Processamento de Linguagem Natural - Uma Abordagem Conexionista

Natural Language Processing - A Connectionist Approach

Prof. João Luís Garcia Rosa[1]
Prof. Dr. Marcio Luiz de Andrade Netto[2]

Abstract

This paper shows an introduction to the neural networks approach of natural language processing where the advantages and characteristics of this approach are discussed and the tendencies for the future are shown. Natural language processing consists of a series of tasks the machine should execute to analyse a text. These tasks are mostly interdependent. Carrying out all these tasks in parallel would increase the performance of the model. Other considerations about his high connectionism model are also discussed.

Key-Words: *Natural language processing, neural network, connectionism, paralellism.*

Resumo

Este artigo mostra uma introdução à abordagem do processamento de linguagem natural, utilizando redes neurais. Discute as vantagens e características dessa abordagem e mostra tendências para o futuro. O processamento de linguagem natural consiste de uma série de tarefas que a máquina deve executar para analisar um texto. Estas tarefas são, na maioria das vezes, interdependentes. A realização das mesmas em paralelo seria muito útil para o desempenho do modelo. Outras considerações a respeito da utilização desse modelo de alto conexionismo são abordados.

Palavras-Chave: *Processamento de Linguagem Natural, Rede Neural, Conexionismo, Paralelismo.*

1. Introdução

O processamento de linguagem natural pode ser definido de formas diferentes. Todas as definições incorporam a noção de armazenamento em computador e manipulação de dados lingüísticos. Entretanto, o ponto de discussão é o grau de sofisticação envolvido; este traduz em uma porção de estruturas lingüísticas inerentes ao texto original, as quais o sistema pode detectar automaticamente, armazenar e manipular. De uma forma mais simples, o processamento de linguagem natural pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia-a-dia.

2. Abordagens do Processamento de Linguagem Natural

O problema central dos sistemas de processamento de

1 Professor do Instituto de Informática - PUCCAMP
2 Professor da Fac. de Eng. Elétrica - UNICAMP.

linguagem natural é a transformação de uma frase de entrada potencialmente ambígua em uma forma não ambígua que possa ser usada internamente por um sistema de computador. Estas representações internas variam, é claro, de uma aplicação para outra.

A transposição de uma frase potencialmente ambígua para uma representação interna é conhecida como "parsing" (análise). A palavra "parse" é derivada do latim: "pars orationis" (parte do discurso). No processamento de linguagem natural, parsing é usualmente um processo de combinar os símbolos de uma frase em um grupo que pode ser substituído por um outro símbolo mais geral. Este novo símbolo pode por sua vez ser combinado em um outro grupo, e assim por diante, até que uma estrutura permitida apareça.

Existem cinco tipos diferentes de parsers (analisadores); por casamento de padrões, baseado em gramática, semântico, baseado em conhecimento, e por redes neurais. Cada um tem uma abordagem única, sem igual, em processamento de linguagem natural. Vai-se abordar aqui os analisadores por redes neurais, também conhecido por analisadores conexionistas.

Abordagem por rede neural: Uma abordagem mais recente em processamento de linguagem natural envolve o estabelecimento de uma rede de unidades de computação parecidas com o neurônio. Cada unidade tem várias entradas, um conjunto pequeno de estados possíveis, e uma saída que é uma função das entradas. Cada entrada para a unidade de computação tem um valor de confiança, que pode variar de -1 a 1. Quando uma unidade é ativada, ela analisa todas as suas entradas e as pondera de acordo com seus respectivos valores de confiança. Se certas condições são encontradas, a unidade gera um valor de saída que é usado como entrada por outras unidades. Note que apenas os valores de confiança das entradas podem ser mudados durante o "aprendizado"; o padrão de conexão é estabelecido previamente.

Este tipo de sistema é usualmente chamado de abordagem por rede neural ou conexionista. A premissa fundamental desta abordagem é que as unidades individuais não transmitem grandes quantidades de dados, portanto ocorre a computação simplesmente por serem conectadas a um grande número de unidades similares.

