MODELO DE CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES NA LÍNGUA PORTUGUESA

Henrique Maia Braum[1], Sandro José Rigo[1], Jorge L. V. Barbosa[1]

[1]UNISINOS – Universidade do Vale do Rio dos Sinos

henriquebraum@gmail.com, rigo@unisinos.br, jbarbosa@unisinos.br

Abstract: The objective of this paper is to present and evaluate a question classifier model for the Portuguese language. Different machine learning algorithms, classifier parameterizations and question preprocessing are used to evaluate the model. The evaluation presents good results and the possibility of future developments in this area, which may help in the development of virtual agents and question answering systems in Portuguese.

Resumo: O objetivo deste trabalho é apresentar e avaliar um modelo de classificação de questões na língua portuguesa. Para a avaliação são utilizados diferentes algoritmos de aprendizagem de máquina, assim como parametrizações dos classificadores e pré-processamento das questões. A avaliação realizada apresenta bons resultados e a possibilidade de avanços futuros nesta área, o que pode auxiliar no desenvolvimento de agentes virtuais e sistemas de perguntas e respostas no idioma português.

1 INTRODUÇÃO

O principal objetivo da área de processamento de linguagem natural é fazer com que os computadores possam realizar tarefas relacionadas à linguagem humana. Um exemplo é uma aplicação para perguntas e respostas baseada na web. Esta é uma generalização de uma simples pesquisa na web, onde ao invés de apenas digitar palavras-chave, um usuário poderia realizar perguntas completas em linguagem natural, desde perguntas simples até as mais complexas (JURAFSKY, 2009).

Para encontrarmos a resposta certa para uma questão é essencial que a pergunta seja processada de maneira correta. Desta forma podemos identificar qual é o tipo de pergunta que está sendo feita, qual o tipo de resposta que é esperada, qual é o foco da pergunta e que palavras-chave compõem essa questão (MOLDOVAN, 2000). O grande volume de dados disponibilizados em formato digital e a ampla adoção de mediação digital na Educação são fatores que geram o interesse em ferramentas que possa apoiar os estudantes em uma interação mais amigável com os atuais sistemas, por exemplo, usando a linguagem natural para a realização de perguntas.

Como a maioria dos avanços e ferramentas da área são desenvolvidas na língua inglesa, além do fato de o processamento de linguagem natural ser um assunto onde ocorreu que as regras que se aplicam para um idioma podem não ser as mesmas aplicadas a outros, este artigo visa explorar o processamento de questões na língua portuguesa. Com isto, este trabalho tem como motivação o fato de: (i) a classificação de questões ser uma área ainda pouco explorada; e (ii) apoiar a divulgação de propostas de modelos de classificação para a língua portuguesa.

O objetivo geral do trabalho é propor e testar um modelo para classificação de questões na língua portuguesa. Para que este objetivo seja alcançado, foram definidos os seguintes objetivos específicos: a) Estudar modelos de classificação de questões já propostos para outros idiomas; b) Treinar o modelo de classificação de questões com uma carga considerável de dados; c) Realizar testes para medir o desempenho do modelo de classificação de questões.

Foram realizadas análises de trabalhos relacionados e estudadas iniciativas na área de classificação de questões, tendo sido adotada a categorização de Li e Roth (2002) como base para testes na utilização de diversas formas de uso de classificadores com o corpus descrito por estes autores. Os testes gerais mostraram bons resultados e permitiram análises de resultados com diferentes abordagens, tanto com classificadores como com atividades de pré-processamento e seleção de características das questões a classificar.

O restante do texto segue estruturado como descrito. Na seção 2 é apresentado o estudo de trabalhos relacionados que apoiaram o desenvolvimento deste trabalho. As seções 3 e 4 apresentam, respectivamente, o modelo proposto e o estudo de caso utilizado para avaliações preliminares da abordagem adotada. Por fim, na seção 5 são apresentadas as conclusões do estudo.

