Capítulo 9

O Cortex e a sua participação no HAREM

Christian Nunes Aranha

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 9, p. 113–122, 2007.

O Cortex é um sistema de inteligência artificial desenvolvido a partir de minha tese de doutorado na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Em minha tese desenvolvi o esboço teórico e implementei a primeira versão a qual participou do HAREM, e hoje já se encontra em sua versão 3.0.

O Cortex nasceu com a ambição de simular as faculdades cognitivas de PLN. Isto significa dizer que seu maior objetivo é a eficiente manipulação da linguagem humana, tanto na leitura, codificação e interpretação de textos como na produção. Acreditamos que se nos aproximarmos cada vez mais do processo cognitivo humano, teremos cada vez melhores resultados.

Nós, da Cortex, entendemos que a produção eficiente tem como pré-requisito uma boa leitura. Sendo assim, não trabalhamos com produção ainda (apenas de resumos). Da mesma forma, para uma boa leitura, é necessário um bom conhecimento das palavras, dos seus significados e da gramática de uma língua, em princípio nesta ordem. Logo, o Cortex é um processador dependente da língua, o que está alinhado com nossos objetivos finais, já que, nós, seres humanos também somos dependentes da língua, porém, com capacidade de aprender novas. Assim como deve ser o Cortex.

9.1 Filosofia

Em psicologia do desenvolvimento humano vemos que bebês/crianças manifestam espontaneamente a capacidade de adquirir (e não aprender) a linguagem sozinhas, simplesmente ao ouvir frases e pequenos textos falados provenientes em grande parte de seus pais. Mais tarde, utilizando essa linguagem "adquirida", irão então, não adquirir, mas "aprender" (por exemplo, na escola) a língua escrita. Aprender porque precisam de um professor para ensinar. Seres humanos não costumam ter a capacidade espontânea de ler e escrever.

Adicionalmente, parece que a explicação natural para a ordem do áudio-visual, ou seja assimilar primeiro o som e só depois a imagem, está contida no domínio biológico já que existe uma conversão quase que direta entre uma mensagem falada e uma escrita. Isso nos leva a crer que, se existe um processo para adquirir a fala, há de haver um para adquirir textos também.

Inspirado nestas observações empíricas, o sistema Cortex surge, então, para responder à seguinte pergunta: Que "programa" haveriam estes bebês de processar para adquirir a linguagem através do som? E mais, que programa seria rodado, para com isso adquirir novas palavras?

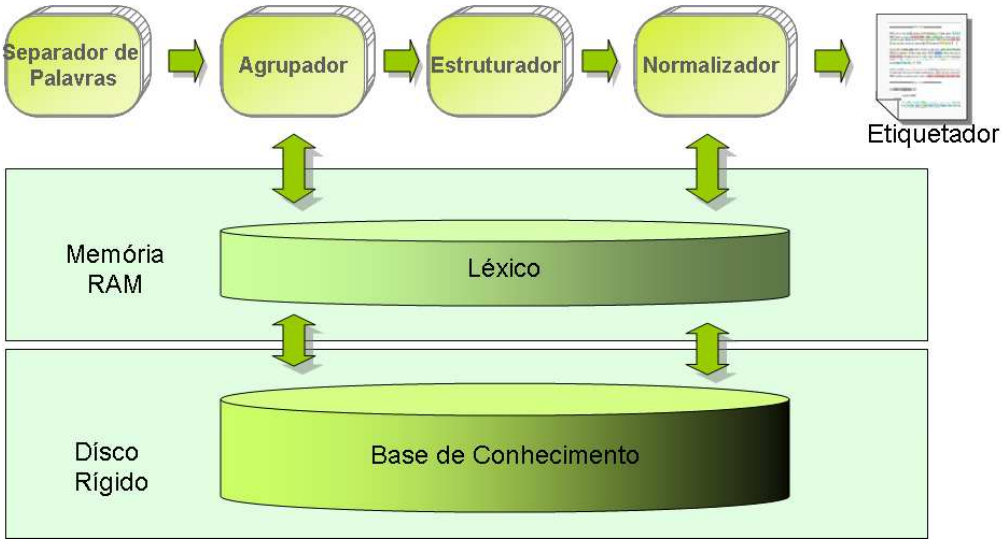

Figura 9.1: Etapas de processamento do texto no Cortex.

