Pontifícia Universidade Católica
do Rio de Janeiro

Thiago Carneiro Ribeiral

Uso de Long short-term memory para identificar diferentes segmentos de vagas de emprego

Relatório Projeto Final I

DEPARTAMENTO DE INFORMÁTICA

Orientador: Prof. Eduardo Sany Laber

Rio de Janeiro
04 de 2020

Introdução

A busca e triagem de candidatos é muito importante no meio profissional para a contratação de novos funcionários. Entre as principais ferramentas na indústria para a descoberta de candidatos e de oportunidades são as postagens de currículos profissionais e vagas de emprego respectivamente.

Nestes casos há uma certa convergência no mercado fazendo com que ambos sejam normalmente divididos em certas seções. No caso dos currículos é comum a presença de segmentos como dados pessoais, introdução, experiência profissional, educação e competências. Já para vagas de emprego é comum a presença de campos como responsabilidades, requisitos, benefícios e às vezes informação sobre a empresa contratante.

O objetivo deste projeto é o desenvolvimento de um programa que utiliza técnicas de *machine learning* capaz de identificar os diferentes segmentos de vagas de emprego automaticamente. Ou seja, uma segmentação de textos não estruturados especializada no domínio de vagas profissionais. Existe espaço para melhorias nesse âmbito, já que as soluções existentes tem certas limitações, especialmente na língua portuguesa.

A identificação automática facilita o preenchimento, a visualização e comparação entre vagas. Além de ser um passo importante para a interpretação do conteúdo de uma vaga com o processamento de linguagem natural. O que por sua vez permite automatizar a busca de candidatos, a correspondência de currículo com a vaga de emprego e pode aconselhar profissionais sobre qual competência deve ser desenvolvida para aumentar suas chances de conseguir um emprego em determinada área. Ou aconselhar empresas sobre competição pela mão de obra e iniciativas de treinamento. É importante notar que pela semelhança de contexto, abordagens que avancem a segmentação de vagas podem ajudar também a avançar na segmentação de currículos e vice versa.

Situação Atual

Os grandes problemas envolvidos neste trabalho são os de segmentação e de classificação de textos. O domínio é o de mercado de trabalho na língua portuguesa, mais especificamente sobre vagas de emprego que são divulgadas para anunciar oportunidades

e atrair candidatos especializados. Enquanto a abordagem utilizada é a divisão do texto em sentenças e a utilização de redes neurais Long short-term memory (LSTM) para a classificação supervisionada destas sentenças, identificando a qual segmento elas pertencem e assim identificando a quebra de segmentos.

Em relação a segmentação de textos ou áudios existem diversas abordagens como a simples Coesão Semântica, por comparações locais como em (1). Modelos Generativos por geração de probabilidade como em (2) e (3), Algoritmos em Grafos como em (4) e (5). Além de Redes Neurais como a convolucional (CNN) em (6) e LSTM em (7).

Já em relação a classificação o mais simples é a Regressão Logística utilizado em (8) e (9). Outras opções são Máquina de Vetores de Suporte como em (10), Árvore de Decisão e Floresta Aleatória como em (11) e (12). Porém segundo (13) “Recentemente, métodos que se baseiam em redes neurais profundas tem chamado atenção por obter resultados melhores que todos os algoritmos anteriores de aprendizado de máquina”, do qual foi retirado a seguinte tabela:

Artigo	Modelo	Dataset	Acurácia (%)
XLNet: Generalized Autoregressive Pretraining for Language Understanding (14)	XLNet	DBPedia (15)	99.38
		AG News Corpus (15)	95.51
Universal Language Model Fine-tuning for Text Classification (16)	ULMFiT	DBPedia	99.20
		AG News Corpus	94.99
Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings (17)	CNN	DBPedia	99.16
		AG News Corpus	93.43

A respeito do domínio existem soluções que já extraem informações do currículo e as estruturam como em (18) e (19). Porém estas não foram desenvolvidas para textos em português, uma língua complexa, com diversas particularidades e com vasto vocabulário no âmbito profissional. Então é de vital importância, tratando de currículos e vagas brasileiras, um algoritmo especializado na língua portuguesa.

Neste quesito duas soluções foram encontradas, a primeira para currículos, o projeto final de Lucas Pavanelli (20) que inspirou o presente. Neste trabalho foi utilizado Modelos Generativos, mais especificamente uma modelagem por Hidden Markov Models (HMM) e “um algoritmo inspirado no algoritmo de Viterbi para encontrar o caminho mais provável no HMM”. Embora o projeto teve um resultado satisfatório, muito ainda pode ser melhorado na acurácia da solução, especialmente em segmentos de Educação.

