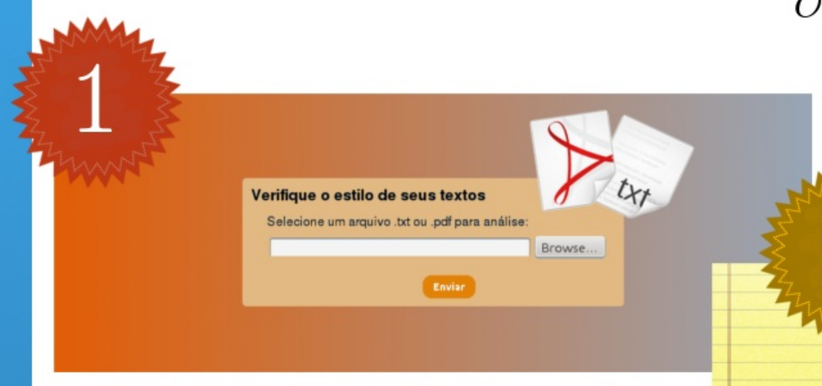
Collaborative Academic Text Advisor

Um verificador de **estilo** para textos acadêmicos de Computação

Ana Luiza Domingues Fernandez Basalo
Orientada pelo Professor Marco Aurélio Gerosa



Envie um arquivo .txt ou .pdf
Selecione um arquivo de texto sem formatação ou um PDF e envie para verificação.

Alternativas aos problemas de estilo
CATA marca os problemas de estilo de seu texto e oferece sugestões para corrigi-los.

Motivação

Escrever textos acadêmicos é uma tarefa árdua para muitos dos alunos de Computação: esses estudantes não foram habituados a lidar com a escrita durante sua formação e, além disso, há questões de terminologia e tradução inerentes à área. Disso decorrem diversos problemas de escrita, que levam orientadores a empregar um longo tempo corrigindo textos e prejudicam a disseminação dos resultados das pesquisas desenvolvidas. Editores de texto oferecem corretores ortográficos, gramaticais e, às vezes, de estilo. Contudo, não são encontrados corretores específicos para estilos e termos da área de Computação. Com esta motivação, foi desenvolvido o **Sistema CATA: Collaborative Academic Text Advisor.**

Objetivo

Produzir um sistema de software que seja capaz de verificar o estilo de textos acadêmicos da área de Computação, aperfeiçoando sua avaliação e a qualidade de suas análises a partir de informações fornecidas voluntariamente por seus usuários.

Para cumprir estes objetivos, foram estudadas soluções nos campos de Processamento de Linguagens Naturais (PLN) e Aprendizagem Computacional. Ambos são ramos da Inteligência Artificial: o primeiro preocupa-se em desenvolver teorias e técnicas para analisar, entender e gerar computacionalmente textos em línguas que humanos usam naturalmente (as chamadas línguas naturais). O último dedica-se a algoritmos que possibilitam que computadores "aprendam" - onde "aprender" significa inferir padrões generalizações a partir de dados para fazer previsões sobre outros dados que poderão ser encontrados no futuro.

Estilo

Em Linguística, "estilo" pode ser entendido como o conjunto de características que definem a "personalidade" de um texto e é resultado de uma série de escolhas linguísticas feitas pelo autor. Em particular, em textos acadêmicos, essas escolhas devem ser feitas de maneira a produzir clareza e objetividade. Assim, expressões ou figuras de linguagem que poderiam trazer mais expressividade a outros tipos de textos, para textos acadêmicos são consideradas problemas. Alguns desses problemas são: alteração da organização sintática natural, repetição excessiva das mesmas palavras ao longo do texto, uso de expressões mal traduzidas, entre muitos outros. Para este trabalho, os "erros" de estilo foram restringidos ao uso de expressões ou termos inadequados - como estrangeirismos, traduções incorretas, clichês, pleonasmo, etc. Por exemplo, escrever "o algoritmo retorna um determinado valor" (tradução de "return"), quando mais elegante seria "o algoritmo devolve um determinado valor", ou "testes unitários" como tradução para "unit tests", em vez da forma mais adequada "testes de unidade".

