Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos

Carlos Eduardo Dantas de MENEZES
e-mail: carlos.edmenezes@sp.senac.br
LIAR – Laboratório de Inteligência Artificial e Robótica,
Faculdade de Ciências Exatas e Tecnologia, SENAC
São Paulo, SP, CEP 01506-000, Brasil

João JOSÉ NETO
e-mail: jjneto@pcs.usp.br
Departamento de Computação e Sistemas Digitais (PCS), EPUSP, Universidade de São Paulo
São Paulo, SP, CEP 05508-900, Brasil

RESUMO

Este trabalho tem por objetivo propor um método de construção de um analisador morfológico de textos em linguagem natural que apresente uma arquitetura híbrida: o etiquetador conterá um conjunto de regras escritas à mão, que cuidam de um grande número de casos, e permitirá que outras regras sejam aprendidas, sendo assim, treinável a partir de exemplos (textos manualmente anotados ou corpora etiquetados). Trata-se de um sistema de aprendizado automático, que infere informações lingüísticas, relativas a aspectos lexicais e contextuais de todo um corpus de treinamento. Estas informações são armazenadas, codificadas com base em autômatos adaptativos, e posteriormente utilizadas para a tarefa de classificação ou etiquetação morfológica.
Os autômatos adaptativos mostraram-se adequados tanto para o fluxo de controle da heurística de aprendizado, como também para nele codificar todos os dados necessários.
Todos os testes foram realizados com textos da língua portuguesa.

Palavras-chave: processamento de linguagens naturais, processamento automático da língua portuguesa, etiquetador morfológico, autômatos adaptativos, aprendizado automático.

1. INTRODUÇÃO

Um etiquetador morfológico tem por função associar a cada palavra uma etiqueta, que corresponda a sua categoria morfológica. Sua aplicação encontra-se em sistemas de tradução automática, em sistemas de auxílio à criação de corpora lingüísticos anotados, entre inúmeras outras tarefas do processamento de linguagens naturais [5, 8 e 10].
A principal dificuldade existente na tarefa da classificação morfológica encontra-se em sua susceptibilidade à ambigüidade. Um etiquetador morfológico robusto deve levar em conta não apenas as informações lexicais da palavra a ser anotada, mas também informações a respeito do contexto em que esta palavra se encontra.

O estado da arte em etiquetadores morfológicos

Basicamente, pode-se dizer que quatro paradigmas ou métodos constituem o estado da arte na classificação morfológica de textos em linguagem natural: o estatístico [3], o que se utiliza de regras escritas manualmente [7], o baseado em regras inferidas automaticamente [2] e o com base em exemplos memorizados [4]. Todos eles conseguem uma taxa de acerto em torno dos 96% para textos na língua inglesa.
É possível observar idéias lingüísticas semelhantes em todos os paradigmas de etiquetadores morfológicos treináveis citados. Todos utilizam-se de três fontes de informação lingüística, extraídas de um corpus de treinamento:

- uma lista de palavras associadas à categorias morfológicas (léxico), para fornecer informações sobre palavras conhecidas;
- os sufixos de palavras, como parte do processo de inferência da etiqueta morfológica de palavras desconhecidas;
- contexto próximo do item lexical que se quer etiquetar (2 ou 3 etiquetas ao redor), para refinar a escolha de sua etiqueta.

Autômatos Adaptativos

Os autômatos adaptativos (AA) constituem um formalismo para a representação de linguagens dependentes de contexto [6]. A base estrutural de um AA é um autômato de pilha; o que os diferencia é que um AA pode ter, associado a cada uma de suas transições, funções adaptativas.
As funções adaptativas são constituídas de um conjunto de **ações adaptativas elementares** que possibilitam modificar o autômato como decorrência da execução de uma transição, através do acréscimo e retirada de estados e transições. As **ações adaptativas elementares** podem ser de três tipos: Inspeção, Eliminação e Inserção. São estes dois últimos tipos de ações adaptativas elementares