A segunda solução, para vagas de emprego, é a dissertação de mestrado de David Martins (13). Foi este trabalho que inspirou a divisão do texto em sentenças e seguinte classificação delas para identificar a quebra de segmento. A dissertação implementou soluções de Naïve Bayes, Regressão Logística Multinomial, Máquinas de Suporte de Vetores e Florestas Aleatórias. E a Regressão Logística teve a melhor performance, mas todas tiveram bons resultados. As fontes de vagas utilizadas foram a Catho, LinkedIn e VAGAS.com, sendo o treino do conjunto de dados realizado com os dados da primeira.

Apesar deste último trabalho já ter obtido ótimos resultados no mesmo problema acreditamos ainda ser promissora a busca por melhores resultados utilizando redes neurais. Isto pois o próprio trabalho afirmou que redes neurais são o estado da arte de classificação, como ilustrado pela tabela anterior. Além de estarem presentes na literatura de segmentação e de classificação de textos e áudios, como já citado. Mais especificamente o uso de redes neurais LSTM como em (17) para classificação e (7) para segmentação. No mínimo há valor na comparação com as demais abordagens de (13). Além de avaliar a generalização da solução testando em fontes distintas.

Objetivos do trabalho

O produto que está sendo implementado neste trabalho é um segmentador e classificador de secções de uma vaga de trabalho. O programa deve receber como entrada uma vaga de trabalho não estruturada em formato de arquivo de texto. Então o texto deverá ser dividido em sentenças a partir das pontuações e quebra de linhas. Assumindo que cada sentença está exclusivamente inclusa em um dos segmentos possíveis, esta sentença deverá ser classificada como do respectivo segmento. Uma vez que cada sentença esteja classificada é gerado um segmentador de forma trivial, já que a quebra de segmentos será quando duas sentenças seguida foram classificadas como segmentos distintos.

Os segmentos possíveis são: responsabilidades, requisitos, benefícios e outros. O primeiro trata dos deveres e atribuições do funcionário, enquanto o segundo trata de tudo

que é requerido do candidato como habilidades, experiência, nível de educação, certificações e posse de bens materiais. Benefícios tratam das vantagens e recompensas pelo salário, assim como detalhes sobre a contratação. As demais informações são consideradas na secção outros.

Há também a restrição de que esse processo deve ser feito em até 1 segundo em média por vaga de trabalho, para garantir que o algoritmo tenha o mínimo de escalabilidade.

A abordagem que será utilizada neste projeto para classificar as sentenças nos diferentes segmentos de currículos será o uso de uma Long short-term memory (LSTM). Uma arquitetura de redes neurais recorrentes no campo de aprendizagem profunda. A LSTM é capaz de considerar dados em um passado distante que seriam esquecidos em uma rede neural recorrente comum, o que espera ser útil na identificação do contexto da vaga.

A tabela abaixo, retirada de (20), apresenta os resultados conseguidos pelo trabalho mencionado de Pavanelli utilizando HMM's. Estes valores representam a média de precisão e revocação na identificação de termos por segmento de todos currículos e são uma referência para o projeto atual, devido à também ser um projeto de conclusão de curso no mesmo domínio

Segmento	Precisão	Recall
Dado Pessoal	93.71	10.79
Experiência de Trabalho	65.13	99.90
Educação	21.57	20.70
Competência	53.78	50.28
Outros	50.00	0.20

Outra referência é a dissertação de mestrado de Martins (13) que entre os resultados estão uma acurácia 95.58% com os dados do Catho, 88.60% com os dados do VAGAS.com.br e 91.14% com os dados do LinkedIn. Além das métricas de segmentação Pk e WindowDiff obtidas, 3.66% e 4.78% respectivamente. No qual Pk é uma métrica de segmentação que representa a probabilidade de erro na segmentação. Enquanto WindowDiff é uma versão aprimorada de Pk que também considera também quantidade de segmentos.

Apesar de ser o ideal, o objetivo do trabalho não é necessariamente conseguir valores melhores que estes projetos e sim promover um desenvolvimento exploratório. O grande intuito é comparar as abordagens distintas e estudar o desempenho da LSTM para esse problema.

Atividades realizadas

Estudos conceituais

A primeira etapa do projeto foi um estudo bibliográfico para o melhor entendimento do problema e das abordagens que serão utilizadas ou tidas como referência. Esta se deu início por um estudo de revisão sobre o básico de modelos de classificação, visto em (21). Em seguida o estudo mais importante, visando aprender a base teórica sobre LSTM e redes neurais a partir de (22). Houve mais um estudo adicional sobre HMM considerando (23), além de estudar as abordagens e trabalhos no mesmo domínio citados, principalmente (20) e (13).