2 TRABALHOS RELACIONADOS

O trabalho desenvolvido por Xin Li e Dan Roth (2002) apresenta uma abordagem baseada em aprendizado de máquina para a classificação de questões. É utilizado um classificador hierárquico guiado por uma hierarquia semântica em camadas de tipo de resposta Foi definida então uma taxonomia de duas camadas, a qual representa uma classificação da semântica natural para típicas questões encontradas na TREC (*Text REtrieval Conference*). Esta hierarquia contém 6 classes brutas (Abreviação, Entidade, Descrição, Humano, Localidade e Valor

Numérico) e 50 classes refinadas. Cada classe bruta possui um conjunto sem sobreposição de classes refinadas. Cada questão é analisada e representada em uma lista de características à serem tratadas no treinamento para a aprendizagem. Os tipos primitivos de características incluem *words* (as palavras em si), *pos tags* (marcação de *Part of Speech*), *chunks* (frases que não se sobrepõem), *named ent*ities (entidades identificadas pelo nome), *head chunks* (a parte com o primeiro substantivo) e *semantically related words* (palavras que ocorrem em uma classe específica de questões).

O LASSO é um sistema de perguntas e respostas desenvolvido na Southern Methodist University (Texas, Estados Unidos). Para encontrar respostas, esse sistema se baseia em uma combinação de técnicas sintáticas e semânticas (MOLDOVAN, 2000). Sua arquitetura compreende três módulos: Processamento de Questões, Indexamento de Parágrafo e Processamento de Respostas. Para este trabalho iremos analisar apenas o módulo de processamento de questões. A principal função do módulo de questões é encontrar o tipo da questão através da taxonomia das questões construídas no sistema. Além disso ele determina o tipo de resposta esperada, constrói o foco da questão e identifica as suas palavras-chave.
Os resultados obtidos pelo projeto LASSO foram satisfatórios, possuindo um resultado de 55,5% de acertos para respostas curtas e 64,5% de acertos para respostas longas.

3 MODELO PROPOSTO

Nesta seção será discutido o modelo proposto para a execução deste trabalho.

A taxonomia adotada para este trabalho será a mesma proposta por Li e Roth (2002), ela é dividida em 6 classes brutas e 50 classes refinadas.

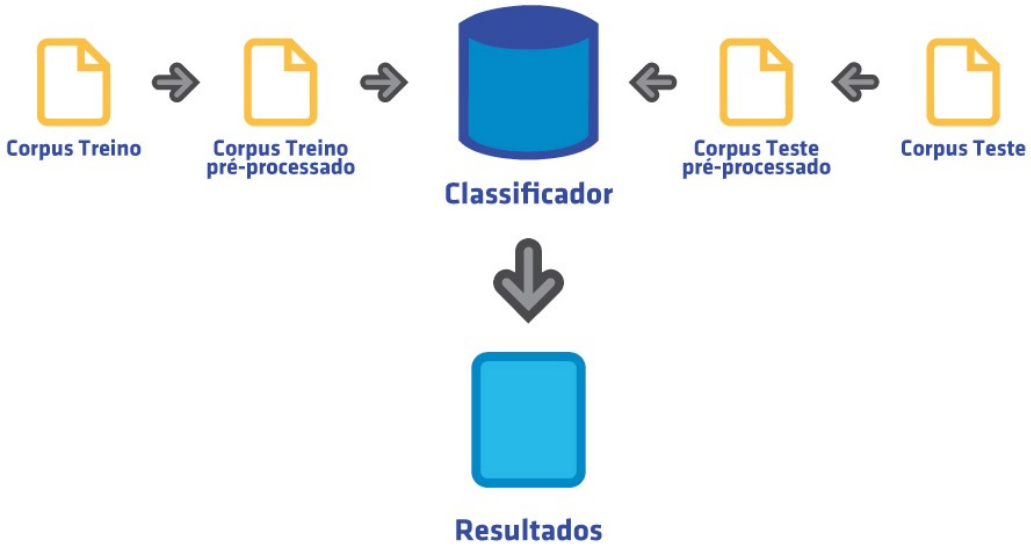
As classes brutas foram definidas da seguinte forma:

- **Abreviação:** Questões relacionadas com abreviação podem ter duas formas, ou o usuário está interessado em saber como abreviar uma expressão, ou ele tem uma expressão e quer saber o que ela significa.
- **Entidade:** A classe de entidade é a maior classe bruta da taxonomia, ela representa diversas entidades como animais, cores, comidas, produtos e afins.

