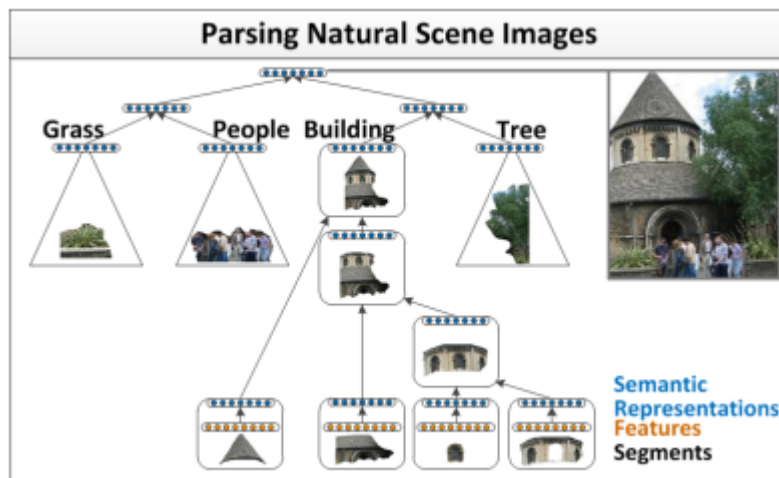
Em contraste com os sistemas probabilísticos apresentados anteriormente, Socher et al. (2011) propõem um sistema utilizando redes neurais recursivas (RNR). Uma RNR é uma variação da rede neural tradicional. Na rede neural tradicional os dados são inseridos um por vez e geram um peso que é utilizado para treinamento da rede (ROSENBLATT, 1958), enquanto a RNR combina duas entradas de dados, para formar uma nova entrada que compartilha o peso combinado das duas entradas (SOCHER et al., 2011). Essa característica se repete ao longo da rede combinando as entradas de forma recursiva.

A ideia de utilização da recursividade está motivada pelas regras sintáticas da linguagem natural, que são conhecidas por possuírem uma estrutura recursiva contendo sintagmas nominais dentro de sintagmas nominais. Essa recursividade também está presente em imagens, por exemplo, um segmento de imagem contendo um carro, ao ser dividido recursivamente, permite encontrar partes como pneu e janela. Considerar essa estrutura recursiva ajuda na segmentação, na anotação, e também na classificação de cenas da imagem.

O sistema proposto por Socher et al. (2011) é ilustrado na Figura 3.22. Para a construção do sistema, o primeiro passo é segmentar a imagem em pequenas regiões obtendo, assim, prováveis pedaços de objetos. Cada região da imagem possui características visuais e sua representação semântica (*semantic representation*). A RNR calcula a probabilidade de combinação entre essas regiões e combina as de maior probabilidade gerando uma nova região de tamanho maior (“super região”) e com sua própria representação semântica. Por fim, as regiões combinadas recebem um rótulo indicando a classe à qual elas pertencem.

Após todas as regiões pertencentes à mesma classe serem combinadas formando novas regiões, as regiões vizinhas são combinadas, formando a imagem inteira. Essas decisões de

Figura 3.22: Estrutura de árvore gerada pela RNN em (SOCHER et al., 2011).



Fonte: (SOCHER et al., 2011)

combinação de regiões são mapeadas em uma árvore binária.

A representação semântica de cada região é obtida usando uma rede neural simples. As regiões da imagem são obtidas a partir do algoritmo de Comaniciu e Meer (2002) e as características de cada região (cor, textura, aparência, formato, etc.) são extraídas seguindo o sistema proposto por Gould, Fulton e Koller (2009). As 119 características extraídas nos experimentos relatados pelos autores são passadas para a rede neural responsável por mapeá-las em uma representação semântica, chamada pelos autores de “vetor de ativação”.

A RNN busca a melhor combinação entre os pares de regiões e gera uma árvore binária contendo as combinações. A árvore só estará correta se todas as regiões adjacentes pertencentes à mesma classe forem transformadas em super regiões antes de serem combinadas com super regiões de diferentes classes. Cada imagem do conjunto de treinamento já possui rótulos vinculados (classes) a cada região da imagem, facilitando o aprendizado da RNN.

Além de gerar a árvore binária de combinações de regiões, a RNN armazena a representação semântica em cada nó da árvore. Como as imagens de treinamento já possuem um rótulo para cada região, é possível adicionar uma camada softmax simples (BISHOP, 2006)⁴, cuja entrada é a representação semântica e a saída é a classe a qual pertence o nó. Desse modo, a softmax treinada é capaz de prever a qual classe pertence um determinado nó.

O sistema proposto foi avaliado nos quesitos de: (1) segmentação e anotação da imagem e (2) classificação de cenas utilizando o banco de dados de Stanford⁵. No primeiro quesito,

⁴A softmax é uma técnica de normalização dos números gerados pela saída da rede. A softmax gera saídas baseadas em porcentagens, o que facilita o entendimento da saída da rede.

⁵Disponível em: <http://dags.stanford.edu/projects/scenedataset.html>

o sistema foi comparado com os trabalhos considerados estado da arte – (TIGHE; LAZEBNIK, 2010) e (GOULD; FULTON; KOLLER, 2009) – obtendo uma precisão maior, conforme mostra a Tabela 3.3. A RNN também foi comparada a um sistema simples que utiliza uma camada de rede neural simples junto a softmax (“Log. Repr. on Superpixel Features” na Tabela 3.3). Algumas imagens resultantes do teste podem ser vistas na Figura 3.23.

Tabela 3.3: Comparação dos valores de precisão obtidos pelo sistema de RNN proposto em (SOCHER et al., 2011) e os sistemas do estado da arte (TIGHE; LAZEBNIK, 2010) (TL) e (GOULD; FULTON; KOLLER, 2009).

Method and Semantic Pixel Accuracy in	%
Pixel CRF, Gould et al. (2009)	74.3
Log. Repr. on Superpixel Features	75.9
Region-based energy, Gould et al. (2009)	76.4
Local Labeling, TL (2010)	76.9
Superpixel MRF, TL (2010)	77.5
Simultaneous MRF, TL (2010)	77.5
RNN (SOCHER et al., 2011)	78.1

Fonte: (SOCHER et al., 2011)

Para o teste de classificação de cena, todas as imagens foram rotuladas entre três tipos de cena (cidade, campo e beira-mar) e treinadas usando um SVM linear (CRISTIANINI; SHAW-TAYLOR, 2000) utilizando a média dos vetores de ativação de cada nó da árvore como entrada para o SVM. O resultado foi uma precisão de 88,1%, superando o estado da arte que consistia nos descritores Gist (OLIVA; TORRALBA, 2001) que obtiveram 84%.

3.2.2.4 Tirilly et al. (2010)

A maioria dos trabalhos de anotação de imagem utiliza um grande acervo de imagens já rotuladas como dados de treinamento. Entretanto, gerar um conjunto grande de imagens anotadas consome um grande tempo. Como alternativa, em (TIRILLY et al., 2010) utilizou-se um sistema não-supervisionado que se baseia em textos de notícias para anotação das imagens e as legendas das imagens de um conjunto de referência são utilizadas para a validação do sistema. Como não há necessidade de obter dados de treinamento confiáveis, é possível utilizar cópulas de diferentes tamanhos sem afetar o desempenho do sistema. Os autores também apresentam uma abordagem supervisionada utilizando dados de treinamento para a realização de uma comparação entre os sistemas.

O sistema não-supervisionado de anotação de imagens proposto em (TIRILLY et al., 2010) consiste em identificar entidades nomeadas no texto e associá-las a faces e logos identificados

Figura 3.23: Resultados da segmentação e anotação do sistema de Socher et al. (2011). Cada *pixel* foi pintado com a cor correspondente à sua classe.



Fonte: (SOCHER et al., 2011)

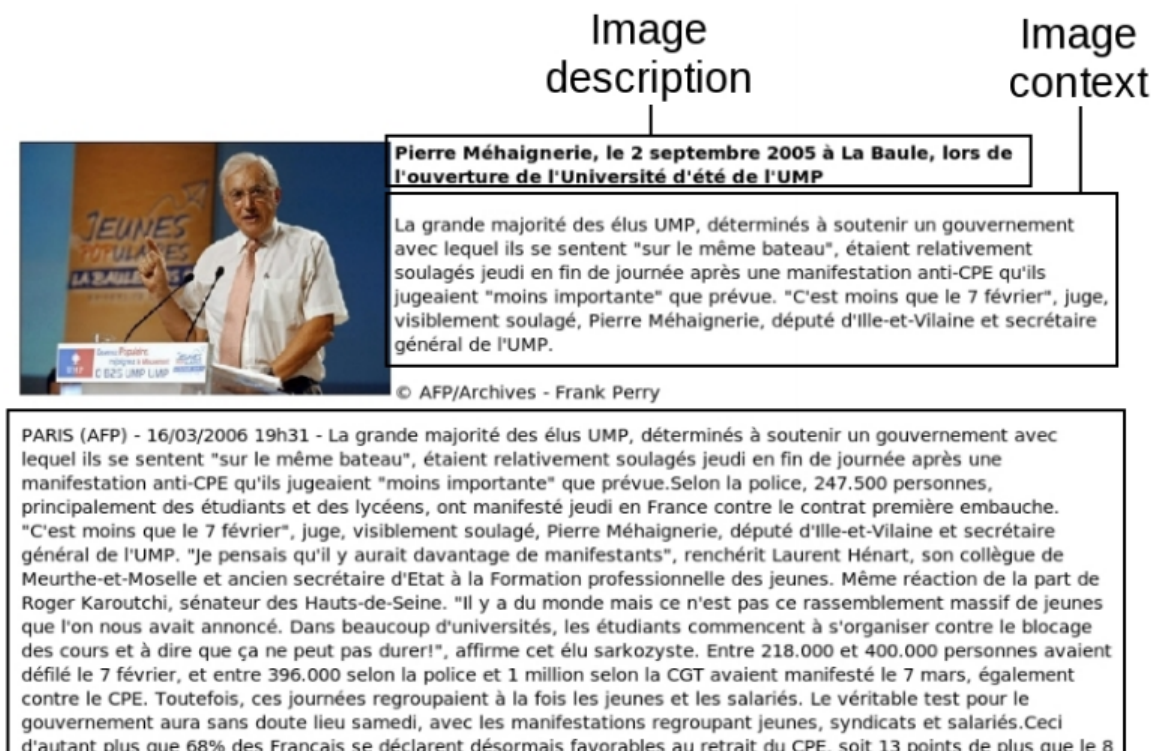
na imagem. O corpus utilizado contém 27.041 artigos do jornal AFP⁶ coletados entre março e novembro de 2006. Os artigos, escritos em francês, contêm 42.568 imagens. Um exemplo de artigo e sua respectiva imagem é apresentado na Figura 3.24. Cada artigo possui uma primeira legenda em negrito que descreve a imagem e é utilizada como referência. A outra parte da legenda refere-se ao contexto da notícia e é seguida pelo texto da notícia propriamente dito.

O primeiro passo do sistema proposto em (TIRILLY et al., 2010) é detectar as faces ou logos presentes na imagem. Para a detecção de faces foi utilizado um detector fornecido pela biblioteca openCV (LIENHART; KURANOV; PISAREVSKY, 2003). Para a detecção de logos utilizou-se um algoritmo desenvolvido pelos autores (TIRILLY et al., 2010), baseado em um *framework* proposto por (SIVIC; ZISSERMAN, 2003). Esse algoritmo foi testado em um conjunto de 413 imagens e obteve 95% de precisão e 60% de cobertura (TIRILLY; CLAVEAU; GROS, 2010).

Para a detecção das entidades nomeadas foi utilizado o NEMESIS (FOUROUR, 2002). De

⁶Disponível em: <https://www.afp.com/en/afp-news>

Figura 3.24: Exemplo de um artigo (composto de imagem e texto) presente no corpus utilizado em (TIRILLY et al., 2010).



Fonte: (TIRILLY et al., 2010)

acordo com os autores, esse algoritmo obteve 95% de precisão e 90% de cobertura na detecção de antropônimos e topônimos no corpus de notícias Le Monde Francês.⁷ As faces são associadas aos antropônimos e os logos aos topônimos pertencentes às categorias de produtos, organizações, empresas, eventos e instituições.

Cada entidade nomeada recebe uma pontuação indicando se é uma boa anotação ou não. Diversas pontuações são calculadas com base em estatísticas similares às de Salton e McGill (1983) e utilizando os valores de:

- Frequência de entidade (f_{ij}) – número de ocorrências de uma entidade i em um documento j ;
- Frequência de documento (df_i) – número de documentos nos quais a entidade i ocorre;
- Frequência de anotação (af_i) – número de imagens anotadas com a entidade i após a etapa de anotação. Esse número engloba as anotações corretas e as incorretas (abordagem não-supervisionada);

⁷Antropônimos são os substantivos que se referem a nomes próprios, e topônimos são os substantivos que se referem a nomes próprios de lugares.

- Frequência de anotação aprendida (laf_i) – número de imagens anotadas com a entidade i após a etapa de treino (abordagem supervisionada).

A Tabela 3.25 lista as diferentes pontuações (f) consideradas por Tirilly et al. (2010).

Figura 3.25: Fórmulas utilizadas para o cálculo da pontuação das entidades nomeadas em (TIRILLY et al., 2010).

f	frequency	f_{ij}
f-idf	frequência e frequência de documento inversa	$f_{ij} \cdot \log\left(\frac{N}{df_i}\right)$
f-df	frequência e frequência de documento	$f_{ij} \cdot \left(1 + \frac{df_i}{N}\right)$
f-iaf	frequência e frequência de anotação inversa	$f_{ij} \cdot \log\left(\frac{N}{af_i}\right)$
f-af	frequência e frequência de anotação	$f_{ij} \cdot \left(1 + \frac{af_i}{N}\right)$
f-ilaf	frequência e frequência de anotação aprendida inversa	$f_{ij} \cdot \log\left(\frac{N}{laf_i}\right)$
f-laf	frequência e frequência de anotação aprendida	$f_{ij} \cdot \left(1 + \frac{laf_i}{N}\right)$

Fonte: (TIRILLY et al., 2010)

Nos experimentos relatados em (TIRILLY et al., 2010), as entidades foram detectadas e caso estivessem presentes nas anotações de referência eram consideradas anotações corretas. A precisão de anotação de cada imagem foi calculada como mostra equação 3.2, onde N é o total de número de imagens anotadas.

$$Precisao = \frac{1}{N} \sum_{i=1}^n \frac{anotacoesCorretasImagem_i}{anotacoesImagem_i} \quad (3.2)$$

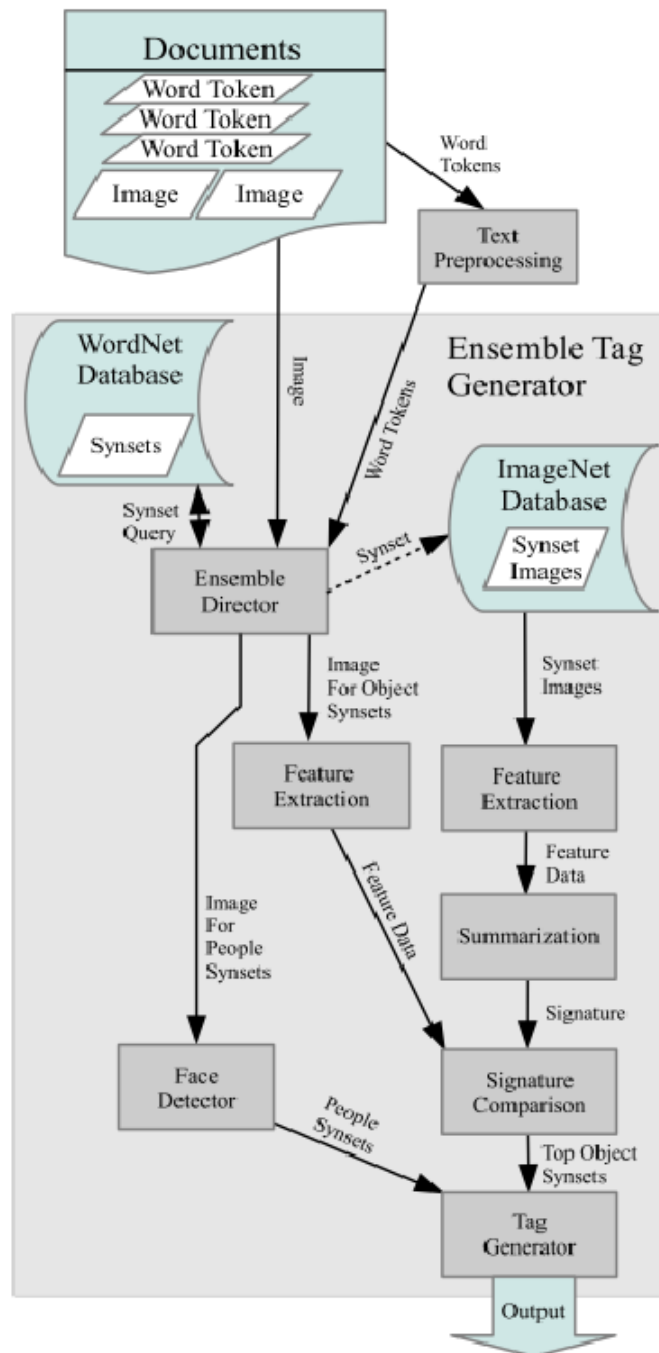
A precisão da tarefa de anotação de imagens no sistema proposto por Tirilly et al. (2010) variou de 30% a 60% para faces e logos. Os autores não relatam os valores de cobertura.