Coleta de dados

Um dos recursos mais importantes para que seja possível implementar o classificador de sentenças é a abundância de dados para treino e teste do algoritmo. Logo foi necessário buscar uma grande quantidade de vagas de emprego anotadas. Há muitos sites que disponibilizam vagas publicamente, então o primeiro passo foi fazer uma seleção entre estes. Dos cem sites analisados onze foram pré-selecionados pela riqueza na descrição das vagas, consistências e presença dos campos principais de requisitos e responsabilidades.

Nesse momento, também foi decidido que os possíveis segmentos seriam: ‘requisitos’, ‘responsabilidades’, ‘benefícios’ e ‘outros’, assim como em (13). Isto foi justificado pela presença quase universal destes campos e pela consequente facilidade de comparar com os resultados de (13). Um outro possível campo que foi um candidato é ‘informações sobre a empresa contratante’, porém foi incorporado ao ‘outros’ por sua presença não ser tão universal nas vagas quanto os demais.

As onze fontes pré-selecionadas consistiam em: Agrobase, Catho, Curriculum, InfoJobs, Vagas.com, Index Empregos, Manager, Balcão de Empregos, Recruta Simples, Emprego Certo e Chances. Houve uma segunda fase da seleção focando na consistência de formatação de vagas da mesma fonte e separação clara entre os diferentes segmentos da mesma vaga. Pois desse modo seria possível acelerar o processo de anotação das sentenças, utilizando a formatação natural da fonte para fazer isso de forma automática. Dessa forma as fontes selecionadas foram: Manager, Agrobase, Curriculum e Recruta Simples. É importante selecionar mais de uma fonte para garantir que o classificador consiga generalizar e não dependa apenas das características individuais de uma fonte.

As vagas de emprego foram fornecidas pela empresa Jobzi Inteligência de Dados na Internet Ltda, desta forma não foi necessário desenvolver um crawler para os sites. Foram adquiridas inicialmente 20.000 vagas da Agrobase e da Manager, assim como 10.000 vagas da Curriculum e Recruta Simples.

Preparação do conjunto de dados

A primeira etapa da preparação do conjunto de dados para aprendizado supervisionado é a quebra do texto de descrição da vaga de emprego em sentenças, identificando o segmento de cada sentença. Para fazer isto foi identificado frases, quebras de linha consecutivas, utilização de letras maiúsculas e palavras específicas que marcam o início de uma determinada secção da vaga de emprego. Estes identificadores são particulares de cada fonte e cada um é feito individualmente e as fontes escolhidas foram selecionadas justamente pela presença e confiança nestes identificadores.

O passo seguinte é considerar que toda sentença entre dois identificadores pertence ao segmento definido pelo primeiro. E final da vaga pertence ao segmento definido pelo último identificador encontrado. Já as sentenças são separadas na presença de quebra de linha e pontuações como ‘.’ e ‘;’.

O exemplo abaixo mostra uma vaga completa retirada do site Agrobase, o primeiro identificador encontrado é a frase “ATIVIDADES A SEREM DESENVOLVIDAS”. Este é considerado um identificador de Responsabilidades e todas as sentenças subsequentes de “Planificação e realização de auditorias de certificação e verificação dos programas” até “O trabalho acontece em um ambiente dinâmico com muitos projetos e clientes diferentes” são consideradas responsabilidades. O mesmo vale para as sentenças depois de “COMPETÊNCIAS NECESSÁRIAS” só que neste caso elas são consideradas como de

Requisitos. “LOCAL DE TRABALHO” identificada sentenças de Outros e tanto “REMUNERAÇÃO:” e “BENEFÍCIOS:” identificam sentenças de Benefícios. Já dentro de “Resumo” tanto “Tipo de Contrato:” quanto “Salário:” são considerados novamente identificadores de benefícios e “Local:” como outros. Todos estes termos só são considerados como identificadores se aparecem logo depois de uma quebra de linha. Assim todas as sentenças foram corretamente separadas e seu respectivo segmento foi identificado para esta vaga. Há outros possíveis identificadores na Agrobase que não aparecem nesta vaga, pois a consistência desta fonte não é perfeita.