O modelo de análise por rede neural contém três níveis de "neurônios". O primeiro nível é o nível léxico, que serve como nível de entrada da rede. Aqui, os neurônios são mapeados em determinadas palavras. No segundo nível, o nível do sentido da palavra, as entradas do nível léxico são combinadas para ativar neurônios que representam o significado das palavras. No terceiro nível, de lógica de caso, os significados são combinados para formar predicados e objetos.

A análise por rede neural se aproxima do modelo humano do processamento de informação lingüística, baseado na evidência neurológica.

3. O Cérebro como Modelo

A idéia de simular o cérebro formava a fundação para muitos trabalhos iniciais em Inteligência Artificial. O cérebro era visto como uma "rede neural", ou seja, um conjunto de nós, ou neurônios, conectados por linhas de comunicação. Atualmente tem havido um crescente interesse no uso de modelos de redes neurais, ou conexionistas, como o campo é chamado. Modelos conexionistas são aplicáveis num nível mais simples, o cérebro funciona da seguinte forma: neurônios ativam ou inibem o disparo de outros neurônios. Se um determinado neurônio dispara ou não depende das entradas inibitórias ou excitatórias de todos os neurônios conectados a ele. De alguma forma, as ativações de todos os neurônios que se comunicam entre si, e a interação do sistema nervoso com o ambiente determinam as lembranças e o pensamento.

O cérebro é relevante? Uma escola forte em Inteligência Artificial sente que estudar o cérebro não é a melhor forma de entender o pensamento. O cérebro representa apenas uma forma de se fazer uma máquina pensante. A Inteligência Artificial tradicional vê o pensamento como uma série de problemas para resolver, e acredita piamente que não há razão filosófica para um computador não poder resolvê-los. A base para esta crença é a tese de Church-Turing, que estabelece que se uma função é computável,

pode-se computá-la com um computador convencional - formalmente, uma máquina de Turing. Esta tese não pôde ser provada, mas é largamente aceita porque ninguém pôde pensar em um contra-exemplo.

Paralelismo: Uma outra razão para se estudar modelos parecidos com o cérebro é seu paralelismo. Os "circuitos" do cérebro são mais lentos do que os de um computador. Para que o cérebro trabalhe o mais rápido possível - os psicólogos mostraram que podemos reconhecer objetos num segundo - muitos neurônios devem trabalhar em paralelo. Em contraste, muitos programas de Inteligência Artificial "rodam" muito lentamente. A esperança é achar formas de "rodar" programas de Inteligência Artificial em paralelismo, para que atinjam resultados num tempo razoável.

A computação paralela tem sido bastante explorada em ciência da computação nos últimos dez anos. As redes neurais representam apenas uma linha de pesquisa em computação paralela. Basicamente, deve-se responder duas questões fundamentais no projeto de um sistema de computador paralelo: como conectar os processadores para propósito de comunicação e quanto de potência computacional e memória cada processador deve ter.

Os pesquisadores de redes neurais pensam que seus modelos, por serem os mais fiéis sobre o cérebro conhecido, mostrarão sucesso. Infelizmente, as redes neurais raramente têm sido construídas em hardware; normalmente elas são simuladas por software. Estas simulações são geralmente muito lentas, pois um processador tem que fazer o trabalho de muitos. Até que construamos hardware de processamento paralelo efetivo, os modelos conexionistas não alcançarão soluções eficientes para problemas de Inteligência Artificial.

4. Variedades de redes neurais

Muitos modelos de redes neurais devem alguma coisa aos percéptrons (redes neurais classificadoras) (LIP86), mas são mais gerais. O modelo típico de rede neural consiste de um conjunto de nós, ou neurônios e conexões. Cada nó tem um número real, que é a sua ativação. Cada conexão contém também um número real, seu peso. Estes números são usualmente positivos e usualmente têm um valor máximo. Algumas unidades são conectadas a entrada e saída. Os pesos representam a força de conexão entre dois neurônios.

Geralmente, a rede neural é um sistema dinâmico, movendo de um estado para o próximo. Como tal, ele tem uma regra matemática que rege esse movimento. Um número infinito de tais regras é possível. Entretanto, usualmente quer-se limitar os modelos a influenciar a ativação de um dado nó baseado apenas nas ativações dos nós conectados a ele e nos pesos das conexões a esses nós.