3.2.2.5 Noel e Peterson (2013)

Em (NOEL; PETERSON, 2013), é apresentado um sistema que utiliza a WordNet (FELLBAUM, 1998) em conjunto com a ImageNet (DENG et al., 2009) para a anotação de imagens. Esse sistema pode ser aplicado a qualquer documento que possua uma imagem rodeada por texto. A Figura 3.26 apresenta a arquitetura completa do sistema proposto em (NOEL; PETERSON, 2013).

Para cada documento, inicialmente são extraídos: (1) a legenda da imagem e (2) um parágrafo antes e um parágrafo depois da imagem. Esses textos são, então, processados (*Text Pre-processing*) para a remoção de *stopwords*. A lematização não é necessária pois a WordNet reconhece várias formas diferentes para cada palavra. Em seguida, é gerada uma lista de *syn-sets* para cada palavra por meio da consulta à WordNet. Mais especificamente, cada palavra

Figura 3.26: Arquitetura do sistema proposto por Noel e Peterson (2013).



Fonte: (NOEL; PETERSON, 2013)

passa pelo *Ensemble Director* (OPITZ; MACLIN, 1999) que é responsável por varrer a hierarquia da WordNet e detectar a qual das duas categorias usadas no trabalho a palavra pertence: pessoas ou demais categorias.

Caso o *Ensemble Director* identifique a palavra como uma pessoa, a imagem associada ao documento é enviada para um detector de face (*Face Detector*). Se uma face for detectada na

imagem, a palavra é marcada como uma palavra em potencial para a anotação. O detector de faces utiliza o algoritmo Haar-like-features (VIOLA; JONES, 2001) a partir de dados treinados da biblioteca OpenCV. Segundo os autores, esse algoritmo detecta com precisão as faces frontais mas tem menos precisão para as faces laterais e imagens pequenas.

Por outro lado, se a palavra pertencer a uma categoria diferente de “pessoa”, a imagem passa pelo processo de extração de características (*Feature Extraction*). Nesse processo, a imagem é segmentada em regiões utilizando o algoritmo de Felzenszwalb e Huttenlocher (2004) e para cada região são extraídas características de cor e textura. Essas características são reduzidas para amostras chamadas de vetores de média, que contêm a representação das características da imagem. A partir desses vetores é construído um histograma normalizado representando a imagem.

No caso das imagens vindas da ImageNet, apenas um histograma normalizado é gerado para cada *synset* proposto pelo *Emsemble Director*, utilizando os vetores de média representando as características comuns entre todas as imagens. Esses histogramas são chamados de assinatura da imagem. Essas assinaturas são comparadas utilizando o algoritmo EMD de (RUBNER; TOMASI; GUIBAS, 2000)(*Signature Comparison*), que funciona muito bem em recuperação de imagem. O EMD fornece uma quantificação numérica entre as diferenças dos histogramas da imagem pesquisada e de cada *synset*. Com o EMD, as palavras melhor classificadas são consideradas mais prováveis de estarem representadas dentro da imagem.

Para a avaliação do sistema, utilizou-se o banco de dados INEX (DENOYER; GALLINARI, 2007). Esse banco de dados possui mais de 600.000 documentos da Wikipedia e a maioria dos documentos possui legendas para as imagens. Anotadores humanos na Amazon Mechanical Turk (AMT) classificaram a relevância da anotação gerada pelo sistema em uma escala de 1 a 5, para cada palavra. Os resultados foram comparados com o ALIPR (LI; WANG, 2008) devido a sua disponibilidade e aceitação na comunidade científica. As 50 palavras melhor qualificadas geradas pelo ALIPR, para cada imagem, foram enviadas aos anotadores humanos da AMT para que fossem classificadas de forma semelhante às imagens do INEX.

O sistema foi, então, avaliado segundo dois quesitos: relevância da anotação e especificidade. A relevância é calculada como a média das notas atribuídas pelos anotadores humanos da AMT às cinco palavras melhor classificadas e quanto maior a relevância de uma palavra, melhor ela descreve o conteúdo da imagem. Nesse quesito, o sistema proposto em (NOEL; PETERSON, 2013) obteve uma relevância de 1,98 contra 1,62 do ALIPR. A especificidade, por sua vez, mede quão precisa é uma palavra para descrever um conceito. Essa medida de especificidade foi calculada a partir da profundidade das palavras na hierarquia da WordNet e quanto maior

for o valor da medida, mais específica é a palavra. Nesse quesito, o sistema proposto em (NOEL; PETERSON, 2013) obteve uma especificidade média de 9,09 contra 6,43 do ALIPR.

A Tabela 3.4 apresenta a quantidade de imagens e os resultados da avaliação do sistema proposto em (NOEL; PETERSON, 2013), para cada categoria. Essas categorias foram escolhidas devido à quantidade de imagens disponíveis e os melhores resultados. No total, há 6894 documentos e 1136 imagens. Nem todos os documentos contêm imagens e alguns documentos possuem múltiplas imagens.

Tabela 3.4: Resultado dos testes por categoria para o sistema de (NOEL; PETERSON, 2013).


INEX Category	Image count	Avg Relev.	Avg Spec.
Elephants	18	1.75	9.15
Mountains	260	1.76	8.44
Aircraft	441	2.07	9.59
Dogs	215	2.00	10.10
Skyscrapers	75	2.00	8.75
Sailboats	23	2.07	8.67
Armored Vehicles	47	2.00	9.61
WWII Ships	22	1.82	8.93
Vegetables	56	2.16	9.02
Flowers	34	2.11	9.25

Fonte: (NOEL; PETERSON, 2013)

A Figura 3.27 apresenta as palavras mais bem classificadas pelo sistema proposto em (NOEL; PETERSON, 2013) (primeira coluna) e o ALIPR (última coluna) para uma imagem difícil de ser analisada. Como pode ser notado, o sistema proposto em (NOEL; PETERSON, 2013) encontrou características de cores e texturas comuns para cachorros, humanos e rua.

Figura 3.27: Exemplo de uma imagem e as respectivas anotações geradas pelo sistema proposto em (NOEL; PETERSON, 2013) e pelo ALIPR (LI; WANG, 2008).

Word (in order)	Definition	ALIPR words
dog	Domestic dog	people
human	family Hominidae	man-made
street	Thoroughfare	sport
sign	A public display	car
street	Thoroughfare (variant 2)	cloth
control	Operates a machine	plane
sign	Advertising board	guard
retriever	Dog variant	parade
people	Group of humans	sky
blind	A protective covering	race
dog	Supports for fireplace logs	motorcycle



Fonte: (NOEL; PETERSON, 2013)

3.3 Alinhamento texto-imagem

Como consequência da descrição do conteúdo da imagem, investigada pelos métodos de anotação, surge um questionamento: no exemplo dado no início da seção anterior referente à anotação de uma imagem contendo uma menina, um cachorro e uma bola, sabe-se que existe um cachorro, uma menina e uma bola na imagem, mas onde esses elementos estão localizados? Como resposta a esse questionamento, surgiu a proposta do alinhamento de regiões da imagem com palavras. O objetivo desse alinhamento texto-imagem é rotular as regiões da imagem (segmentos) com as palavras que melhor as descrevem.

A tarefa de alinhamento texto-imagem foi inspirada no alinhamento de textos paralelos (como (OCH; NEY, 2003)) amplamente utilizado na geração de tradutores automáticos (como (KOEHN et al., 2007)). Os textos paralelos (ou bi-textos) são textos escritos em uma língua, acompanhados de suas traduções pra outra língua. O alinhamento desses textos paralelos pode ocorrer em diversos níveis como parágrafos, sentenças e palavras. A maior dificuldade do processo de alinhamento é determinar precisamente a correspondência entre as palavras nas sentenças paralelas. Análogo ao alinhamento de textos paralelos, no alinhamento de regiões da imagem utiliza-se como entrada para a ferramenta de alinhamento um par paralelo composto por uma imagem e um texto (DUYGULU et al., 2002). Assim como ocorre nos textos paralelos, sabe-se que o texto está relacionado com a imagem, mas a dificuldade em encontrar as correspondências corretas entre os elementos da imagem e os elementos do texto é ainda maior.

A seguir, são apresentados alguns poucos trabalhos encontrados na literatura com o objetivo semelhante ao do trabalho aqui proposto: alinhar regiões da imagem com palavras presentes no texto que a acompanha. São eles: Pham, Moens e Tuytelaars (2008), que visa alinhar nomes (palavras) nas legendas das imagens com faces (regiões da imagem) presentes nessas imagens; Tegen et al. (2014), que visa ligar as regiões da imagem a palavras que aparecem nas legendas anotadas manualmente por humanos; e Redmon et al. (2016), que criou uma rede convolucional para detecção e localização de objetos em imagens.

3.3.1 Pham, Moens e Tuytelaars (2008)

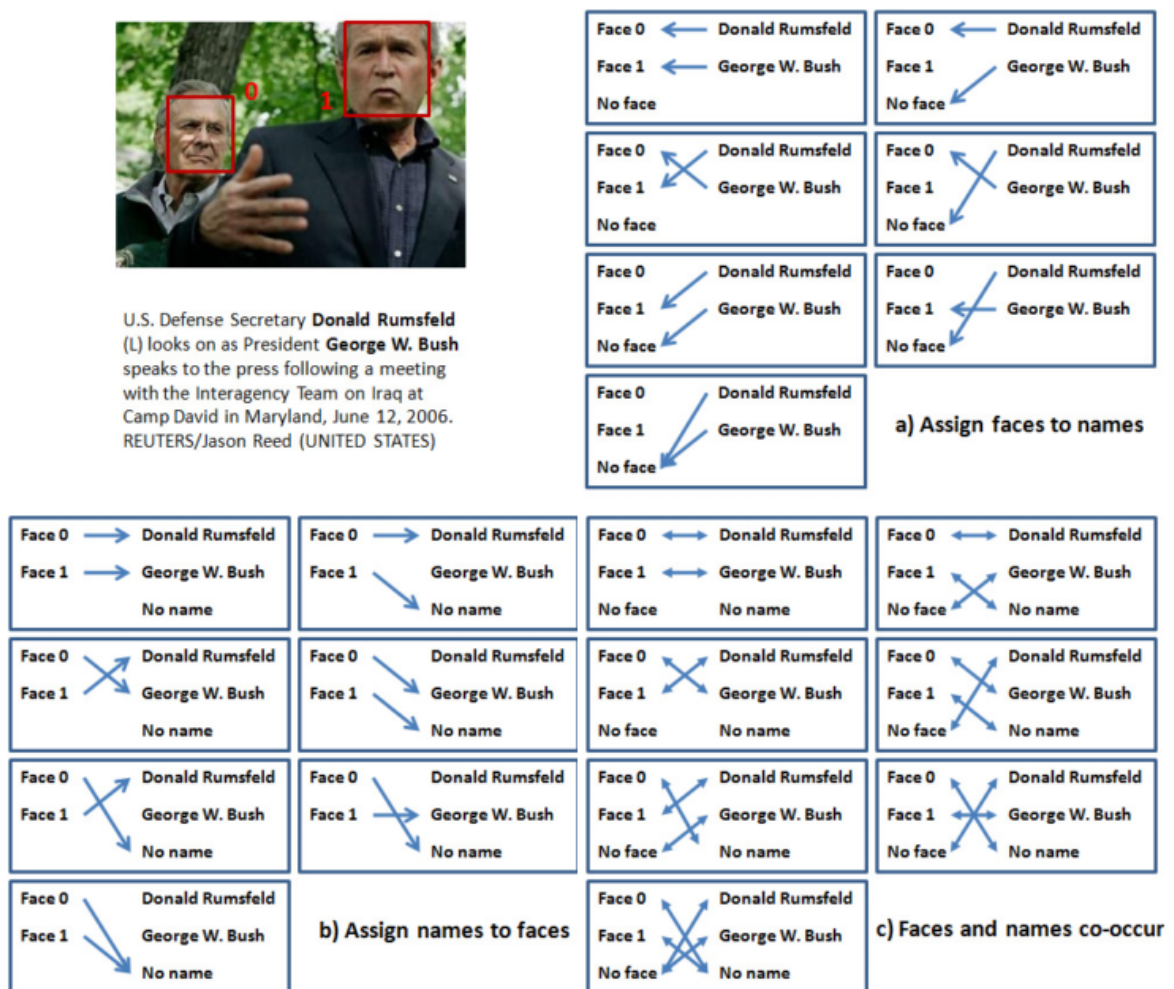
Em (PHAM; MOENS; TUYTELAARS, 2008), o objetivo é alinhar os nomes presentes nas legendas de imagens com suas faces correspondentes nas imagens. Nesse caso, o problema de alinhamentos de omissão (1-0 ou 0-1), encontrado no alinhamento de palavras em textos paralelos, também ocorre aqui. Mais especificamente, como nem todos os nomes que aparecem nas legendas possuem uma correspondência com alguma face na imagem (e vice-versa), alinha-

mentos de omissão podem ocorrer.

Visando a resolução desse problema, a partir de evidências empíricas, três abordagens diferentes para o alinhamento texto-imagem foram propostas por Pham, Moens e Tuytelaars (2008): (a) a primeira trabalha a partir da suposição de que os nomes presentes em um texto geram as faces de uma imagem, (b) a segunda supõe que as faces de uma imagem geram os nomes do texto e (c) a terceira advoga que há uma co-ocorrência entre nomes e faces.

A Figura 3.28 mostra todas as possibilidades de alinhamento de acordo com as três abordagens citadas acima. Além das entidades “face” e “nome” estão presentes também as entidades “nenhuma face” (*no face*) e “nenhum nome” (*no name*) indicando, respectivamente, a existência de nomes no texto que não possuem uma face visível na imagem ou de faces que aparecem na imagem e não possuem um nome no texto.

Figura 3.28: Exemplo de uma imagem acompanhada de um texto e todas as possibilidades de alinhamento texto-imagem de acordo com as três abordagens propostas em (PHAM; MOENS; TUYTELAARS, 2008).



Fonte: (PHAM; MOENS; TUYTELAARS, 2008)

Alguns passos de pré-processamento foram realizados com textos e imagens antes de realizar o alinhamento texto-imagem, são eles:

- Pré-processamento nos textos – Os textos foram processados para permitir: (1) o reconhecimento de entidades nomeadas usando a ferramenta do pacote OpenNLP⁸ em conjunto com nomes extraídos do site Wikipedia⁹ para aumentar a precisão da ferramenta, (2) a resolução de anáfora¹⁰ nos nomes previamente reconhecidos realizada pela ferramenta do pacote LingPipe¹¹ e (3) o agrupamento das entidades reconhecidas no texto usando um algoritmo hierárquico baseado na similaridade de cosseno para a associação de grupos. O agrupamento de nomes é necessário para evitar que dois nomes se referindo à mesma pessoa herdem as faces da imagem, causando distúrbio nos cálculos probabilísticos.
- Pré-processamento nas imagens – As imagens foram processadas para permitir: (1) a detecção de faces utilizando a abordagem da OpenCV (VIOLA; JONES, 2004) e das características faciais com a implementação de (EVERINGHAM; SIVIC; ZISSERMAN, 2006) e (2) o agrupamento das faces, realizado de modo similar ao agrupamento de nomes, utilizando um algoritmo de agrupamento em conjunto com similaridade de cosseno. Nesse processo, foi definido que faces encontradas em uma mesma imagem não podem ser agrupadas no mesmo grupo, evitando assim erros no agrupamento. Além disso, sabe-se que o tratamento de faces é um problema complexo, uma vez que imagens de faces apresentam diferentes condições de luminosidade, mudanças de pose e expressões faciais. Para lidar com esse problema foi utilizado um modelo de faces transformáveis 3D (SMET; FRANSENS; GOOL, 2006). Esse modelo 3D permite a estimativa de poses e luminosidade e elimina mudanças irrelevantes entre as faces. O modelo retorna 40 componentes de textura e 40 para formato.

Vale mencionar que os grupos de faces e textos foram construídos de formas semelhantes para que pudessem ser utilizados pelos cálculos de probabilidade de maneira semelhante. Assim, após o pré-processamento de textos e imagens, o alinhamento texto-imagem foi realizado por meio de uma abordagem probabilística usando o algoritmo *Expectation Maximization* (EM) (DEMPSTER; LAIRD; RUBIN, 1977) tendo como base a co-ocorrência entre faces e nomes.

O sistema foi testado e avaliado no site de notícias Yahoo! News (BERG et al., 2004) com nomes e faces manualmente anotadas por Huang et al. (2007). Esse banco de dados possui

⁸Disponível em: <http://opennlp.sourceforge.net/>

⁹Disponível em: <http://en.wikipedia.org/>

¹⁰A resolução de anáfora ajuda a agrupar nomes como “Lula” e “Luiz Inácio Lula da Silva”.

¹¹Disponível em: <http://www.alias-i.com/lingpipe/>

11.820 pares de texto-imagem com a média de 2 nomes por texto e 1,12 faces por imagem. Após o pré-processamento, foram encontrados 9.793 grupos de nomes reconhecidos com uma precisão de 81%¹² em um conjunto de validação com 100 imagens-texto. Já o agrupamento de imagens retornou 749 faces. Usando um pequeno conjunto de validação, verificou-se que a relação entre os grupos de face gerados e os grupos de faces de um conjunto de referência foi no máximo de 48%.