que dão aos AA o poder computacional para manipular linguagens dependentes de contexto [6].
As **chamadas de funções adaptativas** podem ser de dois tipos: anterior (efetuada sempre antes de uma transição ocorrer) e posterior (efetuada sempre depois que a mudança de estado é realizada).
A característica de poder alterar sua própria topologia, peculiar aos autômatos adaptativos, faz com que eles sejam bastante adequados à modelagem de sistemas de aprendizado automático: um conjunto de exemplos poderia ser inserido em um AA (treinamento) na forma de novas transições; deste modo um AA pode incorporar conhecimento.

2. PROPOSTA

Este trabalho, ainda em andamento, propõe um método para a construção de um etiquetador morfológico, que possa ser usado para um número muito grande de línguas (apesar de ser testado apenas para a língua portuguesa), que apresente uma arquitetura híbrida: o etiquetador conterá um conjunto de regras escritas à mão, que cuidam de um grande número de casos, e permitirá que outras regras sejam aprendidas, sendo assim, treinável a partir de exemplos (textos manualmente anotados ou corpora etiquetados).
Estas duas parcelas da arquitetura cooperam para que seja atingida uma boa precisão no processo de anotação morfológica: as regras pré-definidas responsabilizam-se por abreviar o processo de treinamento, visto que aquilo que já é estabelecido como certo não precisa ser aprendido, além de ter-se a certeza de que não é aprendido um falso padrão; a parcela treinável, contudo, tem a tarefa de aprender todos os padrões que não forem contemplados pelas regras pré-definidas. Em uma situação limite, na qual nenhuma regra escrita à mão exista, esta parte do etiquetador aprenderá todas as regras a partir dos exemplos presentes em corpora lingüísticos anotados.
Os pré-requisitos para que este etiquetador possa ser treinado para uma determinada língua são os seguintes:

- disponibilidade de corpus com anotações morfológicas;
- as palavras desta língua devem ter uma estrutura do tipo:

  Palavra = Prefixos + Radical + Sufixos
- a língua não deve ser aglutinante (como é o caso do alemão, por exemplo), visto que este método ainda não leva em conta a existência de itens lexicais com mais de um radical.

A arquitetura básica do etiquetador morfológico treinável proposto neste trabalho segue o publicado por E. Brill, que divide-se em três módulos: o primeiro, que cuida da etiquetação inicial de palavras conhecidas, o segundo, que cuida da etiquetação inicial de palavras desconhecidas, e um terceiro e último, que promove um refinamento contextual [2].
Cada um dos módulos armazenará informações extraídas de um corpus de treinamento, e, a partir destas informações, procederá a etiquetação sem a inferência de regras explícitas, o que está mais próximo da proposta de W. Daelemans [4].
A forma de se codificar as regras pré-definidas é um tanto quanto simples: basta construir-se um corpus artificial com exemplos selecionados, o que guiará o processo de aprendizado automático para um estado inicial desejável, o qual contém um primeiro núcleo de regras conhecidas.
Autômatos adaptativos (AA) serão usados como base de implementação e como estrutura de dados para o armazenamento das informações necessárias a cada módulo [6].

Primeiro módulo: obtenção da etiqueta mais provável para as palavras conhecidas

A estrutura de dados concebida como base para este módulo é a de uma árvore *n*-ária de letras, utilizada para armazenar o léxico, contendo uma lista ligada associada a cada uma de suas folhas (cada folha representa o final de uma seqüência completa que compõe um item lexical). Esta lista é utilizada para armazenar as várias etiquetas morfológicas possíveis, em ordem decrescente de freqüência de aparecimento. Uma vantagem, inerente a esta estrutura em forma de árvore, é que ocorre naturalmente uma compressão do tamanho da base de dados pelo fato de todos os prefixos serem armazenados apenas uma vez na estrutura. A rigor, considerando-se todas as transições deste autômato, inclusive aquelas que processam os separadores do texto de entrada, ele constitue um grafo orientado cíclico.
Após a fase de treinamento, um texto sem anotações pode ser fornecido ao autômato e este determinará as etiquetas mais prováveis, em ordem decrescente de freqüência, para cada palavra que tenha aparecido no corpus de treinamento (ou seja, uma palavra conhecida); caso uma palavra deste texto não tenha aparecido neste corpus (palavra desconhecida), ela não receberá etiqueta alguma.