As redes neurais não são explicitamente programadas como um computador convencional. Por melhor dizer, elas obedecem leis, ou regras, como um sistema físico. Deve-se programar um computador convencional, mas uma rede neural simplesmente se conduz. Os projetistas de redes neurais vêem isto como uma vantagem, pois isto provê um mecanismo por meio do qual a inteligência pode surgir da lei física.

Uma das mais simples dessas regras é a regra linear. Computa-se a ativação de um dado nó como a soma dos produtos dos pesos de cada nó ao qual está conectado e a força dessa conexão. Essa regra é freqüentemente limitada: valores que passam de um certo limiar são cortados, para evitar os valores de ativação grandes. Existem muitas variantes das regras lineares.

Uma outra regra, sugerida por D. O. Hebb, reforça a conexão entre dois nós que são altamente ativados ao mesmo tempo. Algumas versões da regra de aprendizado Hebbiana permitem entradas, que ensinam, para influenciar a mudança de peso. Este tipo de regra é uma formalização da psicologia associacionista, que assegura que associações são acumuladas entre coisas que ocorrem juntas.

Aprendizado Competitivo: O aprendizado é, talvez, o fenômeno mais importante em psicologia. Os primeiros pesquisadores em redes neurais eram ansiosos para mostrar como as redes podiam aprender padrões de entrada apresentados a elas - ou seja, como elas podiam vir a perceber esses padrões, por elas mesmas.

Um dos métodos que vários pesquisadores têm planejado através dos anos é o aprendizado competitivo. Este método tem um primeiro nível, de unidades de entrada que contêm o padrão a ser entrado no sistema. O nível acima das unidades de entrada consiste de clusters de unidades. Cada unidade num cluster compete com as outras unidades no cluster pelo direito de reconhecer um padrão de entrada. Depois de um período de aprendizado, cada unidade num cluster reconhece um subconjunto dos padrões apresentados a ela. Portanto, cada cluster representa uma classificação, ou grupo, de padrões de entrada.

No aprendizado competitivo, cada unidade em cada cluster é conectado a todas as unidades de entrada. Os pesos das conexões são inicialmente colocados em valores aleatórios. Os pesos aleatórios fazem com que certas unidades nos clusters comecem a responder mais a determinados padrões de entrada, pois os pesos das conexões a essas unidades de entrada são mais fortes para alguns do que para outros.

No decorrer do aprendizado, os pesos mudam. Como determinadas unidades no cluster se tornam sensíveis a determinadas unidades no padrão de entrada, os pesos que conectam os pares associados de unidades aumentam, a custa de pares não associados de unidades. Unidades diferentes no mesmo cluster se inibem, de tal forma que apenas uma unidade num cluster "ganha" o direito de reconhecer um dado padrão.

Assim, com o tempo, unidades diferentes num cluster "reconhecem" propriedades diferentes de padrões de entrada. Por exemplo, um cluster de duas unidades pode separar todos os padrões de entrada naqueles que têm a maioria das suas unidades altamente ativadas e naqueles que estão na maioria desligados. Os clusters maiores fariam mais classificações discriminatórias.

Máquinas de Boltzmann: Uma importante classe de redes neurais simula o comportamento de sistemas físicos. Os sistemas físicos têm uma tendência de se moverem para dentro de estados de energia potencial mínima. Um exemplo simples disto é uma bola rolando num vale entre duas colinas. No alto da colina, a energia potencial é alta; no vale, é baixa.

Este processo é chamado de relaxação. John Hopfield mostrou que uma certa regra evolucionária simples para uma rede neural levará à relaxação. Sistemas como os de Hopfield, que remontam aos sistemas termodinâmicos, como os átomos em uma sala, são chamados de máquinas de Boltzmann. As máquinas de Boltzmann são muito usadas em várias aplicações de redes neurais.

Representações distribuídas : Uma importante característica de muitos modelos de redes neurais é sua natureza distribuída. Uma rede semântica padrão, como aquelas usadas nos primeiros esquemas de representação do conhecimento, consiste de um conjunto de nós conectados de alguma forma. Cada nó representa uma única palavra ou conceito. Se a rede estiver "pensando" na palavra "gato ", o nó para "gato" é ativado, e todos os outros nós não. Esta é uma representação local.