A Figura 3.29 mostra os resultados do sistema de alinhamento texto-imagem proposto em (PHAM; MOENS; TUYTELAARS, 2008). Nessa Figura, as indicações “a)” e “e)” representam a primeira abordagem (atribuição de faces aos nomes), as indicações “b)” e “f)”, a segunda abordagem (atribuição de nomes às faces) e as demais se referem à terceira abordagem (coocorrência entre nomes e faces). A precisão para a primeira abordagem foi calculada como a razão entre as faces corretamente nomeadas sobre o número de nomes. Para a segunda abordagem, a precisão foi calculada como a razão entre nomes corretamente atribuídos sobre o número de faces. Para a terceira abordagem, a precisão foi calculada como a razão entre os pares de face-nome corretos sobre o número total de pares face-nome. Baseado nos resultados da Figura 3.29 nota-se que a primeira abordagem obteve melhores resultados.

Figura 3.29: Valores de precisão para as três abordagens propostas (para a primeira iteração do algoritmo EM e para o EM completo).

Accuracy	After 1 iteration	full EM
Likelihood type $L^{(n \rightarrow f)}$ a)	71.15%	71.24%
Likelihood type $L^{(f \rightarrow n)}$ b)	58.71%	60.15%
Likelihood type $L^{(n, f)}$ using $P(f n)$ c)	69.51%	69.58%
Likelihood type $L^{(n, f)}$ using $P(n f)$ d)	60.98%	62.70%
Likelihood type $L^{(n^* \rightarrow f)}$ e)	71.49%	71.61%
Likelihood type $L^{(f \rightarrow n)}$ f)	57.83%	59.16%
Likelihood type $L^{(n^*, f^*)}$ using $P(f n)$ g)	69.58%	69.70%
Likelihood type $L^{(n^*, f^*)}$ using $P(n f)$ h)	60.49%	61.90%

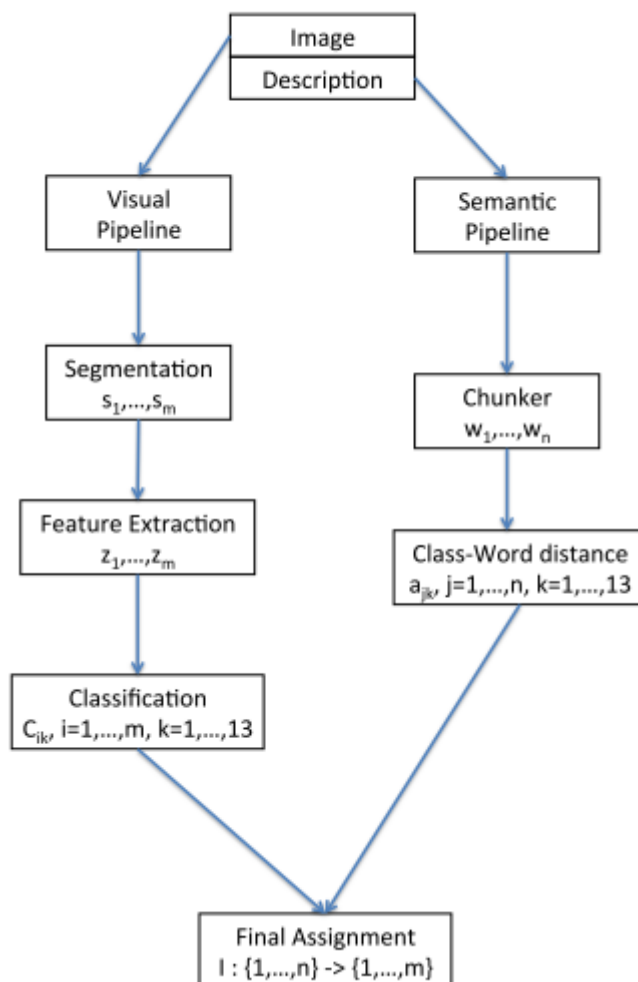
Fonte: Adaptado de (PHAM; MOENS; TUYTELAARS, 2008)

3.3.2 Tegen et al. (2014)

Visando aprender a forma como os humanos descrevem uma imagem nas legendas, em (TEGEN et al., 2014) é proposto um sistema cujo objetivo é ligar as regiões da imagem às palavras que aparecem nas legendas. Para tanto, o sistema proposto, descrito na Figura 3.30, utiliza características visuais da imagem e características textuais da legenda da imagem.

¹²Os autores não informaram os valores de cobertura.

Figura 3.30: Arquitetura do sistema proposto para alinhamento texto-imagem em (TEGEN et al., 2014).



Fonte: (TEGEN et al., 2014)

O trabalho foi realizado com base no banco de imagens SAIAPR TC-12 (ESCALANTE et al., 2010; GRUBINGER et al., 2006) que contém 20.000 imagens anotadas. Além das anotações de referência presentes no SAIAPR, anotações adicionais foram feitas para a obtenção de um conjunto de validação confiável (conjunto de referência). Para tanto, as regiões das imagens do SAIAPR foram rotuladas manualmente com um vocabulário de 282 rótulos. Devido à variada frequência com que esses rótulos aparecem, foram escolhidas as 100 palavras mais frequentes e estas foram divididas em 13 classes visuais¹³: água, céu, vegetação, construção, humanos, objetos da casa, solo, animal, veículo, montanha, estrada, piso e tecidos.

O sistema proposto realiza duas etapas: treinamento e anotação das imagens. Na etapa de

¹³Apenas 13 classes foram consideradas devido ao agrupamento de palavras semelhantes (que possuem o mesmo significado) numa mesma classe.

treinamento ocorre: (1) o treinamento utilizando as regiões anotadas para estimar à qual das 13 classes visuais pertence a região (*Visual Pipeline*, ilustrado na Figura 3.30) e (2) a remoção de palavras que estão na legenda mas não foram anotadas manualmente (*Semantic Pipeline*, ilustrado na Figura 3.30).

O treinamento das regiões inicia-se com a obtenção das características visuais, por meio da segmentação das imagens em regiões utilizando o segmentador CPMC (CARREIRA; SMINCHISESCU, 2010). O CPMC segmenta a imagem a partir da similaridade entre *pixels* adjacentes e a partir de uma classificação de qualidade dos segmentos e agrupa as regiões com maior potencial em uma imagem utilizando uma implementação do Random Forest (BREIMAN, 2001) disponível na biblioteca LIBLINEAR (FAN et al., 2008). Como parâmetro para o CPMC foi definido um número de regiões de 500 à 1.000.

Em seguida, para cada região da imagem, são extraídas características visuais como: largura e altura, média e desvio padrão dos eixos x e y, bordas, convexidade, desvio padrão e assimetria dos espaços de cores RGB e CIE-Lab. Essas características também foram utilizadas por Carbonetto (2003). As características extraídas são, então, usadas no treinamento de utiliza as regiões já anotadas anteriormente como dados de treinamento e estima a probabilidade de cada região pertencer a uma das 13 classes.

A segunda etapa de treinamento (via *Semantic Pipeline*) objetiva remover as palavras de uma legenda que não foram anotadas manualmente, significando que essas palavras removidas não possuem características visuais. Essas palavras são armazenadas e toda vez que aparecem em uma legenda são ignoradas na tarefa de anotação da imagem. Para a extração das palavras pertencentes à legenda das imagens, foi aplicado um analisador sintático (*chunker*) (LAI; HOCKENMAIER, 2014) que separa o texto em sintagmas. Os sintagmas nominais foram escolhidos para a extração das palavras-chave pois geralmente fazem menção a objetos e lugares. A palavra mais à direita é escolhida dentro de cada sintagma. As palavras que não foram descartadas são consideradas palavras-chave.

A etapa final do sistema consiste em anotar as regiões de uma imagem com as palavras-chave contidas na legenda. Cada região da imagem encontrada pelo segmentador CPMC é classificada em uma das 13 classes visuais pelo regressor LIBLINEAR. Em alguns casos há regiões que não correspondem a nenhuma das classes visuais. Quando isso acontece, a região não é anotada. Para cada palavra-chave é calculada a probabilidade de pertencer a uma das 13 classes visuais. Essa probabilidade é estimada utilizando a hierarquia da WordNet para calcular a distância entre a palavra-chave e as classes visuais.

Cada região da imagem é anotada com a palavra-chave que possui a classe visual idêntica à

da região. Caso mais de uma palavra remeta a uma mesma classe visual, escolhe-se a palavra-chave que possui um maior valor de distância calculado pela WordNet.

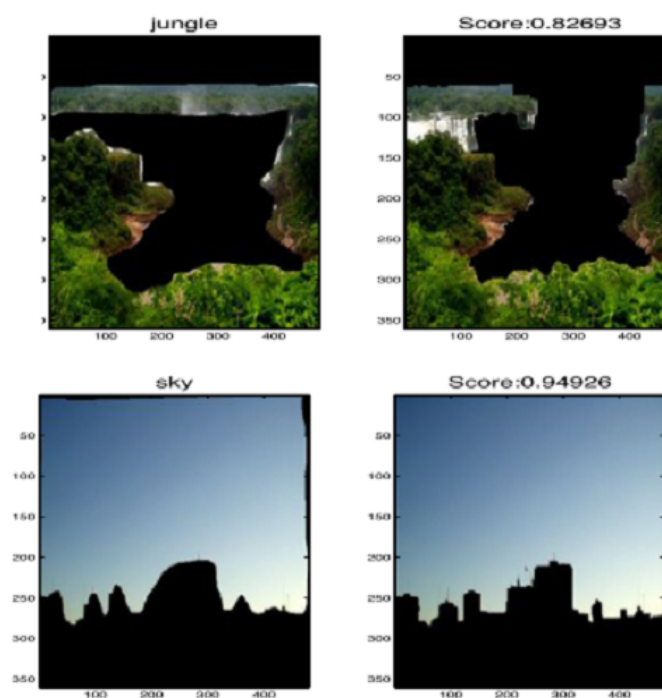
O sistema foi testado com 142 imagens do banco SAIAPR. Para essas imagens, foram encontradas 1.109 palavras-chave das quais apenas 754 possuíam regiões anotadas no conjunto de referência. Das 754 regiões, o sistema produziu anotação para apenas 482. Dessas, 466 palavras-chave foram atribuídas às mesmas regiões que as anotações de referência e 16 palavras foram atribuídas a regiões pelo sistema mas nas anotações a palavra foi anotada como “palavra sem região”. Levando em consideração todas as 754 palavras-chave, o sistema deixou de anotar 288 palavras que verdadeiramente possuíam uma região.

Para a avaliação da ligação entre as regiões da imagem geradas pelo CPMC e as regiões do conjunto de referência foi utilizado o índice de Jaccard (1901) que é calculado pela equação a seguir:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.3)$$

sendo A o conjunto de *pixels* pertencentes às regiões de referência e B o conjunto de *pixels* gerados pelo sistema. O índice de Jaccard tem como valor mínimo 0 e valor máximo 1. Quanto maior o valor do índice, mais semelhantes são as regiões avaliadas. Alguns resultados com as pontuações medidas pelo índice de Jaccard podem ser visualizados na Figura 3.31.

Figura 3.31: Alguns resultados de classificação do sistema proposto em (TEGEN et al., 2014) e o respectivo valor para o índice de Jaccard (1901)



Fonte: (TEGEN et al., 2014)

Capítulo 4

O ALINHADOR TEXTO-IMAGEM LINKPICS

O alinhador LinkPICS, descrito neste documento, tem a tarefa de alinhar os elementos presentes no texto de uma notícia com os elementos presentes na imagem associada. A seção 4.1 descreve a arquitetura do LinkPICS e, em suas subseções, cada uma das etapas envolvidas no processo de alinhamento texto-imagem proposto neste trabalho. Por fim, a seção 4.2 ilustra a saída gerada pelo LinkPICS.

4.1 Arquitetura do LinkPICS

A arquitetura proposta para o LinkPICS está baseada em (NOEL; PETERSON, 2013), que faz distinção entre “pessoa” e outras categorias, aqui chamadas de “objeto”, como ilustrado na Figura 3.26; e (TEGEN et al., 2014), que divide o alinhamento em processamento de imagem e de texto, como no LinkPICS, como ilustrado na Figura 3.30.

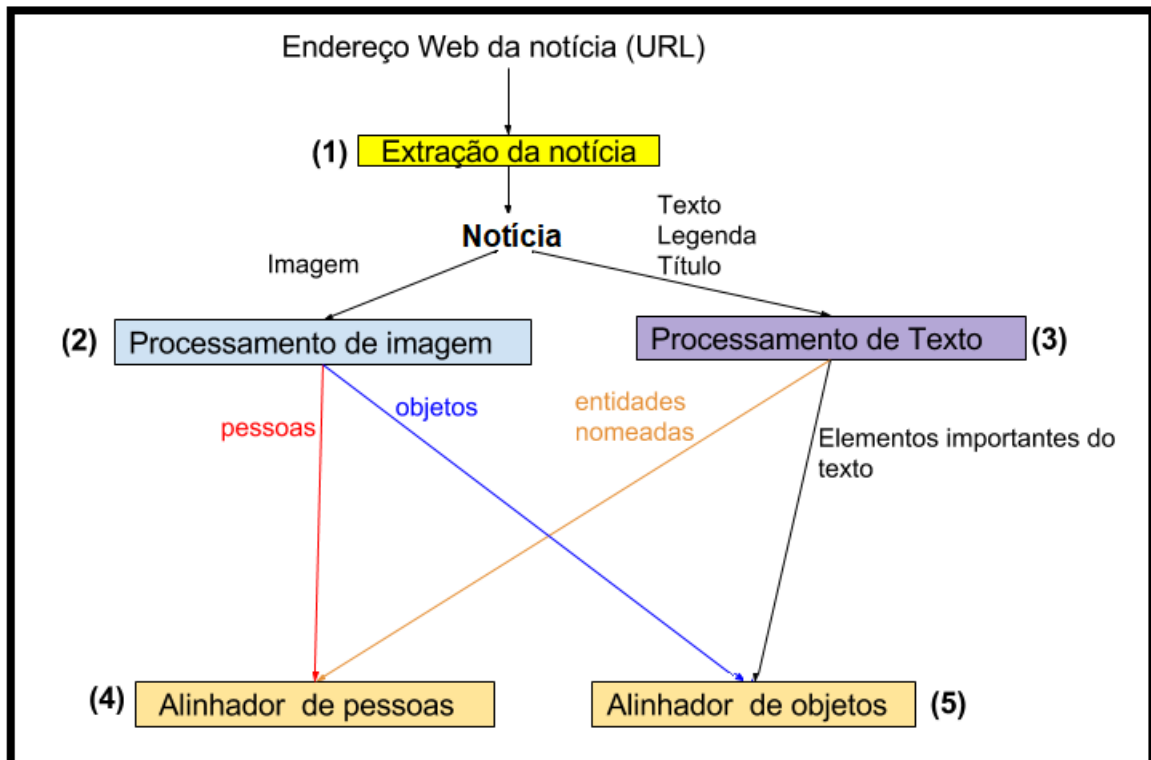
No LinkPICS, o alinhamento texto-imagem é realizado em cinco etapas, conforme ilustrado na Figura 4.1 e detalhado nas próximas subseções: (1) extração da notícia, (2) processamento de imagem, (3) processamento de texto, (4) alinhador de pessoas e (5) alinhador de objetos.

4.1.1 Extração da notícia

Esta etapa consiste na obtenção da notícia. A partir do endereço *web* da notícia de um jornal *online* são extraídos: o texto da notícia, seu título, a legenda da imagem e a imagem associada ao texto. A Figura 4.2 ilustra essas 4 partes da notícias extraídas do site do jornal Folha de São Paulo¹.

¹Disponível em: <http://www1.folha.uol.com.br/internacional/en/brazil/2016/05/1767623-olympic-games-will-be-a-major-success-says-rousseff.shtml>. Acesso em: 05 nov. 2017.

Figura 4.1: Arquitetura do LinkPICS. O alinhador é dividido em etapas que lidam com texto e imagem, separadamente.



A imagem obtida nesta etapa é, então, enviada como entrada para o processamento de imagem (descrito na seção 4.1.2) enquanto os demais elementos textuais extraídos são fornecidos como entrada para a etapa de processamento de texto (descrito na seção 4.1.3).

4.1.2 Processamento de imagem

Nesta seção, o objetivo é descrever o processo de detecção de objetos e pessoas em uma imagem utilizando as ferramentas descritas a seguir. A Figura 4.3 ilustra os passos do processamento de imagem.

4.1.2.1 Detecção de pessoas e objetos usando a YOLO

A YOLO² (veja mais detalhes sobre a YOLO na seção 2.2.1.1) é uma CNN capaz de detectar pessoas, veículos, animais, aparelhos eletrônicos e outros objetos. Cada detecção possui uma localização na imagem (*bounding box*) e um rótulo que descreve o que foi detectado.

Para avaliar a aplicabilidade da CNN YOLO neste trabalho, todas as imagens do corpus da Folha Internacional (veja seção 5.1.1) foram submetidas à YOLO e as saídas foram analisadas

²Disponível em: <http://pjreddie.com/darknet/yolo/>

Figura 4.2: Elementos extraídos da notícia de jornal: (1) título da notícia, (2) texto da notícia, (3) imagem associada ao texto e (4) legenda da imagem.


'Olympic Games Will Be a Major Success', Says Rousseff (1)

05/04/2016 - 10H31

Two-time Olympic volleyball champion Fabiana Claudino runs holding the torch, and around her, in the landscape framed by the Planalto Palace, on the left side, and the National Congress, on the right, protesters follow the athlete screaming. They display banners in favor of President Dilma Rousseff and against her.