9.2 Classificação de entidades mencionadas no Cortex

O Cortex é um sistema computacional para o processamento da língua, cujo algoritmo reproduz alguns comportamentos lingüísticos de um falante, como sua adaptabilidade, flexibilidade, e capacidade de antecipar, pressupor e rever suas hipóteses.

Dessa maneira, o processamento do Cortex é feito em várias etapas, como mostra a Figura 9.1. Cada etapa é capaz de rever os passos anteriores e influir sobre os subseqüentes. Após a separação inicial das palavras, a etapa seguinte consiste em reconhecer as entidades que possam ser constituídas por mais de uma palavra. Substantivos compostos e locuções são descobertos nesse momento. O processo de reconhecimento dos termos é feito com o auxílio de um autômato escrito para identificar padrões de formação de entidades compostas com base num repertório de regras. O resultado dessa etapa é adicionada ao conhecimento existente no léxico, e posteriormente à base de dados.

O próximo passo constitui na classificação dos termos previamente extraídos. Sabendo-se que a criatividade lingüística é de suma importância na produção textual, o Cortex recorre a um banco de informações lexicais com certa parcimônia. As informações armazenadas sobre uma palavra (sua classe, significado, etc.) são tomadas apenas como um dado *a priori*, que pode ser questionado e reavaliado por outras circunstâncias a que esta palavra se vê envolvida no texto. O resultado disso é que o Cortex se torna um mecanismo provido de experiência, ou seja, quanto mais texto processa mais conhecimento lingüístico ele acumula e mais poder de inferência ganha para processar novas informações/textos.

Além disso, o Cortex obtém as informações de quatro fontes de dados, como mostra a

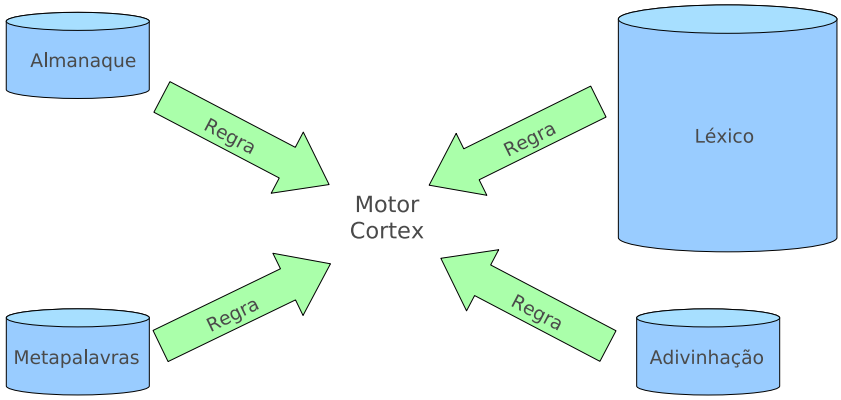

Figura 9.2: Fontes de dados do Cortex.

Figura 9.2: o Almanaque, que contém uma lista de entidades de uma determinada categoria provenientes de uma fonte enciclopédica; o Metapalavras, constituído por uma lista de termos que aparecem nas vizinhanças das entidades, por exemplo, *pianista*, *jogador*; a Adivinhação, que contém um conjunto de termos que constituem as entidades mencionadas, por exemplo, *Prof.*, *Dr.*, *Presidente*; e o Léxico, que armazena todo o conhecimento aprendido pelos textos já processados pelo Cortex.