Segundo módulo: etiqueta para palavras desconhecidas, com base em sufixos

Com base nas últimas letras dos itens lexicais encontrados no corpus de treinamento e nas etiquetas morfológicas associadas a cada um deles, este módulo infere um mapeamento que é usado na etiquetação de itens lexicais que nunca apareceram no corpus de treinamento (palavras desconhecidas).
A heurística por trás deste módulo tem um embasamento lingüístico: é sabido que, nas línguas cujas palavras apresentam a estrutura PREFIXO + RADICAL + SUFIXO, o sufixo de uma palavra tem uma forte correlação com a sua categoria morfológica.
Em princípio, deve-se fazer um pré-processamento no corpus de treinamento, para que o mesmo seja reduzido a apenas terminações de palavras com as respectivas etiquetas. Optou-se por reduzir cada item lexical a apenas suas três últimas letras; é verdade que na língua Portuguesa há sufixos menores e maiores que 3 letras,

contudo, a escolha deste número é arbitrária e este valor pode ser facilmente alterado.
Deve-se também levar em conta que este módulo propricia uma forma de **extrapolação** quando faz uma comparação de sufixos, ou seja, qualquer comparação a ser feita não necessita ser exata, podendo ser parcial. Assim, uma palavra que tenha o sufixo "mente" receberia a etiqueta ADV (advérbio), mesmo que o módulo de palavras desconhecidas tenha associado apenas as últimas três letras "nte" a esta etiqueta.
Para ilustrar o processo de treinamento, pode-se observar o pequeno corpus hipotético, mostrado abaixo, que é, de fato, um subconjunto do corpus real. Nota-se que este já foi pré-processado.

```
ava/D-BV     ava/D-TE     aso/F-JDA
```

Este corpus traduz os seguintes fatos: que o sufixo "ava" (talvez advindo de palavras como "trabalhava", "estava", por exemplo) pode estar associado às etiquetas "VB-D" e "ET-D", e que o sufixo "osa" (por exemplo, das palavras "carinhosa" e "custosa") pode estar associado à etiqueta "ADJ-F".
Para que o autômato do segundo módulo possa ser usado no processo de etiquetação, é proposto um conjunto de transformações neste que podarão as transições ligadas ao seu crescimento (processo de aprendizagem) e acrescentarão novas. Estas novas transições cuidarão de encontrar a etiqueta mais provável para o respectivo sufixo.

Terceiro módulo: refinador contextual

Os dois módulos anteriores cuidam apenas de informações meramente lexicais extraídas de um corpus. Já o terceiro módulo serve como um refinador do serviço que os dois primeiros prestam. Ele é responsável por escolher, dentre as várias etiquetas possíveis para uma dada palavra, aquela que mais se adapte ao contexto em que esta palavra se encontra.
Este módulo é treinado com base em informações referentes à seqüência relativa em que as anotações morfológicas se acham no corpus de treinamento.
O corpus hipotético abaixo (somente formado por etiquetas morfológicas) ilustrará a heurística de treinamento adotada.

```
P  SR  ADV-R  CONJ  ADV-R   ....
P  N  VB-D  ....   P  N  SR
```