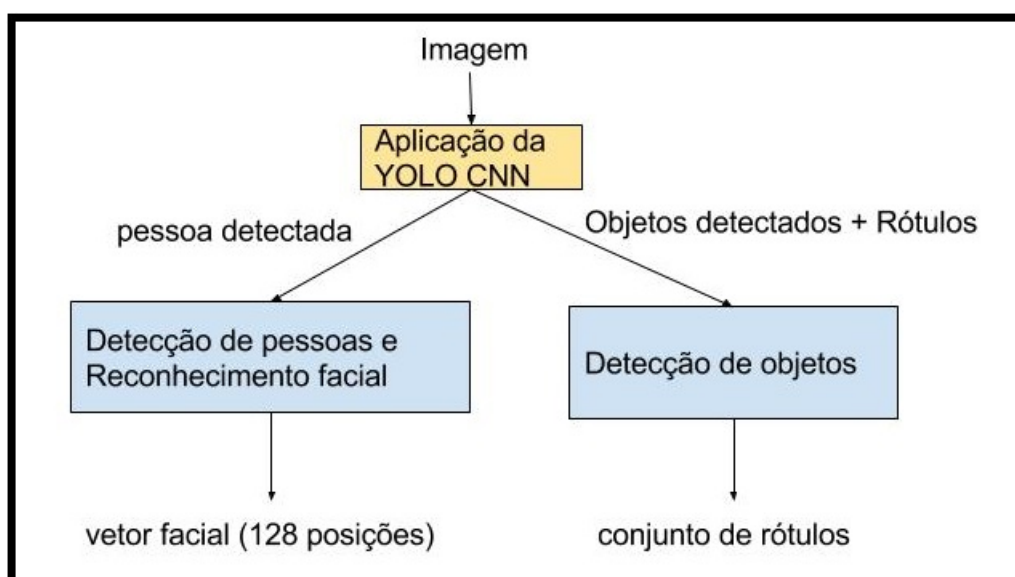
Thus the start of the Olympic torch relay on Tuesday (3), in Brasília, brought side by side the political crisis in the country and the Olympics, universes that Olympic officials try so much to separate.

Before passing the Olympic fire to the first torchbearer on national soil, volleyball player Fabiana, Rousseff said in a speech that the country will be able to successfully promote the Rio Games, in spite of the crisis.



Brazilian volleyball player Fabiana Claudino (L) applauds as Brazilian President (4) Dilma Rousseff holds the Olympic torch

Figura 4.3: Etapa de processamento da imagem do alinhador LinkPICS.



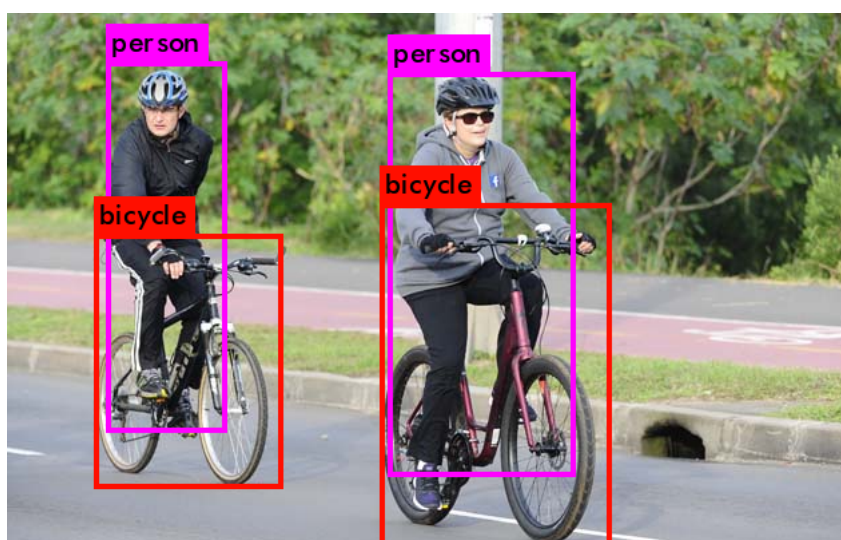
manualmente. Como resultado constatou-se uma precisão superior a 95% na detecção de

peças. Essa precisão é de suma importância para o trabalho proposto, devido ao corpus conter muitas notícias relacionadas a pessoas.

Na detecção de objetos os resultados foram satisfatórios. As categorias de objetos “bicicleta”, “avião” e “carro” foram detectadas com uma precisão superior a 90%. A Figura 4.4 traz um exemplo com a detecção de bicicletas e pessoas utilizando a YOLO. Uma característica da YOLO é a detecção de múltiplos objetos em uma mesma imagem. Note que a YOLO conseguiu detectar tanto as bicicletas, como as pessoas que estavam utilizando-as.

No LinkPICS, a saída da YOLO é a *bounding box* para cada pessoa detectada e a *bounding box* e o rótulo associado para cada objeto.

Figura 4.4: Exemplo de aplicação da YOLO na detecção de bicicletas e pessoas.

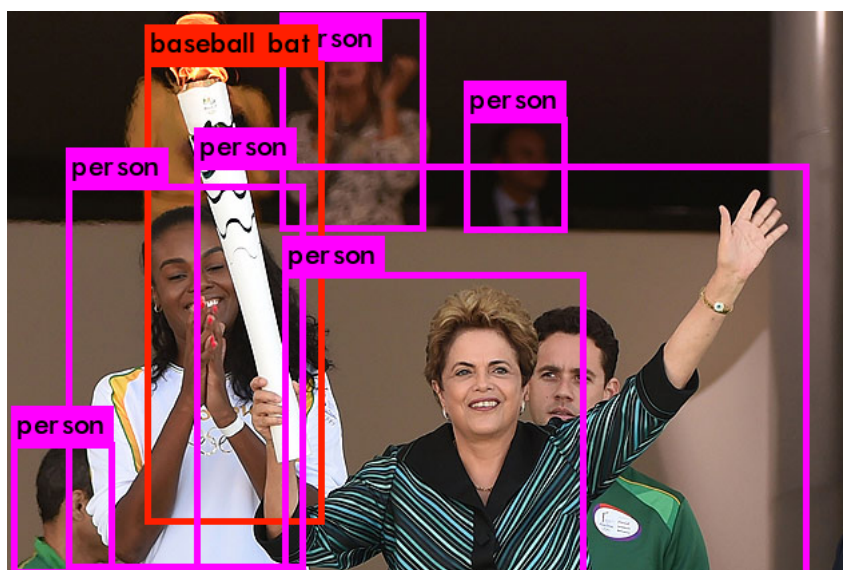


Embora a detecção de objetos realizada pela YOLO seja muito boa, ela não é perfeita. Por exemplo, a Figura 4.5 contém várias pessoas e também um objeto “tocha olímpica”. A YOLO detectou todas as pessoas da foto, entretanto, a tocha foi detectada e identificada como “taco de baseball”. Podemos notar que as características de um taco de baseball são realmente semelhantes às características de uma tocha olímpica. Esse erro de classificação de objetos não diminuiu a força da utilização da YOLO, pois, a partir de outras técnicas detalhadas na seção 4.1.2.3, é possível a troca do rótulo “taco de baseball” pelo rótulo “tocha olímpica”.

4.1.2.2 Detecção de pessoas e Reconhecimento Facial

A partir das detecções de pessoas da YOLO, foi aplicada a técnica de detecção e reconhecimento de faces. A ferramenta utilizada para essa tarefa é um módulo da biblioteca de

Figura 4.5: Exemplo de aplicação do YOLO. Nessa imagem o objeto “tocha olímpica” foi detectado com o rótulo “taco de baseball”.



aprendizado de máquina DLIB³ que obteve uma precisão de 99,38% no conjunto de dados *LFW-labeled faces in the wild* (HUANG et al., 2007).

A Figura 4.6 explica o funcionamento da DLIB. O primeiro passo é a detecção de faces da imagem. Em seguida, são extraídas as características faciais (vetor com 128 posições) de cada face. Essas características são comparadas com as características faciais extraídas das imagens do banco LFW utilizando a distância euclidiana. Caso alguma face semelhante seja encontrada no banco (neste caso considera-se semelhante uma comparação que resulte em uma distância menor ou igual a 0,6), a face objeto da consulta é atribuída à pessoa semelhante do banco LFW.

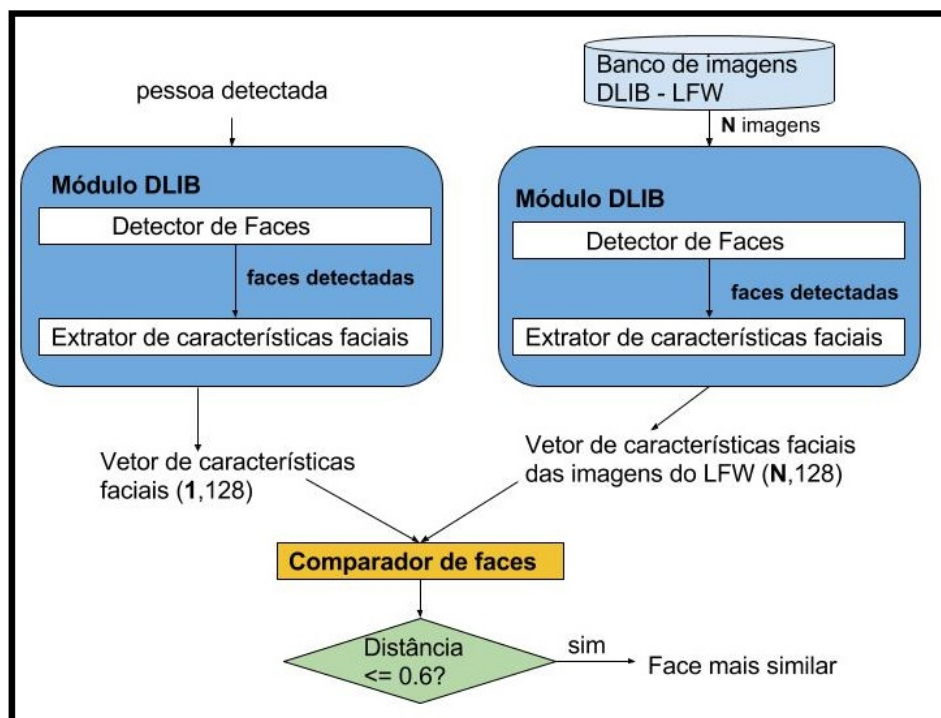
4.1.2.3 Detecção de objetos

Como a YOLO detecta somente 80 classes de objetos, optou-se também pela utilização de outras três CNNs para a detecção dos objetos presentes na imagem. As redes Extraction, DarkNet e DenseNet foram treinadas com base nas imagens da ImageNet (veja seção 2.2.3) e são capazes de classificar uma imagem em 1000 classes. A característica principal dessas redes é analisar a imagem e gerar a probabilidade de cada classe estar presente na imagem. As redes utilizadas nesse trabalho estão descritas a seguir:

- Extraction – Essa CNN foi desenvolvida a partir do modelo GoogleNet (SZEGEDY et al., 2015) e obteve a precisão de 72,5% TOP-1 e de 90,8% TOP-5 no conjunto de dados da

³Disponível em: http://dlib.net/face_recognition.py.html

Figura 4.6: Funcionamento da biblioteca DLIB para reconhecimento facial.



ImageNet⁴.

- Darknet19 448x448 – Modificação da CNN Extraction que obteve uma precisão de de 76,4% TOP-1 e 93,5% TOP-5 no conjunto de dados da ImageNet⁵.
- DenseNet 448x448⁶ – Rede proposta por (HUANG et al., 2016) que, segundo aqueles autores, obteve uma precisão de 77% TOP-1 e 93,7% TOP-5 no conjunto de dados da ImageNet.

Devido à característica das redes de analisar e classificar uma imagem inteira, optou-se por aplicá-las somente nos objetos (ou seja, nas *bouding boxes* de objetos) detectados pela YOLO. Por exemplo, as redes foram aplicadas no objeto “taco de baseball” da Figura 4.5 e obtiveram os seguintes resultados:

- TOP 5-Darknet : *torch* (tocha), *pole* (bastão), *whistle* (apito), *spray* (spray), *plunger* (êmbolo)
- TOP 5-Extraction: *torch* (tocha), *whistle* (apito), *band-Aid* (curativo adesivo), *candle* (vela), *maillot* (maiô)

⁴Disponível em: <http://pjreddie.com/darknet/imagenet/>.

⁵Disponível em: <http://pjreddie.com/darknet/imagenet/>.

⁶Disponível em: <http://pjreddie.com/darknet/imagenet/>.

- TOP 5-DENSENET: *torch* (tocha), *pole* (bastão), *spray* (spray), *whistle* (apito), *candle* (vela)

Todas as redes conseguiram detectar a presença de um objeto “tocha” (*torch*). Isso favoreceu a ideia de combinar os resultados da YOLO com os resultados das CNNs, aumentando a probabilidade de uma detecção de objetos mais confiável.

No LinkPICS, as três redes foram aplicadas em cada detecção de objeto da YOLO gerando cada uma, 5 palavras como possíveis rótulos de objetos dentro imagem. As 15 palavras foram, então, combinadas com o rótulo fornecido pela YOLO formando uma lista com até 16 rótulos.⁷

4.1.3 Processamento de texto

Essa seção descreve o passo a passo do processamento de texto e as ferramentas de PLN utilizadas no LinkPICS, conforme ilustrado na Figura 4.7. Todas as ferramentas foram aplicadas para tratar textos em inglês. Optou-se pela aplicação das ferramentas no idioma inglês por ser este o idioma com a maior disponibilidade de ferramentas e recursos de boa qualidade para o PLN. Contudo, vale ressaltar que toda a metodologia proposta neste trabalho é independente de idioma e pode ser facilmente replicada para outras línguas se houver a disponibilidade de ferramentas para elas.

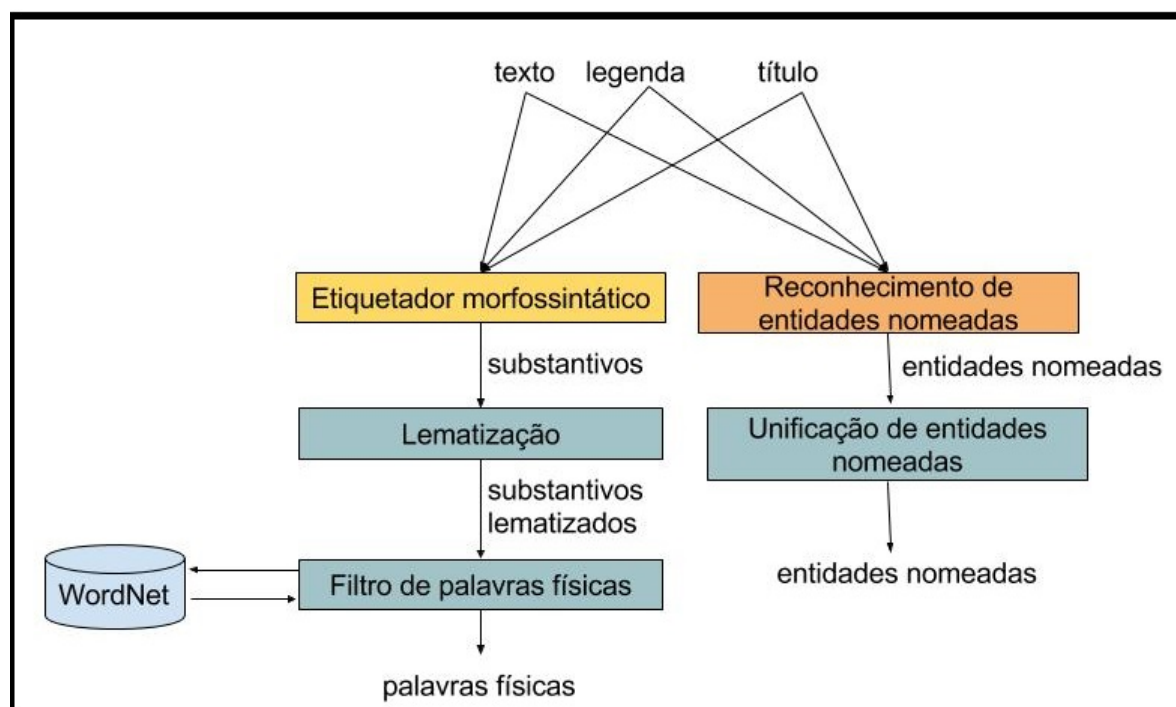
Um dos passos do processamento de texto é a aplicação do etiquetador morfossintático TreeTagger (SCHMID, 2013) no texto, legenda e título da imagem, para a identificação de todos os substantivos presentes nesses textos. Esses substantivos são lematizados pelo TreeTagger para a unificação das palavras que possuam a mesma forma canônica (veja seção 2.1.1).

Os substantivos foram escolhidos devido aos objetos ilustrados em uma imagem pertencerem a essa classe gramatical. Apesar de parte dos substantivos representarem objetos que podem ser ilustrados em imagens (entidades físicas), alguns substantivos do texto não possuem essa característica. Como tentativa de filtrar apenas entidades físicas, de modo similar à (SOCHER; FEI-FEI, 2010), foi utilizada a hierarquia da WordNet, que separa as palavras em duas classes: objeto físico e objeto abstrato. As palavras da classe objeto físico são utilizadas pelo LinkPICS para o alinhamento de objetos.