Cada uma das fontes infuencia a tomada de decisão do Cortex quanto à identificação e classificação de EM. Cada regra traz consigo uma probabilidade associada, que é usada pelo Motor Cortex. Em paralelo a esse sistema existem máquinas de estimação de novas regras e probabilidades. Exemplos de aplicação das quatro fontes de dados são:

Categoria Pessoa

```
Entrada: O acordeonista Miguel Sá(...)
Saída:   O acordeonista <PESSOA TIPO="INDIVIDUAL">Miguel Sá</PESSOA>(...)
```

onde *acordeonista* é um termo obtido da fonte de dados Metapalavras, associado à pessoa *Miguel Sá*.

```
Entrada: Na pesquisa do Dr. Lewis(...)
Saída:   Na pesquisa do <PESSOA TIPO="INDIVIDUAL">Dr. Lewis</PESSOA>(...)
```

onde *Dr.* é uma evidência obtida através da lista Adivinhação que indica probabilidade para nome de pessoa. No modelo original do Cortex, *Dr.* não faz parte da EM. A entidade final é *Vernard Lewis* obtida pela regra de co-referência. Especialmente para o HAREM adicionamos um novo conjunto de regras que juntava `TITLE + NOME` e produzia a etiqueta SGML final.

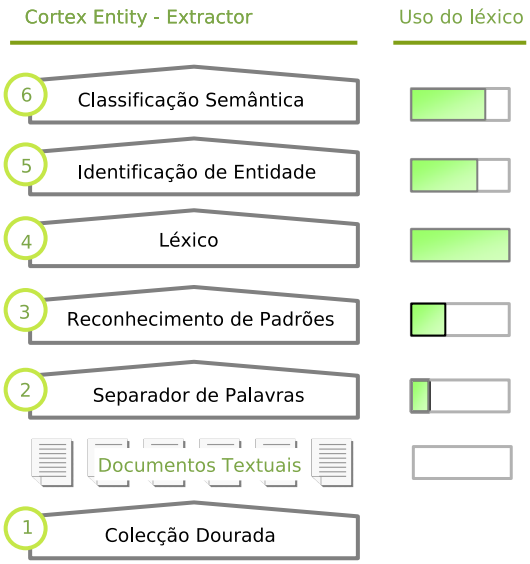

Figura 9.3: Etapas de avaliação de documentos no Cortex.

Categoria Local

```
Entrada: (...)Pela primeira vez no Haiti um padre foi assassinado
         por motivos políticos(...)
Saída:   (...)Pela primeira vez no <LOCAL TIPO="ADMINISTRATIVO">Haiti</LOCAL>
         um padre foi assassinado por motivos políticos(...)
```

onde *Haiti* pode ser primeiramente aprendido pela fonte Almanaque e depois passa para a fonte Léxico.

A Figura 9.3 apresenta todas as etapas as quais os documentos são submetidos ao Cortex, em particular o corpus Coleção HAREM, para se obter sua classificação. Na coluna à direita da figura é apresentado o percentual de uso do Léxico nas diferentes etapas.

O Cortex é composto pelo Separador de Palavras, que identifica cada termo (simples ou composto) como uma palavra; Reconhecimento de Padrões, que reconhece categorias ou classes de termos; o Léxico, que armazena as informações lingüísticas de cada termo; Identificador de Entidades, que identifica os limítes de cada entidade mencionada; o Classificador de Entidades, que finaliza o processo de reconhecimento da entidade atribuindo a ela um rótulo semântico dentro de uma ontologia pré-definida, gerando uma etiqueta SGML correspondente como formato de saída.

Medida	Cenário: TOTAL 1º Lugar	Cenário: SELECTIVO 1º Lugar	Cenário: TOTAL Resultados Cortex
Precisão	ELLE (80,64%)	PALAVRAS (78,50%)	CORTEX_NO (65,57%)
Abrangência	SIEMES1 (84,35%)	SIEMES1 (84,35%)	CORTEX_NO (86,69%)
Medida F	PALAVRAS (0,8061)	PALAVRAS (0,8061)	CORTEX_NO (0,7466)

Tabela 9.1: Vencedores da tarefa de identificação do HAREM (considerando apenas saídas oficiais), e resultados da saída não-oficial do Cortex.