Em contraste, numa rede distribuída, os nós não têm um único significado; ou seja, um conceito individual é representado por um padrão por todos os nós. Por exemplo (Zeidenberg [Zei87]), se há dez nós, ativando os nós 1, 3, 4 e 6 pode-se representar o conceito '"gorila" enquanto que ativando os nós 2, 4, 5 e 7 pode-se representar o conceito próximo "chimpanzé". Conceitos que são próximos têm representações similares.

Uma rede de processamento paralelo distribuído, uma rede neural que usa representação distribuída, oferece a vantagem de generalização automática. Se se quer representar o conceito "gorilas são cabeludos", reforça-se a conexão entre todos os nós que compõem o conceito "gorila" e todos os nós que compõem o conceito "cabeludo". Como resultado, desde que a maioria dos nós em '"gorila" são também usados em "chimpanzé", uma associação é também feita entre "chimpanzés" e "cabeludo". É assim que a generalização automática trabalha. Numa representação local, onde "gorila" e "chimpanzé" são representados por nós separados, uma conexão entre "gorila" e "cabeludo" não implicaria numa conexão entre "chimpanzé" e "cabeludo".

Uma outra vantagem de uma representação distribuída é sua insensibilidade a danos.

Numa representação local, se o sistema perde o nó que representa "avó", ele perde seu conceito de avó. Em uma representação distribuída, para perder um conceito, deve-se perder todos os nós que o representa. Se se perde apenas um ou dois nós, o conceito pode se degradar, mas ainda está lá. Isto é mais próximo ao tipo de memória perdida vista em adultos mais velhos.

Esquemas: Uma crítica aos modelos de redes neurais é que eles não são tão flexíveis na representação do conhecimento como os métodos padrões são. Os métodos padrões incluem a rede semântica local.

Os psicólogos cognitivos, notavelmente Jean Piaget, usam o conceito de esquema ("schema"). Um esquema é um espelho - na mente - de uma situação real. Como crianças, e como adultos, nós aprendemos novas associações e relações entre objetos e os integramos ao nosso esquema.

Não é imediatamente claro como um modelo de rede neural pode considerar o conhecimento representado em um esquema; entretanto, Rumelhart, Paul Smolensky, McClelland, e Geoffrey Hinton mostraram que é possível ([McR86]).

Hierarquias cognitivas: Freqüentemente, modelos de redes neurais são ordenados em hierarquias. Muitos níveis existem numa hierarquia, cada um composto de um conjunto de unidades. Tipicamente, as unidades que recebem a entrada estão no fundo do sistema, e as unidades que dão saída estão no alto. Num sistema bottom-up, as unidades em cada nível são conectadas a outras unidades no seu próprio nível e influenciam as unidades em níveis acima deles. Num sistema top-down, as unidades novamente se conectam a unidades no seu próprio nível mas influenciam as unidades em níveis abaixo.

Top-down e bottom-up são conceitos familiares em ciência cognitiva. Por exemplo, na percepção de sentença, estes termos referem a como elementos lingüísticos de tamanhos diferentes, o fonema (som), morfema (elemento palavra), palavra, frase e sentença, interagem um com o outro.

Uma rede de leitura paralela: Um problema na criação de uma rede de leitura é que as pessoas tendem a ler mais de uma palavra de cada vez. Como uma rede simples lê apenas uma palavra, não funciona. Como solução, McClelland propõe cópias duplicadas de redes. Redes de reconhecimento de palavras individuais duplicadas teriam conexões programáveis ao invés de conexões fixas por hardware ("hardwired") entre letras e palavras.

Processamento de sentenças: Um importante aspecto do entendimento de sentença consiste em determinar os varios papéis que as partes diferentes de uma sentença têm. Por exemplo, considere as seguintes sentenças:

O macaco morreu.

O macaco quebrou.

Na primeira sentença, o macaco é um animal, pois morrer é uma característica dos seres vivos; na segunda, o macaco é uma ferramenta de trocar pneus, pois um animal não pode "quebrar". De alguma forma, o modelo deve discernir seus papéis diferentes.