Outra etapa do processamento de texto é a obtenção das entidades nomeadas utilizando o reconhecedor de entidades nomeadas *Stanford Named Entity Recognizer* (FINKEL; GRENAGER; MANNING, 2005). O reconhecedor foi aplicado no texto, na legenda e no título da notícia.

⁷O tamanho máximo do conjunto de rótulos é de 16, uma vez que as palavras repetidas são removidas resultando em um conjunto de palavras únicas.

Figura 4.7: Etapa de processamento de texto do alinhador LinkPICS.



O passo seguinte foi unificar os nomes encontrados, como por exemplo na notícia da Figura 4.2 no qual o nome “Dilma Rousseff” aparece no texto e no título da notícia é tratado como “Rousseff”.

Para unificar os nomes, o primeiro passo é verificar se a palavra à direita de cada entidade nomeada também representa uma entidade nomeada. Em caso positivo, a entidade vizinha é agregada à entidade nomeada formando um nome composto. Se uma entidade nomeada possuir mais de uma entidade nomeada vizinha, todas as outras serão agregadas à primeira entidade.

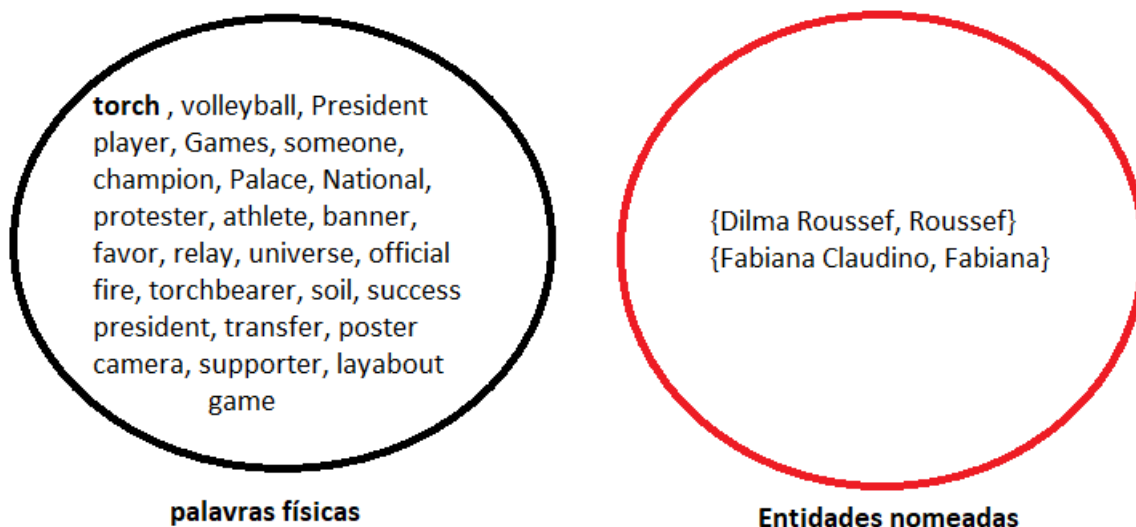
O segundo passo é encontrar e agrupar entidades da notícia que se referem a uma mesma pessoa. Como se referem à mesma pessoa, esses nomes devem ser tratados como uma única entidade. Para cada entidade, é verificado se ela está contida nas demais entidades nomeadas. Se estiver contida, ela é agrupada com sua respectiva entidade nomeada. Todas as entidades do grupo são alinhadas pelo LinkPICS a uma mesma região da imagem contendo pessoa.

O resultado obtido após a aplicação do processamento de texto à notícia da Figura 4.2 pode ser visualizado na Figura 4.8.

4.1.4 Alinhador de pessoas

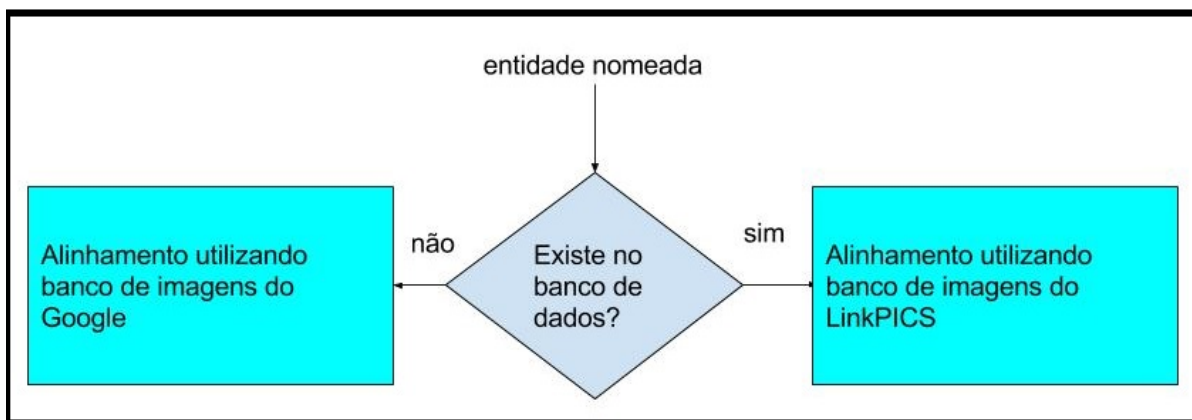
Similar à (PHAM; MOENS; TUYTELAARS, 2008) e (TIRILLY et al., 2010), no LinkPICS as entidades nomeadas são alinhadas às faces reconhecidas na imagem. Assim, após o pré-processa-

Figura 4.8: Exemplo de palavras físicas e entidades nomeadas extraídas da notícia da Figura 4.2 após o processamento do texto.



mento de imagem e textos da notícia, como descrito nas seções anteriores, o alinhador de pessoas segue o fluxo apresentado na Figura 4.9. Em seu processamento, ele faz uso de dois recursos: (1) o banco de imagens do LinkPICS e (2) o banco de imagens do Google.

Figura 4.9: Estrutura do alinhador de pessoas.

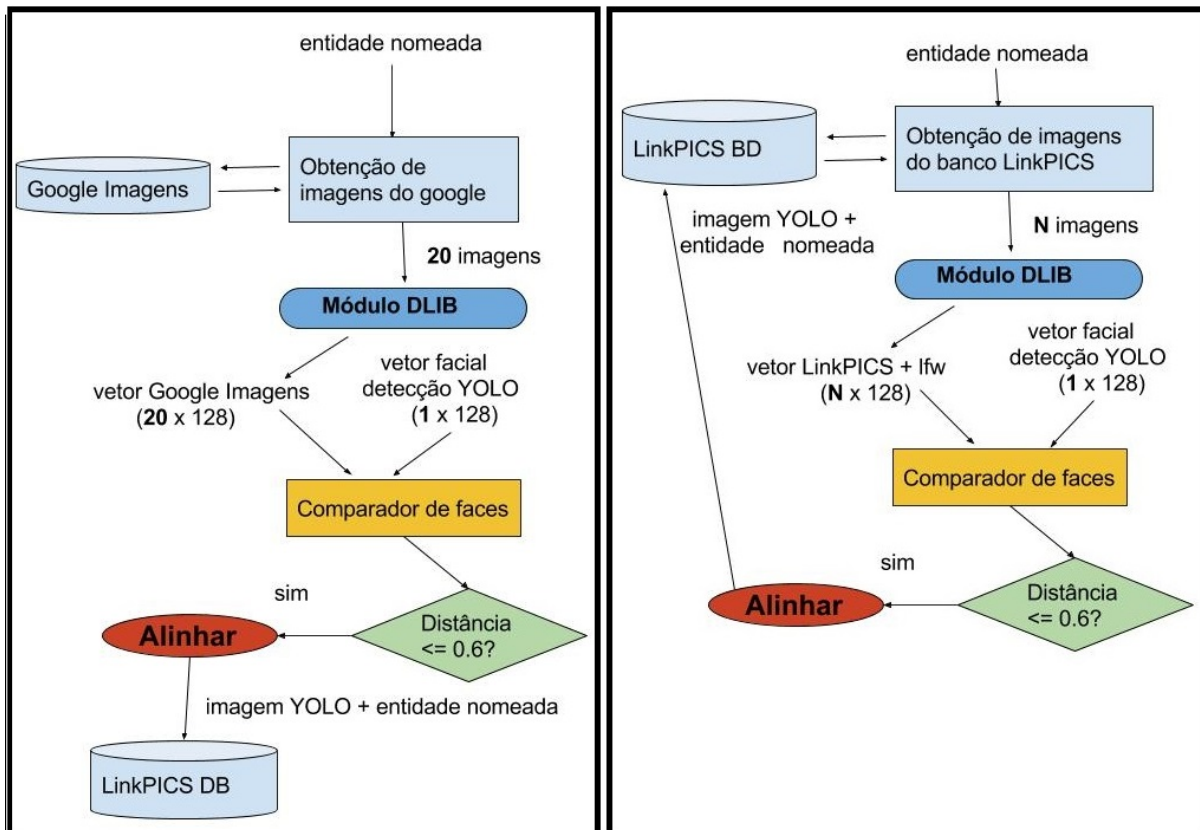


O banco LinkPICS é um recurso desenvolvido neste projeto. Trata-se de um conjunto de faces e nomes associados construído a partir das imagens do conjunto LFW (HUANG et al., 2007) e constantemente incrementado com as imagens alinhadas pelo LinkPICS, conforme ilustrado na Figura 4.10. O banco de imagens do Google, por sua vez, é composto pelas imagens melhor ranqueadas em uma busca feita no Google imagens⁸, como descrito a seguir.

A partir das entidades nomeadas encontradas no processamento do texto (veja seção 4.1.3), para cada entidade, verifica-se se a mesma está presente no banco do LinkPICS. Em caso afir-

⁸As imagens são retiradas do site <https://www.google.com.br/> na aba imagens.

Figura 4.10: Estrutura do alinhador de pessoas utilizando o banco de imagens do Google (à esquerda) e o banco LinkPICS (à direita).



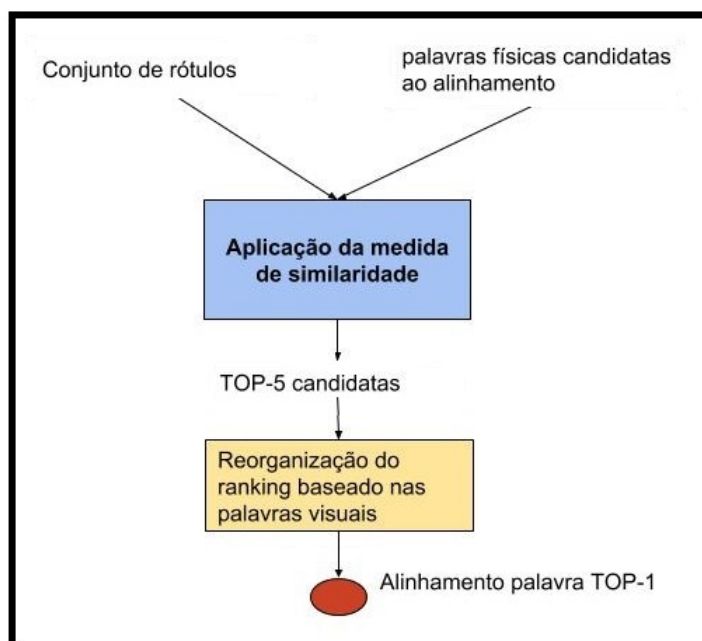
mativo, todas as N imagens faciais presentes no banco LinkPICS referentes à entidade nomeada são transformadas pelo módulo DLIB em um vetor facial (*vetorLinkPICS*) de N posições, como ilustrado na Figura 4.10 (à direita). As características de cada posição do *vetorLinkPICS* são comparadas com as características faciais de cada pessoa encontrada na imagem (veja seção 4.1.2.2) utilizando a distância euclidiana. Se a distância entre as faces for menor ou igual a 0,6, considera-se que a face contida no banco LinkPICS corresponde à mesma pessoa encontrada na imagem, resultando no alinhamento entre a entidade nomeada e a face. A cada alinhamento, o rosto da pessoa detectada na imagem é salvo no banco de faces, com o respectivo nome associado.

Se, por outro lado, a entidade nomeada fornecida como entrada para o alinhador de pessoas não estiver presente no banco LinkPICS, é feita uma busca no site Google imagens utilizando a entidade nomeada como chave para a pesquisa. Como ilustrado na Figura 4.10 (à esquerda), as 20 primeiras imagens retornadas nesta pesquisa são transformadas em vetores faciais (*vetorGoogleImagens*). Esses vetores faciais são comparados com as características faciais de cada pessoa encontrada na imagem, seguindo os mesmos procedimentos do alinhamento citado acima. Novamente, as faces alinhadas são armazenadas no banco LinkPICS.

4.1.5 Alinhador de objetos

Esse alinhador utiliza os rótulos de objetos gerados no processamento de imagem da notícia (veja seção 4.1.2.3) e as palavras físicas candidatas ao alinhamento obtidas no processamento de texto (veja seção 4.1.3). A Figura ?? ilustra os passos do alinhamento de objetos.

Figura 4.11: Alinhador de objetos utilizando os rótulos obtidos do processamento da imagem e as palavras do processamento do texto.



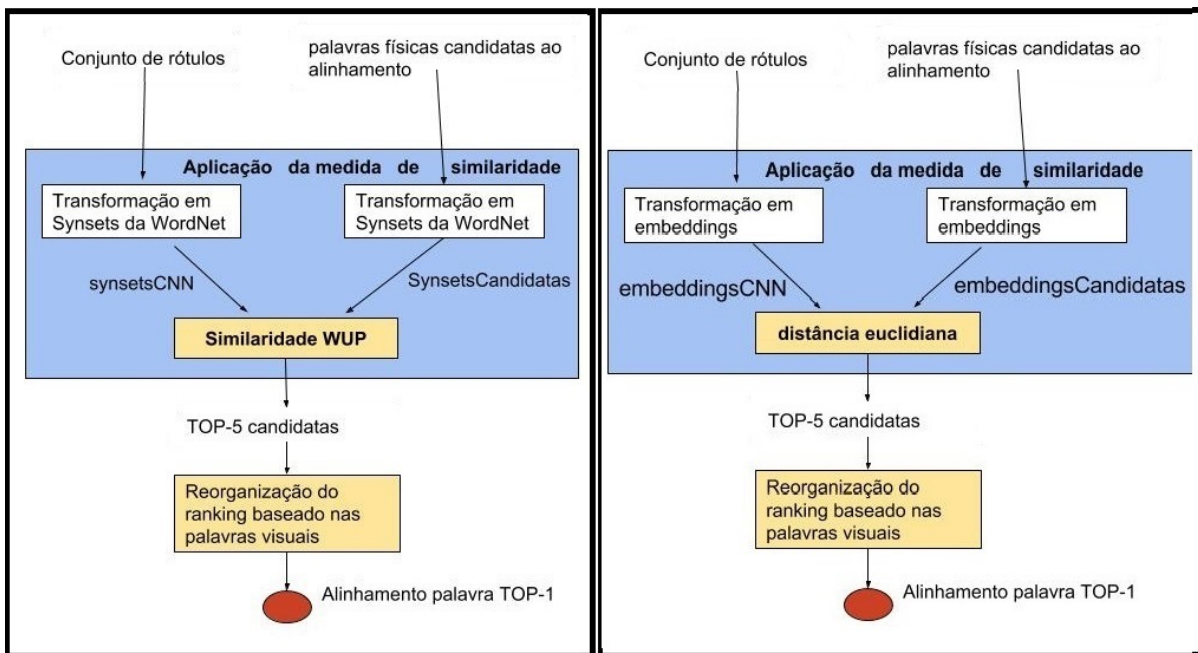
Cada palavra candidata é comparada com cada rótulo por meio de uma medida de similaridade. As duas medidas de similaridade implementadas no LinkPICKS são: a similaridade WUP (veja seção 2.1.5.1) e a similaridade baseada em *word embeddings* (WE) (veja seção 2.1.5.2).

A similaridade WUP é calculada com base na estrutura da WordNet. Como cada palavra na WordNet pode estar associada a vários *synsets*, não é possível determinar inequivocamente qual o *synset* correspondente à palavra do texto. Portanto, nessa estratégia, é necessário calcular a proximidade entre todos os *synsets* possíveis para a palavra do texto com o conjunto de rótulos possíveis para a *bounding box* do objeto.

A estratégia de alinhamento utilizando a similaridade WUP pode ser vista na Figura 4.12. Primeiramente, os conjuntos de rótulos são transformados em *synsets* utilizando a WordNet (*SynsetsCNN*). Em seguida, as palavras candidatas ao alinhamento também são transformadas em *synsets* (*SynsetsCandidatas*). Na sequência, é calculada a similaridade WUP entre os elementos presentes no *SynsetsCandidatas* e os elementos do *SynsetsCNN*. As cinco palavras candidatas com maior similaridade compõem o TOP-5 de palavras candidatas ao alinhamento.

Cada palavra na lista de TOP-5 é buscada na lista de palavras visuais (descrita na seção 5.1.3) e caso esteja presente na lista, essa palavra sobe de posição no TOP-5. Essa estratégia favorece o alinhamento de palavras que podem representar melhor um objeto, aumentando a probabilidade de sucesso no alinhamento. Por fim, na estratégia de alinhamento de objetos adotada no momento, realiza-se o alinhamento da palavra melhor ranqueada (TOP-1) com a *bounding box* do objeto.

Figura 4.12: Estrutura do alinhador de objetos utilizando a similaridade WUP (à esquerda) e as *word embeddings* do GloVe (à direita).