9.3 A participação do Cortex no HAREM

O Cortex foi submetido à avaliação do HAREM nas seguintes tarefas e categorias:

- Tarefas efetuadas: identificação e classificação semântica de EM.
- Cenário seletivo: `PESSOA`, `ORGANIZACAO`, `LOCAL`, `TEMPO`, `ACONTECIMENTO` e `VALOR`.

O principal erro cometido foi, conjuntamente, a baixa flexibilidade do formato de saída de nosso sistema e a má interpretação das regras do HAREM. Não tinhamos muito tempo, começamos a estudar e trabalhar na avaliação poucos dias antes. Foi quando nos deparamos com a diferença entre a saída de nosso sistema e o formato padrão do HAREM.

O Cortex se aproximava da versão 1.0 e não tinha flexibilidade nenhuma de configuração das etiquetas de saída. A solução foi improvisar uma transformação do arquivo através de uma substituição manual, o que ocupava um tempo bastante grande. O Cortex imprimia a saída como `PESSOA`, se a entidade fosse classificada como pessoa, `GEOGRAFIA`, se a entidade fosse `LOCAL`, e `ORGANIZAÇÃO` idem, mas se não conseguiu classificar imprimia apenas `NOME`. Achávamos que só poderiamos concorrer nas tarefas de identificação e classificação semântica, e `NOME` não existia nas directivas do HAREM, sendo assim, optamos por retirar as entidades com marcação `NOME` e não marcar nada. No dia seguinte, lendo as regras com mais calma descobrimos a existência da etiqueta `<EM>`. Fizemos tudo novamente e entramos na avaliação não-oficial.

O prejuízo no resultado oficial foi grande porque nosso sistema de identificação estava razoável para a época, porém, nosso sistema de classificação tinha uma abrangência muito fraca e eliminou várias entidades que poderiam ter sido identificadas. Enfim, fazendo as contas considerando nosso resultado não-oficial, não ficaríamos em primeiro lugar total da medida F por outros problemas que explicarei a seguir, mas pelo menos ganharíamos o primeiro lugar em termos de abrangência no cenário seletivo, com 86,69% (acima de 84,35%, como mostra a Tabela 9.1).

Quanto ao desempenho por Género, apenas nos textos *correio eletrônico* teríamos obtido primeiro lugar na medida F. Em média, teríamos ficado em quarto lugar geral com nossa saída não-oficial.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: TOTAL Forma: RELATIVO 1º Lugar	Cenário: SELECTIVO Forma: ABSOLUTO 1º Lugar	Cenário: SELECTIVO Forma: RELATIVO 1º Lugar
Categorias	PALAVRAS (0,6301)	CORTEX2 (0,7171)	PALAVRAS (0,6301)	CAGE3 (0,8161)
Tipos	-	ELLE (0,8497)	-	NOOJ1 (0,8917)
Combinada	PALAVRAS (0,5829)	ELLE (0,6812)	PALAVRAS (0,5829)	ELLE (0,7327)
Plana	PALAVRAS (0,5293)	ELLE (0,6548)	ELLE (0,5487)	ELLE (0,7044)

Tabela 9.2: Vencedores para tarefa de classificação semântica do HAREM.

O resultado para a classificação semântica (Tabela 9.2) nos mostrou que a classificação tinha uma boa precisão, obtendo o primeiro lugar no cenário total relativo. Os outros problemas de padronização da saída que tivemos foi com relação aos números por extenso que não apresentam letra maiúscula são marcados como entidade do tipo valor pelo Cortex e não pelo HAREM, assim como as referência a tempo (por exemplo, *ontem* e *segunda-feira*). Em contrapartida perdemos muitos pontos pela identificação de *R:* nos textos de gênero *entrevista* que foi marcado porque tinha letra maiúscula, e de fato não faz sentido ser entidade. Finalmente, a titulação das pessoas como por exemplo *Sr.*, *Dom* ou *Dr.* são excluídas da entidade pessoa pelo Cortex, já que esses lexemas são classificados como metapalavras e não fazem parte da entidade, uma mera questão de configuração de saída, e foram consideradas pelo HAREM como parte da pessoa. Veja o exemplo:

```
HAREM:  Na pesquisa do <PESSOA TIPO="INDIVIDUAL">Dr. Lewis</PESSOA>(...)
CORTEX: Na pesquisa do Dr. <PESSOA TIPO="INDIVIDUAL">Lewis</PESSOA>(...)
```

Conclusão, o sistema como estava implementado, sem flexibilidade de configuração, seria impossível fazer essas modificações para o HAREM. Sendo assim, deu-se início ao trabalho do refatoramento para construir a versão 2.0.

