3
Tarefa

Esse capítulo começa dissertando sobre as vantagens de se agrupar as palavras em classes, como elas são agrupadas em *part-of-speechs* e suas aplicações. Em seguida é apresentado o Anotador Morfossintático como um etiquetador de *part-of-speechs*. Por último, discutem-se as dificuldades de implementação de um etiquetador para a linguagem português-twitter.

3.1
Classificação de palavras

A primeira vantagem de se definir classes é que se torna possível fazer afirmações gramaticais com o máximo de economia [6]. Por exemplo, podemos generalizar que verbos regulares, da primeira conjugação, tem a terminação '*-as*' na segunda pessoa do singular. Além disso, a observação dos membros de uma mesma classe nos permite descobrir novas características em comum. Às classes nas quais se agrupam palavras e unidades morfossintáticas de uma linguagem natural, de acordo com seu comportamento sintático no contexto, é dada o nome de classes gramaticais, classes de palavras, ou mais comumente de *part-of-speechs*.

No processamento de linguagem natural, o *part-of-speech* é utilizado para auxiliar na resolução de tarefas linguísticas mais complexas, tais como *phrase chunking* [7] e *named entity recognition* [8]. Também é utilizado em aplicações de TI como, por exemplo, na indexação e recuperação de textos, onde substantivos e adjetivos são melhores candidatos para índices que outras classes. Outra tarefa em que o *part-of-speech* é bastante utilizado é no processamento de fala. Palavras com a mesma escrita podem ter pronúncia diferente, dependendo da classe.

3.2
Anotador Morfossintático

Nossa tarefa é criar um Anotador Morfossintático para *tweets* em português. Também conhecido como Part-Of-Speech Tagger, o Anotador Morfossintático é um classificador que atribui uma etiqueta de POS

a cada palavra de um texto. Esta etiqueta[1] indica a que classe gramatical pertence aquela palavra.

A Figura 3.1 mostra o exemplo de um texto anotado com etiquetas POS. A frase anotada é *A casa caiu.* Nela, a palavra *A* está marcada com a etiqueta *ART*, indicando que ela é um artigo. A palavra *casa* está etiquetada com *N*, significando que ela é da classe dos substantivos e a palavra *caiu* está com a etiqueta *V*, indicando que ela é um verbo.

A/ART casa/N caiu/V ./.

Figura 3.1: Frase com anotação morfossintática.

O conjunto de etiquetas POS é escolhido de acordo com a aplicação que é feita dele. As principais variações ocorrem na granularidade das classes. Por exemplo, a classe das conjunções pode ser representada pela etiqueta *K* ou dividida em duas subclasses: a das conjunções coordenativas *KC* e a das conjunções subordinativas *KS*. O conjunto de etiquetas utilizadas nessa dissertação é apresentado na Seção 5.2

3.3 Anotador Morfossintático para o Português-Twitter

Uma estratégia simples para gerar uma anotação morfossintática para *tweets* em português-twitter é aplicar um Anotador Morfossintático para o português nesses textos. A acurácia dos melhores processadores para essa anotação no português é de aproximadamente 96%. Adicionalmente, é necessária uma etapa preliminar de normalização dos *tweets* onde são eliminadas variações morfológicas elementares, sem o que a qualidade das anotações deteriora muito. Adotando esta estratégia, observamos uma acurácia de 82,76%. Isso, porque o corpus de teste já está *tokenizado*, simplificado e com a caixa das palavras corrigidas. Caso contrário, o valor poderia ser ainda menor. Na Figura 3.2, apresentamos um exemplo ilustrativo dessa anotação, onde comparamos a anotação dourada (GOLD) de um *tweet* com a gerada pelo Anotador Morfossintático do português (PT). É fácil perceber uma alta taxa de erros desse anotador automático simples. Considerando que a anotação morfossintática é utilizada para auxiliar na resolução de tarefas linguísticas mais complexas, tais como *phrase chunking* [7] e *named entity recognition* [8], um valor de 17,24% para a taxa de erro prejudica muito a qualidade das tarefas subsequentes.

[1]Em inglês, as etiquetas são também chamadas de *tags*.