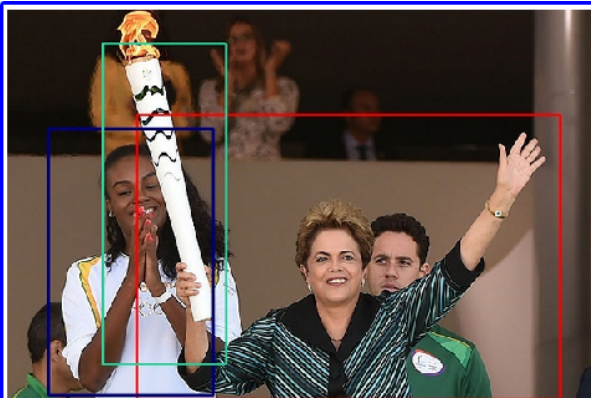


Na estratégia de alinhamento utilizando a distância euclidiana calculada a partir de WEs, o processo é semelhante, a grande diferença destacada em azul na Figura 4.12) é que os conjuntos de rótulos e as palavras candidatas são transformados em *embeddings* (*embeddingCNN* e *embeddingCandidatas*, respectivamente) e na sequência, é calculada a distância euclidiana entre as *embeddings*. As cinco palavras candidatas com a menor distância compõem o TOP-5 de palavras candidatas ao alinhamento e a definição da palavra melhor ranqueada (TOP-1) segue a mesma estratégia utilizada na similaridade WUP.

4.2 Saída do Alinhador

A Figura 4.13 ilustra a saída do alinhador LinkPICS. As palavras alinhadas da notícia são destacadas da mesma cor que a *bounding box* correspondente na imagem.

Figura 4.13: Saída do alinhador LinkPICS para a notícia de entrada apresentada na Figura 4.2. As palavras alinhadas da notícia são destacadas na mesma cor que a *bounding box* correspondente na imagem.

	<p>Título</p> <p>Olympic Games Will Be a Major Success Says Rousseff</p> <p>Legenda</p> <p>Brazilian volleyball player Fabiana Claudino (L) applauds as Brazilian President Dilma Rousseff holds the Olympic torch</p> <p>Texto</p> <p>Two-time Olympic volleyball champion Fabiana Claudino runs holding the torch, and around her, in the landscape framed by the Planalto Palace, on the left side, and the National Congress, on the right, protesters follow the athlete screaming. They display banners in favor of President Dilma Rousseff and against her. Thus the start of the Olympic torch relay on Tuesday (3), in Brasilia, brought side by</p>
--	--

Capítulo 5

EXPERIMENTOS E AVALIAÇÃO

Neste capítulo são apresentados: (seção 5.1) os recursos utilizados nos experimentos, (seção 5.2) o alinhador texto-imagem desenvolvido para servir de *baseline* de comparação com o LinkPICS, (seção 5.3) a plataforma criada para a avaliação dos alinhamentos e os resultados dos experimentos com o alinhador de pessoas (seção 5.4) e objetos (seção 5.5).

5.1 Recursos

Nos experimentos descritos neste documento foram utilizados dois *córpus*, descritos na subseção 5.1.1. Além desses recursos indispensáveis para a avaliação do LinkPICS, outros dois foram utilizados pelo alinhador de objetos: (seção 5.1.2) a WordNet, utilizada para realizar o filtro de palavras físicas e para o cálculo de similaridade lexical usando a WUP, e (seção 5.1.3) a lista de palavras visuais utilizada para o ranqueamento de palavras candidatas ao alinhamento da lista TOP-5.

5.1.1 *Córpus*

Para a realização dos experimentos descritos neste trabalho, dois *córpus* de notícias foram utilizados. O primeiro deles é o *córpus* composto por notícias retiradas do jornal Folha de São Paulo Internacional¹. Nesse *córpus*, os textos estão escritos em inglês e geralmente possuem uma imagem associada. Cada texto tem um *link* que permite visualizar a página com a notícia original, escrita em português. Uma característica desse *córpus* é que a maioria das notícias contém pessoas. A Figura 4.1.1, apresentada no capítulo anterior, é um exemplo de notícia pertencente a esse *córpus*.

¹Disponível em: <http://www1.folha.uol.com.br/internacional/en/>. Acesso em: 29 dez. 2017.

Para a coleta dos artigos de notícias para formar esse *córpus* foi desenvolvido um *crawler* na linguagem *Python* que coleta diariamente as notícias do jornal. Para os experimentos apresentados neste documento, foram coletados 256 notícias (pares de texto-imagem).

Outro *córpus* utilizado neste trabalho é o composto por notícias do jornal BBC NEWS². Esse *córpus* também está no idioma inglês e contém muitas notícias relacionadas a objetos ou outras categorias (animais, etc.) diferentes de pessoas. No total foram coletadas 87 notícias (pares de texto-imagem). Um exemplo de notícia presente no *córpus* BBC é ilustrado na Figura 5.1.

Figura 5.1: Exemplo de notícia extraída do *córpus* BBC contendo: (1) título da notícia, (2) texto da notícia, (3) imagem associada ao texto e (4) legenda da imagem.



5.1.2 WordNet

Outro recurso utilizado nos experimentos descritos neste documento é a WordNet. A WordNet é um banco de dados léxico do inglês. As características principais da WordNet já foram mencionadas na seção 2.1.4. A WordNet está disponível em versão *online* e para *download* no site: <https://wordnet.princeton.edu/>.

Neste trabalho, a WordNet foi utilizada para filtrar palavras físicas (seção 4.1.3) e para o cálculo da similaridade WUP (seção 2.1.5.1), ambos durante o alinhamento de objetos (seção

²Disponível em: <http://www.bbc.com/news>. Acesso em: 29 dez. 2017.

4.1.5).

5.1.3 Palavras Visuais

As “palavras visuais” são substantivos que podem ser representados em uma imagem (BOIY; DESCHACHT; MOENS, 2008). Na seção 4.1.3, foi apresentada uma solução utilizando a WordNet para filtrar esses substantivos em um texto, resultando nas “palavras físicas”. Entretanto, há muitas palavras que mesmo filtradas não podem representar um objeto em uma imagem como “*country*” (país), “*wife*” (esposa) e “*young*” (jovem).

Como tentativa de melhorar a precisão do alinhamento de objetos, neste trabalho foi criada uma lista com 1056 palavras visuais. Essa lista foi criada manualmente a partir da coleta de palavras contidas na página *web 3.000 Core Vocabulary Words*³ que contém palavras pertencentes ao vocabulário inglês. Para a escolha das palavras visuais, foram consideradas as classes de vocabulário que podem ser representadas em uma imagem como: animais, veículos, objetos da casa, eletrodomésticos, etc. Assim, classes como: membros da família, profissões e clima foram descartadas.

A lista de palavras visuais final gerada neste trabalho contém 1.056 palavras e um pequeno trecho dela pode ser observado na Figura 5.2.

Figura 5.2: Trecho da lista de palavra de visuais gerada neste trabalho.

960	snowplow	972	hovercraft	984	bus	1004	tanker
961	sports car	973	icebreaker	985	car	1005	C-clamp
962	tank	974	kayak	986	dogsled	1006	chisel
963	taxi	975	motorboat	987	horse	1007	electric drill
964	taxicab	976	outrigger	988	motorcycle	1008	file
965	tractor	977	rafts	989	scooter	1009	sander
966	wrecker	978	rowboat	990	streetcar	1010	hacksaw
967	barge	979	sampan	991	subway	1011	jigsaw
968	dredger	980	submarine	992	taxi	1012	large level
969	dugout	981	tanker	993	train	1013	large screwdriver
970	galley	982	tugboat	994	truck	1014	snips
971	gondola	983	bicycle	995	airbus	1015	brush

5.2 Alinhador baseline

Para servir de base de comparação com o alinhador LinkPICS proposto neste trabalho, criou-se um alinhador *baseline* que não faz uso de muitos recursos. O *baseline* segue a mesma

³Disponível em: <http://learnersdictionary.com/3000-words>

arquitetura definida para o LinkPICS e apresentada na Figura 4.1. Contudo, a regra de alinhamento estabelecida para o *baseline* é alinhar as regiões mais importantes da imagem com as palavras mais relevantes do texto. No caso deste trabalho, a definição da importância de uma região da imagem (*bouding box*) e de uma palavra ou entidade foi feita com base nos seguintes critérios:

- **Região da imagem contendo pessoa:** A importância das regiões da imagem contendo pessoas está diretamente relacionada a sua posição na imagem: quanto mais próxima do centro inferior, mais importante é a região da imagem, ou seja, mais importante é a pessoa nesta região da imagem. Esse critério foi adotado após a observação de que as personagens mais importantes de uma notícia contendo pessoas são, frequentemente, colocadas no centro da imagem e próximas ao centro inferior da imagem, ganhando destaque.
- **Região da imagem contendo objeto:** A importância das regiões da imagem contendo objetos, por sua vez, está relacionada ao tamanho do objeto detectado na imagem da notícia: quanto maior o objeto, maior é sua importância. Esse critério também foi definido empiricamente após a análise das notícias nos corpúscos.
- **Palavras e entidades:** As palavras e entidades são classificadas por relevância de acordo com os seguintes critérios:
 1. Presença da palavra / entidade na legenda da imagem e no título da notícia.
 2. Presença da palavra / entidade na legenda da imagem.
 3. Presença da palavra / entidade no título da notícia.
 4. Frequência da palavra / entidade.

Assim, as palavras mais importantes são aquelas que estão presentes na legenda e no título da notícia simultaneamente. Em seguida, vêm as palavras presentes na legenda da imagem e, posteriormente, as palavras que aparecem no título da notícia. Caso haja mais de uma palavra que atenda alguns desses critérios, o desempate é realizado considerando-se a frequência, ou seja, palavras mais frequentes são consideradas mais importantes. Por fim, as palavras que não aparecem na legenda e no título da notícia são ordenadas de acordo com a frequência da palavra no texto.

Assim, após serem geradas as listas de regiões da imagem e palavras/entidades ordenadas decrescentemente por seu grau de importância/relevância, seguindo uma abordagem similar a (TEGEN et al., 2014),

o alinhador *baseline* realiza o mapeamento um-para-um na ordem dessas listas. Mais especificamente, o alinhador de pessoas *baseline* alinha as regiões da imagem mais importantes contendo pessoas com as entidades nomeadas mais relevantes. De modo similar, o alinhador de objetos *baseline* alinha as regiões da imagem mais importantes contendo objetos com as palavras físicas mais relevantes.

5.3 Plataforma de Avaliação

Como forma de avaliar os alinhamentos do *baseline* e do LinkPICS, foi criada uma plataforma *web* que pode realizar o alinhamento do texto e da imagem presentes em uma notícia. A Figura 5.3 ilustra a plataforma construída para a avaliação dos alinhadores *baseline* e LinkPICS.

Figura 5.3: Interface web da plataforma construída para a avaliação dos alinhadores *baseline* e LinkPICS.

Alinhamento de notícia

Link da página

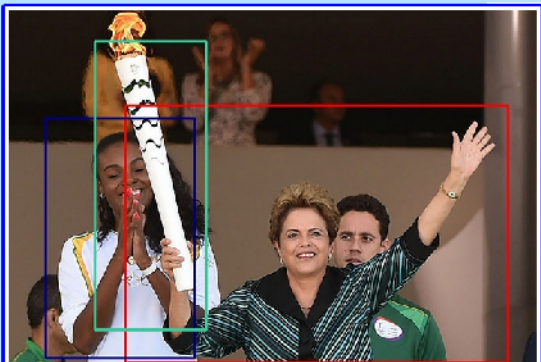
Escolha o arquivo

Browse... brazil_filtrada.txt

Carregar Links

Alinhamento texto-imagem

Similaridade: WUP WE



Dados da notícia

Título

Olympic Games Will Be a Major Success Says **Rousseff**

Legenda

Brazilian volleyball player **Fabiana Claudino** (L) applauds as Brazilian President **Dilma Rousseff** holds the Olympic **torch**

Texto

Two-time Olympic volleyball champion **Fabiana Claudino** runs holding the torch, and around her, in the landscape framed by the Planalto Palace, on the left side, and the National Congress, on the right, protesters follow the athlete screaming. They display banners in favor of President **Dilma Rousseff** and against her. Thus the start of the Olympic **torch** relay on Tuesday (3), in Brasilia, brought side by

Bounding Box	Avaliação - Alinhamento
Dilma Rousseff	<input type="radio"/> Sim <input type="radio"/> Não
Fabiana Claudino	<input type="radio"/> Sim <input type="radio"/> Não
torch	<input type="radio"/> Sim <input type="radio"/> Não

Finalizar Avaliação

No canto superior esquerdo é possível selecionar o endereço *web* da notícia e escolher qual medida de similaridade será aplicada no alinhamento de objetos. No canto inferior esquerdo,

os dados da notícia (título, legenda, texto) são exibidos e cada palavra/entidade alinhada é destacada com a mesma cor de sua respectiva região da imagem (*bounding box*), presente no canto superior direito. No canto inferior direito, todos os alinhamentos são exibidos com a opção de serem marcados como corretos (Sim) ou incorretos (Não). A avaliação é finalizada pressionando o botão “Finalizar Avaliação” que armazena as avaliações em um banco de dados.

Todos os experimentos foram avaliados nessa plataforma, por um anotador humano conforme descrito nas próximas seções.

5.4 Avaliação do alinhamento de pessoas

Para a avaliação do alinhamento de pessoas, utilizou-se um conjunto de 70 notícias do corpus Folha Internacional contendo pessoas. Nessas notícias, há 73 pessoas possíveis de serem alinhadas. As outras pessoas detectadas foram descartadas da avaliação uma vez que não são mencionadas no texto que acompanha a notícia.

Como medidas de avaliação para o experimento foram utilizadas a precisão, a cobertura e a medida-F (veja seção 2.3). Como mencionado anteriormente, a precisão foi calculada como a razão entre a quantidade de faces alinhadas corretamente e a quantidade de faces alinhadas correta e incorretamente, enquanto a cobertura é a razão entre as faces alinhadas corretamente e o total de faces possíveis de serem alinhadas. A medida-F é a média harmônica entre a precisão e a cobertura. O resultado da avaliação do alinhador *baseline* e do LinkPICS, para pessoas, podem ser vistos na Tabela 5.1

Tabela 5.1: Comparação dos resultados para o alinhamento de pessoas utilizando o alinhador *baseline* e o LinkPICS.

Medidas	<i>Baseline</i>	LinkPICS
Precisão	69,57%	98,41%
Cobertura	65,75%	86,30%
Medida-F	67,61%	91,96%

O alinhador LinkPICS alinhou 63 pessoas e 62 estavam corretas, enquanto o alinhador *baseline* alinhou 69 pessoas e 48 estavam corretas. O alinhador *baseline* teve um bom desempenho quando havia apenas uma pessoa na imagem, entretanto, o alinhador não conseguiu alinhar corretamente duas pessoas ou mais. O LinkPICS não possui essa desvantagem e conseguiu ser efetivo em todos os possíveis casos. A Tabela 5.2 traz a média de entidades nomeadas e regiões contendo pessoas para o corpus de notícias Folha Internacional. Os dados da tabela demonstram que há ocorrências de notícias que contém mais de uma pessoa a ser alinhada e notícias que

Tabela 5.2: Média de entidades nomeadas e regiões contendo pessoas para o corpúsculo de notícias Folha Internacional.

Média de pessoas	Média de entidades nomeadas	Total de notícias
1.83	1.80	70

contém um número entidades nomeadas maior do que o número de pessoas a serem alinhadas.

Vale ressaltar que, para este trabalho, uma maior precisão é mais importante do que uma maior cobertura, uma vez que visa-se enriquecer a notícia estabelecendo uma correspondência correta entre as entidades presentes no texto e as pessoas presentes na imagem. Uma alta precisão significa alinhar corretamente as faces, evitando que o leitor da notícia aprenda algo errado em relação a uma pessoa mencionada no texto.

5.5 Avaliação do alinhamento de objetos

A avaliação do alinhamento de objetos também foi realizada para o alinhador *baseline* e o LinkPICS, contudo, para este último, duas estratégias foram investigadas: uma baseada na medida de similaridade WUP e outra baseada na distância euclidiana calculada para *word embeddings* (WE), conforme descrito na seção 4.1.5.

Inicialmente tentou-se avaliar o alinhamento de objetos presentes nas notícias do corpúsculo Folha de São Paulo. Entretanto, constatou-se que a cobertura dos objetos nesse corpúsculo era de apenas 26% significando que poucos objetos poderiam ser alinhados (total de 16 objetos). Para ilustrar esse problema, a Figura 5.4 contém exemplos de objetos que nunca poderiam ser alinhados por não serem mencionados na notícia.

Devido à baixa cobertura do corpúsculo da Folha, optou-se por utilizar o corpúsculo BBC para a avaliação do alinhamento de objetos. Esse corpúsculo contém 94 objetos que podem ser alinhados por serem mencionados na notícia.

Os resultados da avaliação dos alinhadores podem ser vistos na Tabela 5.3. As medidas cobertura e Medida-F não foram utilizadas pois o número de objetos possíveis de alinhar é exatamente igual ao número de objetos alinhados.