9.4 A participação do Cortex no Mini-HAREM

A participação do Cortex no Mini-HAREM contou com a versão 2.0 de nosso sistema, onde havia principalmente flexibilidade de configuração para adequar a saida aos padrões do HAREM. Com isso conseguimos reduzir enormemente os erros de sobre-geração que tanto nos penalizou na primeira edição.

Para implementar a segunda versão e as seguintes foi necessário, não só o refatoramento da primeira versão, como o apoio de mais três membros.

Além disso, a versão 2.0 contava com um sistema de classificação bem mais evoluído, com mais estratégias cognitivas e também mais conhecimento lexical, dado que o sistema Cortex acumula o conhecimento a cada documento novo lido.

O Cortex foi então submetido à avaliação do Mini-HAREM nas seguintes tarefas e categorias:

Medida	TOTAL 1º Lugar	SELECTIVO 1º Lugar
Precisão	Cortex2CEM (87,33%)	Cortex2CEM (83,87%)
Abrangência	Cortex1REM (87,00%)	Cortex1REM (88,93%)
Medida F	Cortex1REM (0,8323)	Cortex1REM (0,7662)

Tabela 9.3: Vencedores da tarefa de identificação no Mini-HAREM.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
Categorias	Cortex2CEM (0,6157)	Cortex2CEM (0,6839)
Tipos	-	-
Combinada	Cortex2CEM (0,5855)	Cortex2CEM (0,6501)
Plana	Cortex2CEM (0.5525)	Cortex2CEM (0,6145)

Tabela 9.4: Vencedores da tarefa de classificação semântica no Mini-HAREM.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
HAREM	PALAVRAS (0,8061)	PALAVRAS (0,8061)
Mini-HAREM	Cortex1REM (0,8323)	Cortex1REM (0,7662)

Tabela 9.5: Comparação dos resultados HAREM e do Mini-HAREM para a tarefa de identificação.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
HAREM	PALAVRAS (0,6301)	PALAVRAS (0,6301)
Mini-HAREM	Cortex2CEM (0,6157)	Cortex2CEM (0,6839)

Tabela 9.6: Comparação dos resultados HAREM e do Mini-HAREM para a tarefa de classificação semântica, medida por categorias.

- Tarefas efetuadas: identificação e classificação semântica de EM.
- Cenário seletivo: `PESSOA`, `ORGANIZACAO`, `LOCAL`, `TEMPO`, `ACONTECIMENTO` e `VALOR`.

E obteve os resultados mostrados pelas Tabelas 9.3 e 9.4 para as avaliações de identificação e classificação respectivamente das quais participou.

Comparando os resultados do Mini-HAREM e os do HAREM, podemos fazer um *ranking* total, com todos os participantes (embora esta seja uma comparação bastante artificial,

Gênero	Precisão	Abrangência	Medida F
web	76,26%	81,97%	0,7901
correio eletrônico	64,80%	81,50%	0,7220
literário	79,29%	87,12%	0,8302
político	90,83%	90,83%	0,9083
expositivo	90,76%	91,59%	0,9117
técnico	38,81%	69,67%	0,4985
entrevista	93,40%	93,79%	0,9359
jornalístico	90,52%	94,24%	0,9234

Tabela 9.7: Comparativo dos resultados do Cortex segmentado por gênero.