Auditor de Programas de Biocombustíveis

1 dia atrás

ATIVIDADES A SEREM DESENVOLVIDAS:

- Planificação e realização de auditorias de certificação e verificação dos programas: ISCC, 2BSvs, RFS2, LCFS, Renovabio e similares;
- Elaboração dos relatórios de auditoria;
- Contato com o cliente para seguimento dos processos de auditoria;
- Participar de treinamentos internos e externos

O trabalho acontece em um ambiente dinâmico com muitos projetos e clientes diferentes.

COMPETÊNCIAS NECESSÁRIAS:

- Conhecimento e experiência com auditoria e processo de certificação;
- Curso de auditor lider ISO9001/ISO14001/ISO45001;
- Preferível conhecimento em indústria de biocombustíveis;
- Idiomas obrigatórios: inglês avançado;
- Boa comunicação com as partes internas e externas;
- Capacidade para colaborar com colegas nacionais e internacionais;
- Ter boa organização e gestão de tempo.

LOCAL DE TRABALHO:

- Escritório em São Paulo;
- Disponibilidade para viajar dentro e fora do Brasil.

REMUNERAÇÃO:

- A combinar.

BENEFÍCIOS:

- Auxílio Academia (SmartFit);
- Convênio Médico;
- Vale Refeição;
- Auxilio Idiomas (escola CNA);
- Vale Transporte

Além das profissões alvos relacionadas, profissionais com experiência na área de Sustentabilidade também podem aplicar.

Resumo

Tipo de Contrato: Efetivo (CLT)
Salário: não especificado
Local: São Paulo

CANDIDATAR-SE

candidatura através da plataforma de currículos Agrobase

Plano de Ação

A proposta foi desenvolvida durante a etapa de estudo bibliográfico com o seguinte plano de cronograma:

Etapa	**Data de Início**
Estudo Bibliográfico	04/03/2020
Utilização de arquiteturas LSTM	10/05/2020
Definição de um corpus sintético	10/06/2020
Desenvolvimento do segmentador baseado em LSTM	10/07/2020
Avaliação do segmentador no corpus	15/10/2020
Comparação com HMM's	01/11/2020
Escrita do Projeto Final	15/11/2020

A maior diferença em relação à proposta foi modificar o alvo do segmentador de currículos para vagas de emprego. Esta decisão ocorreu pois os currículos são muito mais escassos em relação às vagas, já que as vagas são postadas publicamente diferentemente dos currículos. Assim redirecionando o projeto para vagas podemos utilizar dados reais, de vagas postadas em sites de busca de emprego. Aumentando a confiabilidade do segmentador e de sua avaliação se comprado ao plano original que era construir um corpus sintético.

Com esta alteração, houveram algumas mudanças no cronograma, já que foi necessário adicionar as etapas de coleta e preparação do conjunto de dados ao invés da construção do corpus sintético.

Outro fator impactante foi a publicação de (13) na mesma época do desenvolvimento da proposta. Este trabalho trouxe grande progresso para o domínio e influenciou na nossa decisão de implementação. Desenvolvendo o segmentador através da divisão em sentenças e classificando estas sentenças, como explicado em ‘Objetivos do trabalho’. Porém o trabalho utilizou métodos diferentes para a classificação, então a comparação

destas abordagens com o uso de LSTM ainda segue válido. Além de uma maior generalização buscando fontes diferentes de vagas de emprego.

Cronograma e etapas futuras

Assim, segue o novo cronograma com todas as etapas previstas:

Etapa	**Data de Início**	**Data de Conclusão**
Estudo Bibliográfico	04/03/2020	07/05/2020
Seleção de fontes de vagas	08/05/2020	21/06/2020
Estudo de (9)	20/05/2020	07/06/2020
Coleta dos dados	22/06/2020	02/07/2020
Preparação do conjunto de dados e teste com segmentador por árvore de decisão	03/07/2020	14/07/2020
Utilização de arquiteturas LSTM	15/07/2020	31/07/2020
Desenvolvimento do segmentador baseado em LSTM	01/08/2020	19/10/2020
Avaliação do segmentador	20/10/2020	31/10/2020
Comparação com outros trabalho	01/11/2020	14/11/2020
Escrita do Projeto Final	15/11/2020	Prazo Final

A etapa atualmente em desenvolvimento é a de preparação do conjunto de dados e teste com segmentador por árvore de decisão. A preparação do conjunto foi descrita e exemplificada na seção de ‘Atividades realizadas’. O uso inicial de árvore de decisão para classificar as sentenças serve como teste e protótipo antes de iniciar a implementação das redes neurais mais complexas.