- **Descrição:** São o tipo de questões que exigem respostas mais elaboradas, perguntas como “por que” ou “como” na sua maioria se encaixam neste grupo.
- **Humano:** As questões relacionadas com pessoas e organizações são pertencentes a esta categoria. Na maioria das vez são questões que começam com a palavra “quem”.
- **Localidade:** A categoria de localidade cobre todas as questões referentes a algum local específico. Em sua maioria as questões começam com a palavra “onde”.
- **Numérico:** As questões da classe numérica são todas as questões onde a resposta é algum número. Esses números vão desde a quantidade de alguma coisa, datas, idades até códigos postais.

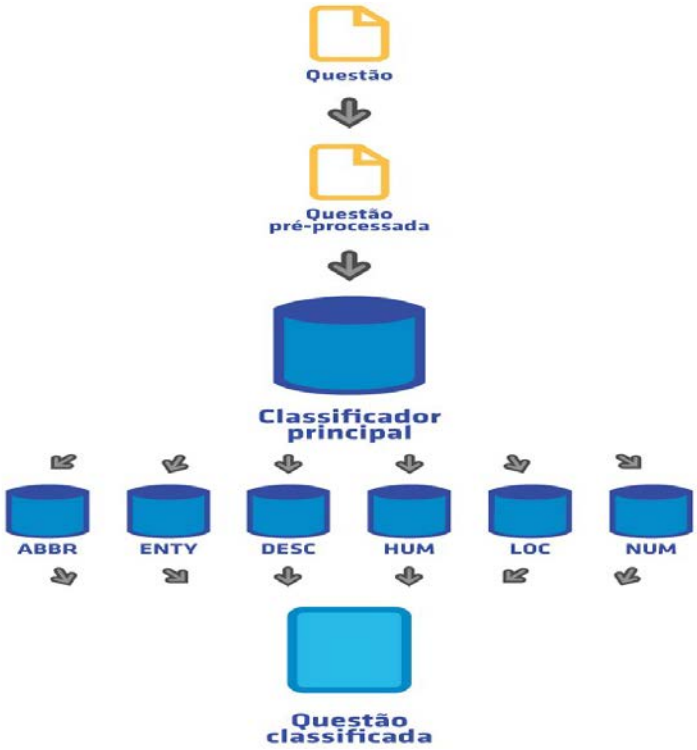
Foram utilizados dois tipos de corpus diferentes para realizar o treinamento do classificador, ilustrados na figura 1. O primeiro corpus é para o treinamento dos classificadores, esse corpus será pré-processado antes de alimentar o classificador com amostras e o resultado esperado para cada uma. O segundo corpus é o corpus para testes. Esse corpus passará pelo mesmo pré-processamento que o corpus de treino, porém seu objetivo é testar o classificador para que seja possível realizar comparativos entre mudanças no pré-processamento, algoritmo e parâmetros do classificador.

Figura 1 – Fluxo para treinamento e teste do classificador



Para auxiliar na tarefa da classificação da classe e subclasse, será utilizado um classificador hierárquico. O classificador hierárquico possui um classificador principal que faz a classificação da classe bruta e um classificador para cada uma dessas classes brutas. A vantagem de utilizar uma estrutura de classificação hierárquica para classificar uma subclasse é que esses classificadores de subclasses são treinados especificamente para cada classe, fazendo com que ao invés de ter que classificar uma subclasse tendo 50 outras opções, cada classificador possui apenas as subclasses referentes a sua classe principal.

Figura 2 – Fluxo para classificação de uma questão



Como podemos ver na figura 2, antes de ser classificada a questão passará pelo mesmo pré-processamento das questões utilizadas para treinar os classificadores. Isso faz com que a questão esteja no mesmo padrão que as questões que o classificador conhece. Após o pré-processamento, o classificador principal irá identificar a qual classe principal esta questão pertence, e baseado nessa classificação irá direcionar a questão para o classificador secundário correto. O classificador secundário irá então identificar a subclasse desta questão, entregando como resultado a questão devidamente classificada.

Foi implementada uma interface web que permite a utilização da implementação que foi realizada para o modelo. Esta interface foi desenvolvida para disponibilizar um formulário no qual um usuário pode digitar em língua natural, em português, uma frase contendo uma pergunta. Nesta mesma interface o usuário recebe como resultado a frase classificada, de acordo com o conjunto de classes utilizadas.

A Figura 3 exemplifica a forma empregada para esta implementação e o seu uso, com um dos resultados como exemplo.