Saída	Precisão (%)	Abrangência (%)	Medida F	Erro Combinado	Sobre-geração	Sub-geração
cortex3	57,12	73,54	0,6430	0,4969	0,3492	0,1743
cortex2cem	57,12	73,54	0,6430	0,4969	0,3492	0,1743

Tabela 9.8: Resultado para categoria `QUANTIDADE`.

porque compara desempenho sobre textos diferentes, de diferentes versões dos mesmos sistemas). Mas admitindo que essa comparação é válida, os resultados das Tabelas 9.5 e 9.6 mostram que o sistema Cortex obteve o primeiro lugar no cenário total absoluto para a tarefa de identificação, e o primeiro lugar no cenário selectivo absoluto para a tarefa de classificação semântica.

Nessa seção analisaremos os pontos críticos apontados pelos relatórios disponibilizados pela Linguateca. Esses serão os pontos de melhora para as próximas versões na intenção de aumentar a medida F.

O primeiro ponto crítico que vale a pena ressaltar foi o desempenho do Cortex no gênero *técnico*. A Tabela 9.7 mostra como o desempenho foi bem inferior aos demais.

Isso se deu em grande parte pelo reconhecimento dos subtítulos como entidades. Além de nomes de teorias e pessoas que acabaram dificultando a tarefa.

O segundo ponto crítico foi o desempenho semântico do Cortex na categoria `VALOR`, mostrado na Tabela 9.8. Analisando o arquivo de alinhamento, descobrimos que o Cortex considera *80 anos* (por exemplo) como `TEMPO` e não como `VALOR TIPO="QUANTIDADE"`, o que ocasionou uma baixa significativa na medida F.

Além desses pontos, vale destacar que o Cortex é treinado na língua portuguesa do Brasil e portanto, diversos verbos diferentes foram encontrados no início de frase, provocando uma confusão com uma entidade desconhecida.

Finalmente, cargos em letra maiúscula também foram descartados e serão configurados como `GRUPOCARGO` a partir de agora e números referentes a artigos que foram considerados como número e irão pra categoria `OBRA` para a próxima edição do HAREM.

9.5 Cortex 3.0

Os últimos resultados levam-nos a pensar que a utilização de almanaques é bastante interessante e útil no início do aprendizado do sistema, porém, conforme ele vai adquirindo inteligência gramatical, a utilização destes descresce bastante, e algumas vezes, acaba por prejudicar a precisão do sistema.

Por esse motivo, o foco do sistema Cortex é cada vez mais em cima das informações presentes no texto, ontologias e conhecimento enciclopédico. Procuramos atualmente um modelo de representação para o conhecimento abstrato extraído dos textos e que seje o mais interpretável possível de modo a aumentar o poder de gerenciamento do conhecimento acumulado.

9.6 Conclusões

Este capítulo descreve o sistema Cortex, um sistema baseado em inteligência artificial para o aprendizado, aquisição, reconhecimento e classificação de, não só entidades como também verbos, substantivos e adjetivos. Para as duas primeiras edições do HAREM, trabalhamos principalmente com em textos na língua portuguesa do Brasil.

O sistema foi projetado para integração com mecanismos de indexação, o que o torna completamente escalável para mineração de textos em grandes quantidades de documentos. A abordagem aqui descrita faz parte de um projeto maior de estruturação de dados não-estruturados. Isso significa extrair um modelo de representação semântico para ser usado em domínios como a Web Semântica. Esse mesmo sistema é usado na plataforma de inteligência competitiva da empresa Cortex Intelligence[1].

Para o HAREM foram feitas algumas adaptações ao sistema para atender a especificação da ontologia da avaliação, que difere em parte da utilizada por nós. Mesmo criando um módulo mais sofisticado de configuração da ontologia para o Mini-HAREM, vimos que ainda cometemos erros de transdução.

Os relatórios produzidos pela Linguateca ajudaram em muito o aperfeiçoamento de nosso sistema. Apontando detalhes que nos passavam desapercebidos, mostrando novos domínios de informação a serem explorados, assim como um panorama mundial do tratamento da língua portuguesa. Além, é claro, na produção de um corpus de treinamento para as próximas edições.

Estamos em constante melhoramento de nosso sistema, ainda temos muito a caminhar, principalmente para outras línguas. Em futuras edições do HAREM, gostaríamos de ver avaliações de anáforas e fatos.

[1] www.cortex-intelligence.com