As similaridades WUP e WE utilizam valores que indicam a proximidade entre as palavras. Com base nesses valores, tentou-se determinar um limiar que pudesse ser usado para garantir uma boa precisão dos alinhamentos. De forma empírica foi escolhido o valor 0,82 para o limiar da similaridade WUP e o valor 0,50 para o limiar da similaridade WE. A Tabela 5.4 traz a precisão alcançada considerando-se esses limiares de similaridade para realizar o alinhamento. No

Figura 5.4: Exemplos de objetos que nunca poderiam ser alinhados por não serem mencionados nas notícias presente no corpúsculo da Folha de São Paulo Internacional. Em (a) temos uma cadeira e uma mochila; em (b), o logotipo da empresa; em (c), uma taça de vidro e, em (d), trecho de um documento.



Tabela 5.3: Resultado da avaliação do alinhamento de objetos do alinhador *baseline* e do LinkPICS utilizando WUP e distância euclidiana calculada para *word embeddings* (WE).

Estratégia	Alinhamentos corretos	Possíveis de alinhar	Precisão
<i>Baseline</i>	31	94	32,97%
LinkPICS-WUP	59	94	62,77%
LinkPICS-WE	52	94	55,32%

caso da similaridade WUP, quanto maior o valor, mais similares são as palavras em comparação. Assim, apenas os pares com similaridade maior do que 0,82 foram alinhados. A similaridade WE, por outro lado, especifica que quanto menor o valor, mais próximos (similares) são os elementos em comparação, por isso apenas os pares com valores de similaridade WE menores do que 0,50 foram alinhados.

No cálculo da precisão com base nesses limiares de similaridade para o alinhamento, considerou-se apenas os objetos que satisfazem o limiar e, por este motivo, o número de objetos possíveis de alinhar caiu de 94 para 78 na WUP e de 94 para 32 na WE. A precisão usando os limiares, conforme apresentado na Tabela 5.4, subiu de 62,77% para 71,79% na WUP e de

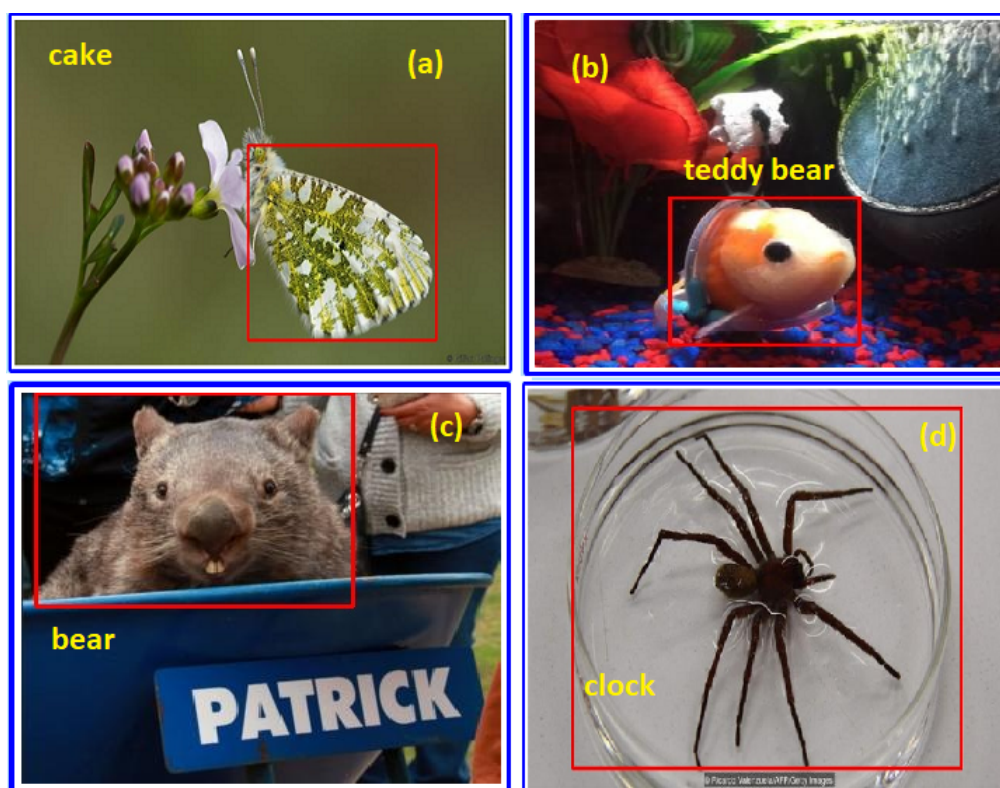
Tabela 5.4: Resultado da avaliação do alinhamento de objetos usando o alinhador LinkPICS com WUP e WE com limiares para o alinhamento.

Estratégia	Alinhamentos corretos	Possíveis de alinhar	Precisão
LinkPICS-WUP (>0,82)	56	78	71,79%
LinkPICS-WE(<0,5)	22	32	68,75%

55,32% para 68,75% na WE. Como pode-se perceber, a utilização dos limiares trouxe um aumento na precisão dos alinhamentos, apesar da cobertura ter diminuído. Contudo, vale ressaltar que a perda na cobertura foi de apenas 17% na WUP, mas de 66% na WE.

Outra conclusão a respeito do alinhamento de objetos é a de que a estratégia de utilizar as redes CNNs (seção 4.1.2.3) em conjunto com a YOLO mostrou-se muito efetiva. Para ilustrar essa constatação, a Figura 5.5 apresenta quatro exemplos de alinhamentos de sucesso graças a essa estratégia. Em todas essas imagens, há animais que a YOLO não consegue rotular corretamente pelo fato de não terem sido treinados na rede.

Figura 5.5: Exemplos de alinhamentos beneficiados pela estratégia de utilização das CNNs em conjunto com a YOLO. Esses objetos foram alinhados corretamente porque o rótulo atribuído pela YOLO (especificado em amarelo em cada imagem) foi expandido pelas outras redes.



Por exemplo, se utilizássemos somente a YOLO, o objeto da imagem (a) seria rotulado como “*cake*”(bolo). Isso provavelmente acontece porque a asa da borboleta possui características semelhantes a alguns exemplos de imagens de “*cake*” utilizadas no treinamento da

rede. Caso a medida de similaridade entre a palavra “*butterfly*” (borboleta) presente no texto e a palavra “*cake*” fosse aplicada, teríamos uma baixa similaridade entre as palavras por se tratarem de objetos totalmente diferentes. A aplicação das redes CNNs somente na região da imagem detectada para esse objeto identificou a presença de um rótulo “*butterfly*” entre os rótulos da saída das redes. Então, aplicando a medida de similaridade entre os rótulos, a palavra “*butterfly*” foi considerada a palavra mais similar.

Por último, foi realizada uma avaliação referente à utilização da lista de palavras visuais. A Tabela 5.5 ilustra o resultado dessa avaliação. Os resultados mostram que o uso das palavras visuais trouxe um aumento na precisão, eliminando as palavras bem ranqueadas que nunca poderiam representar um objeto na imagem.

Tabela 5.5: Precisão do alinhador LinkPICS utilizando as palavras visuais e a precisão do alinhador sem o uso das palavras visuais.

Estratégia	Precisão usando as palavras visuais	Precisão sem as palavras visuais
LinkPICS-WUP	62,77%	45,74%
LinkPICS-WE	55,32%	46,80%

Capítulo 6

CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Este trabalho abordou o alinhamento texto-imagem. O alinhamento texto-imagem é aplicado em textos que possuam uma imagem associada e consiste em deixar clara a correspondência entre elementos representados em uma imagem (pessoas, objetos, animais, etc.) e as palavras ou entidades nomeadas presentes no texto e que referenciam esses elementos.

O principal foco do alinhamento texto-imagem investigado neste trabalho são as notícias de jornal *online*. Muitas notícias não deixam clara para o leitor a correspondência entre elementos do texto e elementos contidos na imagem associada. O alinhamento texto-imagem surge com a intenção de orientar o leitor, trazendo clareza para a notícia e a imagem associada.

Para tanto, neste projeto, técnicas de Processamento de Língua Natural (PLN) e Visão Computacional (VC) foram combinadas no desenvolvimento do alinhador LinkPICS. O LinkPICS realiza o alinhamento texto-imagem por meio de dois processos distintos: (1) o alinhamento de pessoas, no qual regiões da imagem contendo pessoas são alinhadas com as entidades nomeadas identificadas no texto, e (2) o alinhamento de objetos, no qual regiões da imagem contendo objetos, animais, etc. são alinhadas com palavras físicas presentes no texto.

Com relação às técnicas de VC, o LinkPICS utiliza a rede convolucional (CNN) YOLO (REDMON et al., 2016) para detectar pessoas e objetos na imagem associada ao texto da notícia. No alinhamento de objetos, devido ao pequeno número de classes de objetos que a YOLO reconhece (apenas 80), foram utilizadas, também, outras três CNNs para a geração de outros possíveis rótulos para as regiões de imagem contendo objetos. As três CNNs são aplicadas em cada região da imagem na qual um objeto foi detectado e 5 novos rótulos são gerados por cada rede, totalizando 16 rótulos (15 rótulos da CNN + 1 da YOLO).

Com relação às técnicas de PLN, o LinkPICS realiza a etiquetagem morfosintática com o TreeTagger (SCHMID, 2013), importante para o alinhamento de objetos, e a identificação de en-

tidades nomeadas com o *Stanford Named Entity Recognizer*¹, indispensável para o alinhamento de pessoas. O LinkPICS também utiliza a WordNet para filtrar as palavras do texto mantendo apenas as palavras físicas (palavras que podem representar um objeto), ou seja, removendo as palavras abstratas. Além disso, para auxiliar o alinhamento de objetos foi criada manualmente uma lista de palavras visuais que contém palavras pertencentes ao vocabulário inglês que podem ser representadas em uma imagem, como por exemplo, animais, veículos, objetos da casa, eletrodomésticos, etc.

O alinhador de pessoas utiliza técnicas de reconhecimento facial que são aplicadas nas regiões da imagem contendo pessoas detectadas pela YOLO, e também em imagens provenientes do site *Google* imagens e do banco de faces LinkPICS. O banco LinkPICS foi construído a partir da combinação de imagens do conjunto de faces *LFW-labeled faces in the wild* e das faces alinhadas pelo LinkPICS. Isso significa que toda face alinhada é armazenada no banco LinkPICS.

O alinhador de objetos utiliza medidas de similaridade que calculam a distância entre as palavras candidatas ao alinhamento (palavras físicas filtradas como auxílio da WordNet) e os rótulos atribuídos às regiões da imagem contendo o objeto, e gera uma lista com as cinco palavras candidatas com maior similaridade (TOP-5). Em seguida as palavras candidatas contidas na lista de palavras visuais ganham preferência em relação às palavras da lista TOP-5 que não estão na lista de palavras visuais. Por fim, a palavra na primeira posição da lista de TOP-5 (a TOP-1) é alinhada com o objeto.

A avaliação do alinhador de pessoas foi realizada no corpúsculo de notícias Folha de São Paulo Internacional e obteve uma precisão de 98%. A avaliação do alinhador de objetos foi realizada no corpúsculo de notícias BBC News e obteve uma precisão de 72% usando similaridade WUP com limiar mínimo de 0,82.

Além desses valores objetivos, outras constatações subjetivas nesse processo de avaliação são: (1) a estratégia de utilizar as redes CNNs em conjunto com a YOLO no alinhamento de objetos mostrou-se muito efetiva e (2) a lista de palavras visuais que aumentou a precisão ao desconsiderar como candidatas ao alinhamento aquelas palavras que não podem ser representadas em uma imagem.

A seguir, são listadas algumas propostas de trabalhos futuros decorrentes deste (seção 6.1), bem como as principais contribuições deste trabalho (seção 6.2).

¹Disponível em: <https://nlp.stanford.edu/software/CRF-NER.shtml>. Acesso em: 08 jan. 2018.

6.1 Trabalhos Futuros

Nesta seção são detalhados os principais trabalhos decorrentes deste relacionados: (6.1.1) às principais melhorias previstas para o LinkPICS, (6.1.2) à criação de um dicionário visual e (6.1.3) à criação de uma ferramenta não supervisionada para coleta de imagens referentes a objetos.

6.1.1 Melhorias para o alinhador LinkPICS

Entre as melhorias que poderão ser implementadas no LinkPICS, destacam-se as listadas a seguir:

- **Alinhamento de expressões multipalavras** – A versão atual do alinhador de objetos do LinkPICS alinha apenas palavras simples, uma vez que o processo envolve a identificação de substantivos e cada substantivo pode ser alinhado com uma região da imagem. Contudo, há ocorrências de palavras compostas por mais de um substantivo (por exemplo, “*panda bear*”) e adjetivos associados aos substantivos (por exemplo, “*blue car*” e “*red hat*”). Assim, como melhoria futura está prevista a incorporação de um processo automático de identificação de expressões multipalavras de tal maneira que essas palavras sejam tratadas como uma única entidade e, desse modo, seja possível realizar o alinhamento envolvendo multipalavras.
- **Alinhamento n:1** – na versão atual do LinkPICS são gerados apenas alinhamentos 1:1, ou seja, uma palavra ou entidade nomeada no texto está associada a uma região da imagem e vice-versa. Como extensão futura, pretende-se investigar a possibilidade de alinhamentos n:1, nos quais várias palavras ou entidades nomeadas podem ser associadas a uma mesma região da imagem. Isso ocorre, por exemplo, em casos de sinônimos no texto (como “*plane*” e “*airplane*”) e anáforas.² Por exemplo, considere o trecho de texto da Figura 6.1³ e todas as entidades relacionadas destacadas (em amarelo). A resolução de anáfora pode auxiliar na interpretação do texto e trazer clareza ao leitor, alinhando todas as menções de uma pessoa à sua respectiva imagem.
- **Geração de uma interface WEB** – Outra melhoria prevista para o LinkPICS é a adaptação da plataforma de avaliação (veja seção 5.3) para disponibilização do LinkPICS para o

²Anáforas são termos (palavras) de um texto que referenciam a mesma entidade.

³Texto original disponível em: <http://www1.folha.uol.com.br/internacional/en/brazil/2016/07/1792819-lula-leads-2018-election-voter-polls-but-second-round-victory-would-prove-difficult.shtml>

Figura 6.1: Texto original e texto após a resolução de anáfora.

Texto Original	Former president Luiz Inácio Lula da Silva (PT) leads the Datafolha voter polls in the first round of elections for the 2018 presidential campaign. The PT member's lead does not, however, guarantee a second round victory and he could be defeated by former Senator Marina Silva (Rede) or Minister of Foreign Relations, José Serra (PSDB).
Resolução de anáfora	Former president Luiz Inácio Lula da Silva (PT) leads the Datafolha voter polls in the first round of elections for the 2018 presidential campaign. The PT member's lead does not, however, guarantee a second round victory and he could be defeated by former Senator Marina Silva (Rede) or Minister of Foreign Relations, José Serra (PSDB).

público geral em forma de *plugin*⁴ para os sites da Folha Internacional e BBC News.

- **Adaptação do LinkPICS para alinhar notícias em português** – Por fim, pretende-se gerar uma versão do LinkPICS capaz de processar e alinhar textos em português. Para tanto, o maior trabalho será relacionado às ferramentas e recursos de PLN, sendo necessário: (1) um etiquetador morfosintático e (2) um reconhecedor de entidades nomeadas para o português, bem como (3) uma WordNet disponível para esse idioma e (4) uma versão da lista de palavras visuais em português. Com relação às técnicas de VC será necessário, apenas, traduzir os rótulos fornecidos pelas redes convolucionais de inglês para o português.

Em termos das estratégias usadas para o alinhamento propriamente dito, também será necessário utilizar uma WordNet em português para cálculo da similaridade WUP ou gerar *word embeddings* em português para o cálculo da distância euclidiana no alinhamento de objetos.

6.1.2 Criação de um Dicionário Visual

Outro resultado do alinhamento texto-imagem é a possibilidade de criação de um dicionário visual. Um dicionário visual é composto de pares de imagem e termos associados conforme ilustra a Figura 6.2.

Acredita-se que esse dicionário visual pode ser útil para diversas aplicações manuais e automáticas. Por exemplo, pensando em uma aplicação manual, o dicionário visual pode auxiliar no aprendizado de outro idioma e pode ser mais útil que um dicionário tradicional contendo apenas palavras. Pensando em uma aplicação automática, a partir dos pares de palavras e imagens seria possível a construção de outra ferramenta na forma de um *plugin*. Essa ferramenta poderia

⁴Os *plugins* são ferramentas disponibilizadas nos navegadores de internet que executam tarefas específicas.

Figura 6.2: Exemplo de um dicionário visual contendo a imagem e termos associados.

verificar quais palavras do texto estão contidas no dicionário visual para, em sequência, enriquecer o texto destacando essas palavras de tal maneira que quando o usuário clicasse em uma palavra destacada, a imagem correspondente a essa palavra fosse exibida na tela. A utilização dessa ferramenta poderia acelerar o aprendizado do usuário a respeito de palavras desconhecidas.

6.1.3 Criação de uma ferramenta para coleta automática de imagens

Por fim, considerando-se o sucesso da estratégia de alinhamento utilizando CNNs, visa-se a criação de uma ferramenta não supervisionada que percorre os sites da internet e coleta imagens dos objetos de interesse informados pelo usuário.

A estratégia da ferramenta seria a de coletar todos os textos e imagens da página e aplicar o LinkPICS. Os objetos alinhados poderiam, então, ser úteis para a criação de conjuntos de dados para treinamento de redes convolucionais ou para incrementar o dicionário visual.