	O	**q**	**eh**	**q**	**vc**	**q**	**?**
GOLD	PROSUB	PROSUB	V	KS	PROPESS	V	?
PT	ART	N	ADJ	ADJ	ADJ	ADJ	?

Figura 3.2: Anotação automática simples de um *tweet*.

Observando as mensagens escritas no Twitter entendemos esse baixo rendimento do Anotador Morfossintático para o português. Essas mensagens não foram escritas na norma culta do português, mas em uma linguagem originada a partir do português para o Twitter, que denominamos como português-twitter. Na Figura 3.3, apresentamos um *tweet* correspondendo a frase *Eu estou é com desejo de comer brigadeiro.*

Português	Eu	estou	é	com	desejo	de	comer	brigadeiro
Tweet	eu	to	eh	c/	10sejo	d	cume	brigadeiro

Figura 3.3: Mensagem em português-twitter.

3.4 Português-Twitter

A linguagem português-twitter surgiu de maneira espontânea, a partir da adequação do português ao Twitter e a Internet. Como consequência, seu léxico é formado pelo léxico do português em conjunto com terminologias da Internet e do Twitter.

Abaixo, alguns termos e elementos encontrados na linguagem:

Site - Uma página da internet.

URL - Endereço de um site na web.

E-mail - Endereço do correio eletrônico de um usuário.

Emoticon - É uma sequência de caracteres tipográficos que tem como objetivo expressar o estado emotivo de quem os emprega.

Tweet - São as mensagens, ou *posts*, utilizados para atualizar o status do usuário.

Tweeple ou tweeps - São os usuários, membros do Twitter.

Hashtags ou Tópicos - Marcam os *tweets*, possibilitando o acompanhamento em tempo real de assuntos no Twitter.

Retweet ou RT - Também conhecido como RT, é o *reposting* de um *tweet* interessante por outro usuário.

Detweet - Reposting de um *tweet* com um grau de desaprovação.

Follow - É o ato de seguir outro usuário. É o mesmo que se inscrever para receber os *tweets* escritos por ele.

Followers - São os seguidores de um determinado usuário.

Following - São os usuários seguidos por um determinado usuário.

O português-twitter permite que o usuário se expresse de maneira mais livre e direta. Essa liberdade, característica de linguagens da INTERNET, introduz dificuldades na geração de um anotador morfossintático. Por ser uma linguagem muito dinâmica, ela não tem uma gramática bem definida. Outro problema decorre da grafia das palavras no português-twitter. Uma mesma palavra pode ter diversas grafias, decorrente de fatores tais como:

- Repetição de letras;
- Inclusão de fonemas;
- Substituição de partes de palavras por símbolos, fonemas ou números;
- Substituição do acento agudo nas oxítonas pela letra *h*.

Na Figura 3.4 vemos o exemplo de uma mensagem escrita em português (PT), junto com algumas de suas equivalentes escritas em português-twitter (PT-TW), a linguagem escolhida para ser estudada nessa dissertação.

PT	Estou com desejo de comer brigadeiro.
PT-TW	Estou c/ desejo d brigadeiro :PPP
PT-TW	to c 10sejo d brigadeiro :PPP
PT-TW	TO CUM DESEJO DI BRIGADEROOOOOO

Figura 3.4: Exemplo de um *tweet* escrito em português e alguns de seus possíveis equivalentes em português-twitter.

A limitação de caracteres é outro fator importante a ser discutido. Os usuários desenvolvem táticas para economizar caracteres, o que aumenta a dinamicidade da linguagem. Algumas táticas são:

- O uso de *emoticons* para expressar emoções.
- O uso de contrações, abreviações e numerais sempre que possível.
- Pronomes e pontuações muitas vezes são esquecidos.

– Partes de palavras, ou palavras inteiras, são muitas vezes substituídas por símbolos, números e representações fonéticas.

Outras características do português-twitter que o distinguem do português, são:

– Grande presença de estrangeirismo.
– O emprego de letras maiúsculas e minúsculas não segue a regra da língua de origem.
– O uso de acentos não é obrigatório.