A idéia central do método baseia-se na utilização de uma janela de três posições. Percorre-se com ela a seqüência de etiquetas previamente extraída do corpus de treinamento; a primeira posição da janela refere-se a uma etiqueta já consumida anteriormente, mas que é memorizada; a segunda, refere-se à etiqueta que está sendo consumida na ocasião, e a terceira, à etiqueta seguinte, que é apenas consultada, sem ser consumida (um *look-ahead*).
Esta janela desloca-se um passo por vez, sendo que a cada passo da janela, o correspondente trigrama é considerado.
Durante o primeiro passo, são encontradas as etiquetas P SR ADV-R. As duas primeiras etiquetas foram consumidas, enquanto que a terceira foi apenas examinada. Todos estes dados são mantidos em uma estrutura de dados similar a uma árvore.
A heurística de aplicação do conhecimento contextual adquirido parte da suposição de que o processo se inicia a partir de uma etiqueta não ambígua; a etiqueta seguinte, a qual será chamada de Foco, é a que será refinada, tendo em vista as etiquetas anterior e posterior (esta última pode ser ambígua ou não). Portanto, será escolhida uma dentre as várias etiquetas possíveis, de acordo com o contexto, para substituir o Foco.
Ilustrativamente, caso este módulo se defronte com a etiqueta ambígua ADJ/N/SR, sendo a etiqueta anterior (já decidida) P e a próxima VB-D/ADV (também ambígua), percebe-se que existem 6 (1 vezes 3 vezes 2) possibilidades de trigramas sem ambigüidades, conforme listado a seguir:

P	ADJ	VB-D
P	ADJ	ADV
P	**N**	**VB-D**
P	N	ADV
P	SR	VB-D
P	SR	ADV

Suponha-se que apenas o trigrama ressaltado acima (P N VB-D) apareça no corpus de treinamento. Seria, então, natural esperar que o módulo de refinação contextual optasse pela etiqueta N para substituir a etiqueta ambígua ADJ/N/SR.
Se o contexto não oferecer informações suficientes (talvez por causa do treinamento com um corpus de tamanho pequeno, pouco significativo), opta-se por utilizar as informações lexicais vindas dos módulos anteriores (o de palavras conhecidas e o de palavras desconhecidas), considerando-se apenas a etiqueta mais provável.

3. RESULTADOS INICIAIS

Os experimentos que serão relatados não continham regras pré-codificadas, tendo contado apenas com padrões aprendidos dos corpora de treinamento. O primeiro deles indicou um desempenho baixo, no que tange à taxa de acertos; contudo, vale ressaltar que o tamanho do corpus de treinamento usado foi muitíssimo pequeno, praticamente insignificante: 1.684 palavras anotadas.
Mesmo assim, obteve-se uma taxa de acerto de 81,25%, comparável ao trabalho de C. Alves, que atingiu a taxa de 78,28%, com um corpus de treinamento de 5.000 palavras, e ao trabalho de A. Villavicencio, que alcançou 84,5%, com um corpus de treinamento de 14.000 palavras [1 e 11].
Como já era esperado, o aumento do corpus de treinamento propiciou um considerável aumento nesta

taxa de acerto. O experimento seguinte, feito com um corpus de treinamento de 51.017 palavras, atingiu uma taxa de acerto próxima aos 90%, o que é razoável: E. Brill começou a fazer experimentos que produziram resultados práticos com um corpus de 45.000 palavras; W. Daelemans e outros pesquisadores argumentam que o método baseado em exemplos memorizados começa a produzir resultados satisfatórios a partir de um corpus com 300.000 palavras [2 e 4].

4. CONCLUSÕES

Sem dúvida nenhuma, este trabalho constitue uma constatação significativa da adequação dos AA para a representação e manipulação de conhecimento da área de PLN (processamento de linguagens naturais). Mostrou sua viabilidade especialmente para a modelagem de algoritmos de aprendizado automático. Também deve ser ressaltado que o etiquetador para a língua portuguesa gerado é a primeira aplicação prática de larga escala baseada nos AA, mostrando que estes são dispositivos simples, elegantes, eficientes e treináveis.