Figura 3 – Exemplo de uso da interface Web implementada

Esta interface está disponibilizada para uso e testes a partir da seguinte url: http://henriquebraum.pythonanywhere.com/, sendo considerada neste momento como um elemento experimental do trabalho.

4 ESTUDO DE CASO

Esta seção detalha todos os passos seguidos durante a implementação do modelo proposto na seção anterior, desde a criação dos corpus para treino e teste, escolha do algoritmo de classificação, táticas de pré-processamento e avaliação dos resultados.

Para a execução e avaliação dos resultados foram criados três corpus de treino e um corpus de teste. O primeiro corpus de treino foi criado com 500 questões, o segundo com 1000 e o terceiro com 1500, sendo que o conteúdo do primeiro corpus foi incluído no segundo, e o conteúdo do segundo foi incluído no terceiro corpus. O corpus de teste foi criado com 350 questões. A fonte para criação

de todos os corpus foram questões utilizadas por Li e Roth (2002) traduzidas para o português, assim como questões de autoria própria e questões criadas com o apoio do setor de Linguística da Unisinos. A tabela 1 mostra a distribuição das classes principais dentro de cada corpus.

Tabela 1 – Distribuição das classes principais dentro dos corpus

	Corpus Treino 1	**Corpus Treino 2**	**Corpus Treino 3**	**Corpus Teste**
ABBR	15	23	38	14
ENTY	130	254	489	88
DESC	102	203	261	61
HUM	78	158	233	48
LOC	75	149	190	57
NUM	100	213	289	82
Total	**500**	**1000**	**1500**	**350**

Após a criação dos corpus para treino e teste foram escolhidos seis algoritmos para serem testados e avaliados. O principal critério para a escolha destes seis algoritmos foi o suporte a classificação multiclasse (possibilidade de um classificador ser treinado em mais do que duas classes). Além do algoritmo, foi também necessário definir como que os algoritmos iriam ser treinados. Como não é possível treinar um classificador a partir de textos, foram utilizados dois diferentes métodos para extrair *features* (características usadas para treinar o classificador) dos textos. Os dois métodos utilizados foram: a) **Bag-of-words:** Este método consiste em reunir todas as palavras de todos as questões em uma única lista, sem duplicar nenhum palavra. Desta maneira cada questão é representada como uma lista numérica, onde 0 indica que a palavra não está presente na questão e valores maiores que 0 indicam quantas vezes a palavra se repete na questão. b) **Tf-idf:** Este método é uma extensão do *bag-of-words* onde quanto mais vezes uma palavra aparece na questão, menor é a sua relevância para a classificação da mesma.

Cada algoritmo foi avaliado com base em três medidas: a) **Precision:** A habilidade de um classificador em não classificar uma amostra negativa como positiva; b) **Recall:** A habilidade do classificador em encontrar todas as amostras positivas; c) **F1-Score:** É uma medida que considera tanto a *precision* quanto o *recall* para gerar uma pontuação, sendo a melhor medida para se avaliar um classificador.Analisando os resultados obtidos foi possível identificar os três

algoritmos que tiveram o melhor desempenho na tarefa de classificar a classe principal para as questões do corpus de teste. O melhor resultado foi do LinearSVC que obteve uma pontuação de 87% quando treinado com o corpus de 1500 questões e utilizando o método tf-idf para extração das *features*. Os outros dois melhores resultados foram obtidos pelo SVC com *kernel* linear, que possui uma implementação um pouco diferente do LinearSVC, e pelo LogisticRegrassion. Também podemos analisar que, para estes três algoritmos, os resultados foram melhorando conforme o tamanho do corpus utilizado.

N-gramas são excelentes aliados para classificadores de texto pois ajudam a identificar a ordem em que palavras aparecem em uma frase. Para auxiliar na melhoria dos resultados do classificador, os três melhores algoritmos da primeira etapa foram testados com a utilização de unigramas, bigramas e trigramas. Os resultados destes testes podem ser visualizados na tabela 2, sendo que o resultado que aparece para cada um é a pontuação do F1-Score.