Em seguida será realizado um estudo prático de redes neurais, utilizando e experimentando arquiteturas de LSTM's sem precisar estar no contexto específico de vagas de emprego. Então será desenvolvido o produto principal do projeto, o segmentador de currículos baseado em LSTM.

Assim que estiver pronto será avaliado seu desempenho em relação à um conjunto de dados pré-selecionados para testes e os resultados serão comparados com os obtidos com HMM's em (20) e com as abordagens de (13), especialmente a Regressão Logística. Por fim haverá a escrita do relatório do projeto final declarando todo processo, os resultados e as respectivas conclusões.

Referências bibliográficas

[1] Marti A. Hearst. **TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages**, https://www.aclweb.org/anthology/J97-1003.pdf Association for Computational Linguistics, 1997

[2] Pedro J. Moreno, David M. Blei. **Topic segmentation with an aspect hidden Markov model**, https://dl.acm.org/doi/10.1145/383952.384021 SIGIR Set 2001

[3] Matthew Purver, Konrad P. Kording, Thomas L. Griffiths, Joshua B. Tenenbaum. **Unsupervised Topic Modelling for Multi-Party Spoken Discourse**, https://dl.acm.org/doi/pdf/10.3115/1220175.1220178?download=true NAACL, 2013

[4] Igor Malioutov, Regina Barzilay. **Minimum Cut Model for Spoken Lecture Segmentation**, https://www.aclweb.org/anthology/P06-1004.pdf ACL 2006

[5] Goran Glavas, Federico Nanni, Simone Paolo Ponzetto. **Unsupervised Text Segmentation Using Semantic Relatedness Graphs**, https://madoc.bib.uni-mannheim.de/41341/1/S16-2016.pdf ACL 2016

[6] Liang Wang, Sujian Li, Yajuan Lyu, Houfeng Wang. **Learning to Rank Semantic Coherence for Topic Segmentation**, https://www.aclweb.org/anthology/D17-1139.pdf EMNLP 2017

[7] Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, Jonathan Berant. **Text Segmentation as a Supervised Learning Task,** https://arxiv.org/pdf/1803.09337.pdf AACL 2018

[8] David W. Hosmer, Stanley Lemeshow. **Applied Logistic Regression**, volume 85. Out 2004.

[9] HARRELL, JR., F. E.. **Regression Modeling Strategies**. Springer-Verlag, Berlin, Heidelberg, 2006.

[10] HAN, E.-H.; KARYPIS, G.. **Centroid-based document classification: Analysis and experimental results**. Lecture Notes in Computer Science, 1910:424–431, 01 2000.

[11] XU, B.; GUO, X.; YE, Y. ; CHENG, J.. **An improved random forest classifier for text categorization**. Journal of Computers, 7, 12 2012.

[12] CHEN, W.; XIE, X.; WANG, J.; PRADHAN, B.; HONG, H.; BUI, D. T.; DUAN, Z. ; MA, J.. **A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility**. CATENA, 151:147 –160, 2017.

[13] David Evandro Amorim Martins. **Segmentação semântica de vagas de emprego: estudo comparativo de algoritmos clássicos de aprendizado de máquina**, Dissertação de Mestrado PUC-Rio Mar 2020

[14] YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R.; LE, Q. V.. Xlnet: **Generalized autoregressive pretraining for language understanding**, 2019.

[15] ZHANG, X.; ZHAO, J. ; LECUN, Y.. **Character-level convolutional networks for text classification**, 2015.

[16] HOWARD, J.; RUDER, S.. **Universal language model fine-tuning for text classification**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[17] JOHNSON, R.; ZHANG, T.. **Supervised and semi-supervised text categorization using lstm for region embeddings**, 2016.

[18] KACZMAREK, T.; KOWALKIEWICZ, M. ; PISKORSKI, J.. **Information extraction from cv**. 04 2019.

[19] CHEN, J.; ZHANG, C. ; NIU, Z.. **A two-step resume information extraction algorithm**. Mathematical Problems in Engineering, 2018:1–8, 05 2018.

[20] Pavanelli, L., **Algoritmo de segmentação para textos não estruturados usando HMM**, Projeto Final PUC-Rio Jan, 2019.

[21] P. Tan, M. Steinbach, and V. Kumar. **Introduction to Data Mining**, Pearson Education, (2006)

[22] Graves, A., **Supervised Sequence Labelling with Recurrent Neural Networks**, https://www.cs.toronto.edu/~graves/preprint.pdf, Feb, 2012.

[23] LAWRENCE R. RABINER, FELLOW, IEEE. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**, IEEE 1989