Sob o olhar da lingüística computacional ou do PLN, foi construída uma ferramenta que propricia a análise morfológica de textos livres, com taxa de acerto comparável às dos paradigmas que representam o estado-da-arte na área, e com algumas vantagens, como:

- A complexidade computacional do treinamento e da aplicação dos três módulos que formam o etiquetador morfológico é independente do número de etiquetas e linear com respeito à cadeia de entrada. Isto é uma grande vantagem deste método proposto, em relação ao de E. Brill (cuja fase de treinamento tem dependência polinomial com relação à quantidade de etiquetas).
- A possibilidade de explicar ou justificar uma decisão tomada, com base na maior proximidade de um determinado exemplo memorizado.
- Como o autômato adaptativo presente no etiquetador comporta-se de modo praticamente igual a um autômato finito, depois da fase de treinamento (já que as ações adaptativas nesta fase não são tão usadas), o desempenho da implementação pode estar muito perto do melhor possível, que seria o de um autômato finito [9]. Uma exceção a esta afirmação é o módulo responsável pelo refinamento contextual, o qual, mesmo na fase de aplicação, conta com o uso de funções adaptativas na montagem e desmontagem de estruturas auxiliares de transições. Contudo, se o objetivo for a obtenção de alta velocidade de etiquetagem e se houver um bom gerenciamento de memória por parte da aplicação e/ou do sistema operacional, de modo que não haja limitação quanto ao número de transições, a heurística do refinamento contextual poderia ser otimizada de modo que não sejam removidas as estruturas criadas para a resolução de ambigüidade.

5. REFERÊNCIAS

[1] ALVES, C.; FINGER, M. Etiquetagem do Português Clássico Baseada em Córpora. In **Proceedings of IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR99)**, Évora, Portugal, 21-22 September 1999.

[2] BRILL, E. **A corpus-based approach to language learning**. Thesis (PhD) - Department of Computer and Information Science of the University of Pennsylvania, Philadelphia, 1993, 154 p.

[3] CHARNIAK, E. **Statistical language learning**. MIT Press, 1993.

[4] DAELEMANS, W.; ZAVREL, J.; BERCK, P.; GILLIS, S. MBT: A memory-based part of speech tagger-generator. In **Proceedings WVLC**, 1996. Copenhagen.

[5] ISABELLE, P.; BOURBEAU, L. TAUM-AVIATION: its technical features and some experimental results. **Computational Linguistics**, v.11, n.1, p.18-27, 1985.

[6] JOSÉ NETO, J. Adaptative automata for context-dependent languages. **ACM SIGPLAN Notices**, v.29, n.9, p.115-24, Sept. 1994.

[7] KOSKENNIEMI, K. Representations and finite-state components in natural language, p.99-116. In ROCHE, E.; SCHABES, Y. (Eds.) – **Finite-state language processing**. MIT Press, 1997.

[8] MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. Building a large annotaded corpus of English: the Penn Treebank. **Computational Linguistics**, v. 19, n.2, p.313-30, 1993.

[9] ROCHE, E.; SCHABES, Y. Deterministic part-of-speech tagging with finite-state transducers, p.205-39. In ROCHE, E.; SCHABES, Y. (Eds.) – **Finite-state language processing**. MIT Press, 1997.

[10] TBCHP Tycho Brahe Parsed Corpus of Historical Portuguese. **Instituto de Estudos da Linguagem**, UNICAMP, SP, http://www.ime.usp.br/~tycho/corpus, 1998.

[11] VILLAVICENCIO, A.; MARQUES, N.; LOPES, G.; VILLAVICENCIO, F. Part-of-Speech Tagging for Portuguese Texts Introduction, p.323-32. In WAINER, J.; CARVALHO, A. (Eds.) – **Lecture Notes in Artificial Intelligence 991 – Advances in Artificial Intelligence – 12th Brazilian Symposium on Artificial Intelligence, SBIA'95**, Campinas, Brazil, Oct. 10-12, 1995 – Springer-Verlag.