Tabela 2 – Desempenho de n-gramas na classificação de questões

Algoritmo	Corpus	Feature	Unigrama	Bigrama	Trigrama
LogisticRegression	500	bow	70%	70%	68%
		tfidf	66%	68%	68%
	1000	bow	75%	77%	74%
		tfidf	70%	72%	72%
	1500	bow	84%	86%	84%
		tfidf	77%	80%	79%
LinearSVC	500	bow	73%	72%	70%
		tfidf	74%	76%	76%
	1000	bow	78%	78%	75%
		tfidf	79%	82%	80%
	1500	bow	86%	89%	88%
		tfidf	87%	91%	90%
SVC (kernel linear)	500	bow	72%	71%	69%
		tfidf	71%	74%	74%
	1000	bow	76%	78%	74%
		tfidf	79%	79%	79%
	1500	bow	85%	88%	86%
		tfidf	85%	90%	89%

É possível analisar que em geral os melhores resultados foram alcançados por bigramas, enquanto que em diversos casos os trigramas tiveram resultados piores até mesmo que os unigramas. Com base nestes resultados o algoritmo

LinearSVC, com o método de extração de *features* tf-idf e bigramas foram selecionados para serem usados até o final deste trabalho.

O pré-processamento tem como tarefa transformar uma questão em algo que faça mais sentido para o classificador. Os seguintes pré-processamentos foram aplicados neste trabalho: Reconhecimento de entidades, Remoção do gênero de artigos; Remoção do gênero da preposição “de”; Reconhecimento do verbo “Ser”. A tabela 3 mostra o ganho proporcionado pelo pré-processamento das questões para o algoritmo LinearSVC utilizando tf-idf e bigramas.

Tabela 3 – Efeitos da utilização de pré-processamento

Algoritmo	Corpus	Pré-processamento	Precision	Recall	F1-
LinearSVC	500	Não	78%	76%	76%
		Sim	86%	85%	85%
	1000	Não	83%	82%	82%
		Sim	89%	89%	88%
	1500	Não	91%	91%	91%
		Sim	94%	93%	93%

A utilização de pré-processamento melhorou consideravelmente o resultado dos corpus menores e proporcionou um ganho de cerca de três por cento no corpus maior.

5 CONCLUSÃO

O presente artigo estudou a eficiência de um classificador de questões na língua portuguesa, montado com aprendizado de máquina e baseado na taxonomia de Li e Roth (2002). O principal objetivo deste trabalho foi concluído, considerando que foi proposto e avaliado um modelo para classificação de questões na língua portuguesa. Podemos considerar os resultados obtidos pelo classificador como extremamente satisfatórios, tendo uma pontuação de 94% durante a classificação da classe bruta e uma média de 91% entre os classificadores secundários.

Ao final deste trabalho, diversos caminhos diferentes surgiram para trabalhos futuros, entre as quais se destacam a criação de um agente virtual para apoio ao trabalho dos estudantes, baseado em classificadores hierárquicos para identificar o módulo e submódulo que possui a resposta para a pergunta de um usuário. Além disso também foi relacionada a criação de um sistema de perguntas e respostas que utilize o classificador para auxiliar na hora de encontrar a resposta.

6 REFERÊNCIAS

CORTES, C.; VAPNIK, V. N. Support Vector Machines, Machine Learning 20: 273–297, 1995.

HIRSCHMAN, L.; GAIZAUSKAS, R. Natural Language Question Answering: The View from Here. Natural Language Engineering, Volume 7, Edição 4, p.275-300, 2001

JOACHIMS T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of ECML98, 10th European Conference on Machine Learning, 1998.

JURAFSKY, Daniel; James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall, 2009.

LI, Xin; ROTH, Dan. Learning Question Classifiers, Proceedings of the 19th International Conference on Computational Linguistics, p.1-7, 2002.

LIDDY, R. Natural Language Processing, Library and Information Science, Marcel Drecker Inc. New York, USA, 2a Ed. 2003.

MOLDOVAN, Dan. The Structure and Performance of an Open-domain Question Answering System. ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, p.563-570, 2000.

PINTO, D.; BRANSTEIN, M.; COLEMAN, R.; CROFT, W. B.; KING, M., LI, W.; WEI, X. QuASM: A System for Question Answering using Semi-structured Data, Proceedings of the Joint Conference on Digital Libraries, 2002.

WITTEN, I. H.; FRANK, E. & HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques , Morgan Kaufmann , Amsterdam, 2011