Análise dos textos

Dado um texto acadêmico, antes de efetuar "uma busca" pelos problemas de estilo, é necessário, primeiro, processá-lo, seguindo um determinado fluxo de operações:

Segmentação
Não é desejável que detalhes como espaços em branco alterem a verificação de estilo do texto. Assim, o primeiro passo, ao receber um texto, é realizar a chamada "segmentação" (em inglês "tokenization"), para extrair as frases e termos do texto.

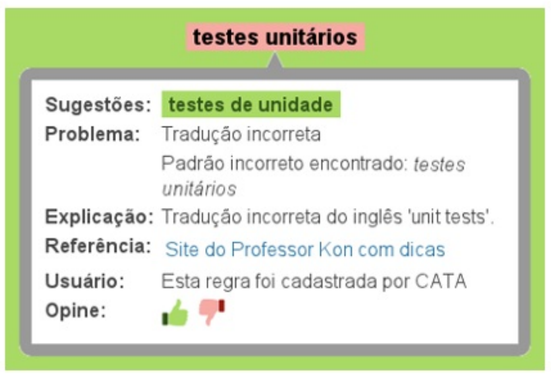
Lematização
Considere o verbo - inexistente na língua portuguesa - "debugar" (tradução incorreta do inglês "to debug"). Se tal verbo ocorresse num texto acadêmico, gostaríamos que o sistema o apontasse como um problema de estilo (a sugestão seria trocar por "depurar"). Mas isso também deveria funcionar para as seguintes flexões do verbo: "debugo", "debuguei", "debugamos", etc. Seria cansativo ter que registrar todas as flexões do verbo para poder detectá-lo como erro e até mesmo ineficiente para o sistema ter que armazenar e buscar todas estas flexões, sendo que, o que se quer, na verdade, é saber se uma determinada palavra deriva ou não do verbo "debugar". Para resolver este problema, é realizada a "lematização" (do inglês "lemmatization") do texto, para obter os lemas dos termos (o lema é a palavra em sua forma não flexionada - e.g. "menino" é o lema de "meninas").

Busca de padrões

Uma vez executado o pré-processamento do texto, é possível partir para a "busca" dos problemas de estilo. Como tais problemas foram restringidos apenas a termos inadequados, esta etapa reduz-se a uma busca de padrões no texto. Para cumprir esta tarefa, foi usado o algoritmo Aho-Corasick, que, basicamente, constrói uma máquina de estados (autômato finito) que assemelha-se a uma "trie" com ponteiros entre os nós internos. Para a busca dos padrões propriamente dita, o tempo consumido é linear no comprimento do texto mais o número de padrões encontrados (o algoritmo "percorre" o texto uma única vez).

Inteligência Coletiva

Apesar do problema da verificação de estilo ter sido simplificado para poder ser tratado por um sistema de software, ainda assim as técnicas descritas até aqui não são suficientes para que a análise dos textos seja sempre adequada. Considere este exemplo: o termo "cores" pode assumir vários significados - pode ser a expressão em inglês para "núcleos" (de processadores) ou o plural do substantivo "cor". O primeiro caso é considerado um estrangeirismo inadequado e o sistema deveria marcá-lo como problema. Contudo, pode ser que o sistema aponte como erro de estilo o termo "cores" num contexto em que a palavra assume o segundo significado. Para lidar com este tipo de situação, foram utilizados conceitos de Inteligência Coletiva - cujo propósito é combinar o conhecimento de várias pessoas para criar soluções e ferramentas mais poderosas. Assim, quando um problema de estilo é indicado, o usuário pode reportar ao sistema se concorda ou não com aquela indicação. Em ambos os casos, o software infere qual a semântica do contexto (de que assunto trata o texto no trecho em que foi detectado o erro) e armazena esta informação para, futuramente, aplicar ou não a sugestão de estilo num termo que está inserido num contexto similar.



Conclusão

Sendo a escrita a principal forma de comunicação científica, melhorar o estilo dos textos acadêmicos significa melhorar a comunicação entre os acadêmicos, o que leva a uma maior colaboração entre eles.
Ainda que as características das linguagens naturais representem grandes desafios, este trabalho mostrou que é possível produzir um sistema de verificação de estilo. Para este problema, a avaliação de um ser humano ainda é bastante superior, mas o uso de um sistema computacional mostra-se eficaz para automatizar, ainda que em parte, a tarefa.

USP github SOCIAL CODING Java vraptor jQuery write less, do more.