6.2 Contribuições

Essa última seção resume as principais contribuições deste trabalho. São elas:

1. **Alinhador texto-imagem LinkPICS** – com 98% de precisão no alinhamento de pessoas e 72% no alinhamento de objetos, o LinkPICS (descrito no Capítulo 4) se mostrou uma boa ferramenta automática para o alinhamento texto-imagem em sites de notícias;
2. **Córpus Folha de São Paulo Internacional** – contendo 256 pares de texto-imagem coletados do site do jornal Folha de São Paulo Internacional⁵, dos quais 70 foram usados nos experimentos de avaliação do alinhador de pessoas;
3. **Córpus BBC News** – contendo 87 pares de texto-imagem coletados do site do jornal BBC NEWS⁶, todos usados nos experimentos de avaliação do alinhador de objetos;
4. **Lista de palavras visuais em inglês** – contendo 1.056 palavras originalmente extraídas da página *web 3.000 Core Vocabulary Words*⁷ e filtradas manualmente neste projeto para conter apenas aquelas que podem ser representadas em uma imagem, como animais, veículos, objetos da casa, eletrodomésticos, etc.;
5. **Banco LinkPICS** – contendo faces e nomes (entidades nomeadas) associados construído a partir das imagens do conjunto LFW (HUANG et al., 2007) e constantemente incrementado com as imagens alinhadas pelo LinkPICS, conforme ilustrado na Figura 4.10;
6. **Dicionário visual** – conjunto de pares de imagens de objetos e palavras relacionadas gerados automaticamente pelo alinhador de objetos do LinkPICS, conforme descrito na seção 6.1.2.

⁵Disponível em: <http://www1.folha.uol.com.br/internacional/en/>. Acesso em: 29 dez. 2017.

⁶Disponível em: <http://www.bbc.com/news>. Acesso em: 29 dez. 2017.

⁷Disponível em: <http://learnersdictionary.com/3000-words>

REFERÊNCIAS

- BAUM, L. E. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, v. 3, p. 1–8, 1972.
- BAUM, L. E.; EAGON, J. A. et al. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, v. 73, n. 3, p. 360–363, 1967.
- BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, JSTOR, v. 37, n. 6, p. 1554–1563, 1966.
- BAUM, L. E. et al. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, JSTOR, v. 41, n. 1, p. 164–171, 1970.
- BAUM, L. E.; SELL, G. Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, Mathematical Sciences Publishers, v. 27, n. 2, p. 211–227, 1968.
- BERG, T. L. et al. Names and faces in the news. In: IEEE. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. [S.l.], 2004. v. 2, p. II–848.
- BISHOP, C. M. Pattern recognition. *Machine Learning*, v. 128, 2006.
- BLEI, D. M.; JORDAN, M. I. Modeling annotated data. In: ACM. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. [S.l.], 2003. p. 127–134.
- BOIY, E.; DESCHACHT, K.; MOENS, M.-F. Learning visual entities and their visual attributes from text corpora. In: IEEE. *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*. [S.l.], 2008. p. 48–53.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- BRILL, E. Unsupervised learning of disambiguation rules for part of speech tagging. In: SOMERSET, NEW JERSEY: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the third workshop on very large corpora*. [S.l.], 1995. v. 30, p. 1–13.
- CAO, L.; FEI-FEI, L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: IEEE. *Proceedings of the IEEE 11th International Conference on Computer Vision*. [S.l.], 2007. p. 1–8.

- CARBONETTO, P. *Unsupervised statistical models for general object recognition*. Tese (Doutorado) — University of British Columbia, 2003.
- CARREIRA, J.; SMINCHISESCU, C. Constrained parametric min-cuts for automatic object segmentation. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. [S.l.], 2010. p. 3241–3248.
- CHANG, F.; CHEN, C.-J.; LU, C.-J. A linear-time component-labeling algorithm using contour tracing technique. *computer vision and image understanding*, Elsevier, v. 93, n. 2, p. 206–220, 2004.
- CHOI, D. et al. A method for enhancing image retrieval based on annotation using modified wup similarity in wordnet. In: *Proceedings of the 11th WSEAS international conference on artificial intelligence, knowledge engineering and data bases, AIKED*. [S.l.: s.n.], 2012. v. 12, p. 83–87.
- CHOI, D.; KIM, P. Automatic image annotation using semantic text analysis. In: SPRINGER. *International Conference on Availability, Reliability, and Security*. [S.l.], 2012. p. 479–487.
- CHOI, D. et al. Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, Elsevier, v. 38, p. 16–21, 2014.
- CHURCH, K. W. A stochastic parts program and noun phrase parser for unrestricted text. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the second conference on Applied natural language processing*. [S.l.], 1988. p. 136–143.
- COMANICIU, D.; MEER, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 5, p. 603–619, 2002.
- CRISTIANINI, N.; SHAWE-TAYLOR, J. *An introduction to support vector machines and other kernel-based learning methods*. [S.l.]: Cambridge university press, 2000.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, JSTOR, p. 1–38, 1977.
- DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. [S.l.], 2009. p. 248–255.
- DENOYER, L.; GALLINARI, P. The wikipedia xml corpus. *Comparative Evaluation of XML Information Retrieval Systems*, Springer, p. 12–19, 2007.
- DEROSE, S. J. Grammatical category disambiguation by statistical optimization. *Computational linguistics*, MIT Press, v. 14, n. 1, p. 31–39, 1988.
- DESCHACHT, K.; MOENS, M.-F. et al. Text analysis for automatic image annotation. In: ACL. [S.l.: s.n.], 2007. v. 7, p. 1000–1007.
- DUYGULU, P. et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Computer Vision—ECCV 2002*. [S.l.]: Springer, 2002. p. 97–112.

- ESCALANTE, H. J. et al. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, Elsevier, v. 114, n. 4, p. 419–428, 2010.
- EVERINGHAM, M. et al. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, Springer, v. 111, n. 1, p. 98–136, 2015.
- EVERINGHAM, M.; SIVIC, J.; ZISSERMAN, A. Hello! my name is... buffy”–automatic naming of characters in tv video. In: *BMVC*. [S.l.: s.n.], 2006. v. 2, n. 4, p. 6.
- FAN, R.-E. et al. Liblinear: A library for large linear classification. *Journal of machine learning research*, v. 9, n. Aug, p. 1871–1874, 2008.
- FELLBAUM, C. *WordNet: An Electronic Lexical Database*. [S.l.]: MIT Press, 1998.
- FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient graph-based image segmentation. *International Journal of Computer Vision*, Springer, v. 59, n. 2, p. 167–181, 2004.
- FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. [S.l.], 2005. p. 363–370.
- FISCHER, A. et al. Transcription alignment of latin manuscripts using hidden markov models. In: ACM. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*. [S.l.], 2011. p. 29–36.
- FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.
- FOUROUR, N. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In: *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN’02)*. [S.l.: s.n.], 2002. p. 265–274.
- FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In: SPRINGER. *European conference on computational learning theory*. [S.l.], 1995. p. 23–37.
- GAELE, J. V.; VLACHOS, A.; GHAHRAMANI, Z. The infinite hmm for unsupervised pos tagging. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. [S.l.], 2009. p. 678–687.
- GATOS, B.; STAMATOPOULOS, N.; LOULOUUDIS, G. Icdar2009 handwriting segmentation contest. *International Journal on Document Analysis and Recognition (IJ DAR)*, Springer, v. 14, n. 1, p. 25–33, 2011.
- GOLDWATER, S.; GRIFFITHS, T. A fully bayesian approach to unsupervised part-of-speech tagging. In: CITESEER. *Annual meeting-association for computational linguistics*. [S.l.], 2007. v. 45, n. 1, p. 744.
- GOULD, S.; FULTON, R.; KOLLER, D. Decomposing a scene into geometric and semantically consistent regions. In: IEEE. *2009 IEEE 12th international conference on computer vision*. [S.l.], 2009. p. 1–8.

- GRUBINGER, M. et al. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: *International Workshop OntoImage*. [S.l.: s.n.], 2006. v. 5, p. 10.
- HE, K. et al. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- HUANG, G. et al. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.
- HUANG, G. B. et al. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. [S.l.], 2007.
- JACCARD, P. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. [S.l.]: Impr. Corbaz, 1901.
- KARPATHY, A.; FEI-FEI, L. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. p. 3128–3137.
- KENT, A. et al. Machine literature searching viii. operational criteria for designing information retrieval systems. *American documentation*, Wiley Online Library, v. 6, n. 2, p. 93–101, 1955.
- KOEHN, P. et al. Moses: Open Source Toolkit for Statistical Machine Translation. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) – Demonstration Session*. Prague, Czech Republic: [s.n.], 2007.
- KORNFELD, E. M.; MANMATHA, R.; ALLAN, J. Text alignment with handwritten documents. In: IEEE. *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*. [S.l.], 2004. p. 195–209.
- KRISHNAMOORTHY, N. et al. Generating natural-language video descriptions using text-mined knowledge. In: *AAAI*. [S.l.: s.n.], 2013. v. 1, p. 2.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105.
- LAI, A.; HOCKENMAIER, J. Illinois-lh: A denotational and distributional approach to semantics. *Proc. SemEval*, 2014.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, IEEE, v. 86, n. 11, p. 2278–2324, 1998.
- LEYDIER, Y. et al. Learning-free text-image alignment for medieval manuscripts. In: IEEE. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. [S.l.], 2014. p. 363–368.
- LI, J.; WANG, J. Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 25, n. 9, p. 1075–1088, 2003.
- LI, J.; WANG, J. Z. Real-time computerized annotation of pictures. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 30, n. 6, p. 985–1002, 2008.

- LI, L.-J.; SOCHER, R.; FEI-FEI, L. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: IEEE. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. [S.l.], 2009. p. 2036–2043.
- LIENHART, R.; KURANOV, A.; PISAREVSKY, V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: SPRINGER. *Joint Pattern Recognition Symposium*. [S.l.], 2003. p. 297–304.
- LIU, P.-Y.; ZHAO, T.-J.; YU, X.-F. Application-oriented comparison and evaluation of six semantic similarity measures based on wordnet. In: IEEE. *Machine Learning and Cybernetics, 2006 International Conference on*. [S.l.], 2006. p. 2605–2610.
- LORIGO, L. M.; GOVINDARAJU, V. Transcript mapping for handwritten arabic documents. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Electronic Imaging 2007*. [S.l.], 2007. p. 65000W–65000W.
- LOULLOUDIS, G. et al. Text line detection in handwritten documents. *Pattern Recognition*, Elsevier, v. 41, n. 12, p. 3758–3772, 2008.
- LOVINS, J. B. *Development of a stemming algorithm*. [S.l.]: MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.
- MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.
- MALISIEWICZ, T.; EFROS, A. A. Recognition by association via learning per-exemplar distances. In: IEEE. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. [S.l.], 2008. p. 1–8.
- MARSH, E.; PERZANOWSKI, D. Muc-7 evaluation of ie technology: Overview of results. In: *Proceedings of the seventh message understanding conference (MUC-7)*. [S.l.: s.n.], 1998. v. 20.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- NEW, B. Lexique 3: Une nouvelle base de données lexicales. In: *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*. [S.l.: s.n.], 2006.
- NOEL, G. E.; PETERSON, G. L. Context-driven image annotation using imagenet. In: *The Twenty-Sixth International FLAIRS Conference*. [S.l.: s.n.], 2013.
- OCH, F. J.; NEY, H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, v. 29, n. 1, p. 19–51, 2003.
- OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, Springer, v. 42, n. 3, p. 145–175, 2001.
- OPITZ, D.; MACLIN, R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, v. 11, p. 169–198, 1999.

- PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *EMNLP*. [S.l.: s.n.], 2014. v. 14, p. 1532–1543.
- PHAM, P.; MOENS, M.-F.; TUYTELAARS, T. Linking names and faces: Seeing the problem in different ways. In: *Proceedings of the 10th European conference on computer vision: workshop faces in 'real-life' images: detection, alignment, and recognition*. [S.l.: s.n.], 2008. p. 68–81.
- PORTER, M. F. An algorithm for suffix stripping. *Program*, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980.
- RAMISA, A. et al. Breakingnews: Article annotation by image and text processing. *arXiv preprint arXiv:1603.07141*, 2016.
- REDMON, J. *Darknet: Open Source Neural Networks in C*. 2013. <http://pjreddie.com/darknet/>.
- REDMON, J. et al. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 779–788.
- REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2015. p. 91–99.
- ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.
- ROTHFEDER, J.; MANMATHA, R.; RATH, T. M. Aligning transcripts to automatically segmented handwritten manuscripts. In: SPRINGER. *International Workshop on Document Analysis Systems*. [S.l.], 2006. p. 84–95.
- RUBNER, Y.; TOMASI, C.; GUIBAS, L. J. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, Springer, v. 40, n. 2, p. 99–121, 2000.
- RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, Springer, v. 115, n. 3, p. 211–252, 2015.
- SADEGHI, M. A.; FORSYTH, D. 30hz object detection with dpm v5. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2014. p. 65–79.
- SALTON, G.; MCGILL, M. J. Introduction to modern information retrieval. McGraw-Hill, 1983.
- SANDHAUS, E. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, v. 6, n. 12, p. e26752, 2008.
- SANG, E. F. T. K.; MEULDER, F. D. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. [S.l.], 2003. p. 142–147.

- SAUVOLA, J.; PIETIKÄINEN, M. Adaptive document image binarization. *Pattern recognition*, Elsevier, v. 33, n. 2, p. 225–236, 2000.
- SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: ROUTLEDGE. *New methods in language processing*. [S.l.], 2013. p. 154.
- SCHMIDT, D. Tilt 2: Text to image linking tool. 2014.
- SHI, J.; MALIK, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 22, n. 8, p. 888–905, 2000.
- SHI, T.; LIU, Z. Linking glove with word2vec. *arXiv preprint arXiv:1411.5595*, 2014.
- SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- SIVIC, J.; ZISSERMAN, A. Video google: A text retrieval approach to object matching in videos. In: IEEE. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. [S.l.], 2003. p. 1470–1477.
- SMET, M. D.; FRANSENS, R.; GOOL, L. V. A generalized em approach for 3d model based face recognition under occlusions. In: IEEE. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. [S.l.], 2006. v. 2, p. 1423–1430.
- SOCHER, R.; FEI-FEI, L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. [S.l.], 2010. p. 966–973.
- SOCHER, R. et al. Parsing natural scenes and natural language with recursive neural networks. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. [S.l.: s.n.], 2011. p. 129–136.
- STAMATOPOULOS, N.; LOULLOUDIS, G.; GATOS, B. Efficient transcript mapping to ease the creation of document image segmentation ground truth with text-image alignment. In: IEEE. *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. [S.l.], 2010. p. 226–231.
- SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. p. 1–9.
- TEGEN, A. et al. Image segmentation and labeling using free-form semantic annotation. In: *ICPR*. [S.l.: s.n.], 2014. p. 2281–2286.
- TIGHE, J.; LAZEBNIK, S. Superparsing: scalable nonparametric image parsing with superpixels. In: SPRINGER. *European conference on computer vision*. [S.l.], 2010. p. 352–365.
- TIRILLY, P.; CLAVEAU, V.; GROS, P. Détection de logos pour l'annotation d'images de presse. In: *Congrès francophone AFRIF-AFIA de reconnaissance de formes et d'intelligence artificielle, RFIA'10*. [S.l.: s.n.], 2010.
- TIRILLY, P. et al. News image annotation on a large parallel text-image corpus. In: *LREC*. [S.l.: s.n.], 2010.

- TIWARI, P.; KAMDE, P. Automatic image annotation and retrieval using contextual information. 2015.
- TOSELLI, A. H.; ROMERO, V.; VIDAL, E. Viterbi based alignment between text images and their transcripts. In: *Proceedings of the Workshop on Language Technology for Cultural Heritage Data*. [S.l.: s.n.], 2007. p. 9–16.
- TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. [S.l.], 2003. p. 173–180.
- VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: IEEE. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. [S.l.], 2001. v. 1, p. I–511.
- VIOLA, P.; JONES, M. J. Robust real-time face detection. *International journal of computer vision*, Springer, v. 57, n. 2, p. 137–154, 2004.
- WIKIPEDIA. *Stemming — Wikipedia, The Free Encyclopedia*. 2016. [Online; accessed 3-August-2016]. Disponível em: <<https://en.wikipedia.org/w/index.php?title=Stemming&oldid=723150663>>.
- WU, Z.; PALMER, M. Verbs semantics and lexical selection. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. [S.l.], 1994. p. 133–138.
- YAN, J. et al. The fastest deformable part model for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 2497–2504.
- YIN, F.; WANG, Q.-F.; LIU, C.-L. Transcript mapping for handwritten chinese documents by integrating character recognition model and geometric context. *Pattern Recognition*, Elsevier, v. 46, n. 10, p. 2807–2818, 2013.
- ZENNAKI, O.; SEMMAR, N.; BESACIER, L. Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. 2015.
- ZIMMERMANN, M.; BUNKE, H. Automatic segmentation of the iam off-line database for handwritten english text. In: IEEE. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. [S.l.], 2002. v. 4, p. 35–39.
- ZINGER, S.; NERBONNE, J.; SCHOMAKER, L. Text-image alignment for historical handwritten documents. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *IS&T/SPIE Electronic Imaging*. [S.l.], 2009. p. 724703–724703.



Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Wellington Cristiano Veltroni, realizada em 02/03/2018:

Helena de M. Caseli

Profa. Dra. Helena de Medeiros Caseli
UFSCar

Ricardo Cerri

Prof. Dr. Ricardo Cerri
UFSCar

Profa. Dra. Vladia Celia Monteiro Pinheiro
UNIFOR

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Vladia Celia Monteiro Pinheiro e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Helena de M. Caseli