McClelland e Alan Kawamoto desenvolveram um sistema conexionista para fazer esta atribuição de papéis ([McR86]). Palavras são descritas por "microcaracterísticas semânticas" - dimensões básicas que descrevem muitos objetos e ações. Por exemplo, duas das microcaracterísticas que descrevem substantivos são 'humano" e "leveza" , que têm os valores "humano, não-humano", e "leve", "pesado", respectivamente. As palavras não são representadas diretamente nas redes do sistema, mas em termos das ativações de unidades representando microcaracterísticas.

O modelo tem um grupo de unidades para cada um dos papéis principais que substantivos diferentes podem ter em uma ação. Estes papéis são Agente (ator), Paciente (agido sobre), Instrumento (coisa usada) e Modificador (palavra adverbial ou cláusula). Por exemplo, a sentença "O homem comeu o sanduíche", ativaria as microcaracterísticas de "comeu" e "homem" no conjunto das unidades que correspondem ao Agente; isto representa o fato de que o Agente para o verbo "comeu" é "homem".

O sistema é treinado em uma série de sentenças. As atribuições do papel correto para as sentenças de treina-

mento são mostradas ao sistema. Estas atribuições correspondem às ativações de nós particulares. O sistema ajusta as conexões entre esses nós de tal forma que eles se reforcem mutuamente.

Depois de ser treinado com um número suficiente de sentenças, o sistema pode fazer atribuições de papel correto para novas sentenças. Ele ainda pode fazer atribuições de papel correto para sentenças com alguma ambigüidade sintática. Por exemplo, na sentença "O homem abateu o garoto com a maleta", o sistema considera que "maleta" é o Instrumento de "abateu" ao invés de pertencer ao "garoto", desde que "maleta" tenha microcaracterística que indique que ela é um instrumento.

O sistema também manipula bem vários outros problemas, e geralmente faz um bom trabalho em atribuição de papéis.

5. Conclusão

O **futuro** - As redes neurais são boas para várias tarefas de processamento de linguagem natural, incluindo reconhecimento de letra, leitura, e entendimento de sentença. Elas também são úteis no armazenamento de conhecimento em esquemas e em recuperar itens da memória. Elas não são milagrosas, mas mostram uma direção para a inteligência artificial e a psicologia cognitiva, forte e biologicamente plausível, para muitos problemas importantes.

Eventualmente, um modelo conexionista será provavelmente construído do processo de entendimento de linguagem natural, desde que, como os psicólogos têm mostrado, envolva conhecimento integrado de muitos domínios, incluindo fonética, morfologia, sintaxe e semântica. Modelos conexionistas são particularmente bons na integração desses tipos de conhecimento.

Bibliografia

[Abu86] Abu-Mostafa, Yaser S. "Neural Networks for Computing?". AIP Conference Proceedings 151 - pp. 1-6. Neural Networks for Computing - Snowbird, UT1986. Editor: John S. Denker. American Institute of Physics - New York, 1986.

[And83] Anderson, James A. "Cognitive and Psychological Computation with Neural Models. IEEE Transactions on Systems, Man, and Cybernetics 13 (5), Sept/Oct 1983, pp. 799-815.

[AWV88] Anderson, J.A. & Wisniewski, E.J. & Viscuso, S.R. "Software for Neural Networks". Computer Architaecture News (USA), vol. 16, nº 1, pp. 26-36 (March 88).

[Cal86] Callatay, Armand M. de. Natural and Artificial Intelligence - Processor Systems Compared to the Human Brain North-Holland, 1986.

[Fel85] Feldman, Jerome A. "Connections" Byte, pages 277-285. April 1985.

[Gro87] Grossberg, Stephen. "Competitive Learning: From Interactive Activation to Adaptative Resonance". Cognitive Science 11, 23-63 (1987).

[Hen89] Hendler, James A. "Marker-passing over Microfeatures: Towards a Hybrid Symbolic/Connectionist Model". Cognitive Science 13, 79-106 (1989).

[Ive84] Iversen, L.L. "Amino Acids and Peptides: Fast and Slow Chemical Signals in the Nervous System?" Proceedings of the Royal Society of London. B. volume 221 - pp. 245-260 - nº 1224 - 22 May 1984.

[KuN76] Kuffler, Stephen W. & Nicholls, John G. From Neuron to Brain - A Cellular Approach to the Function of the Nervous System. 2nd. edition, 1976. Sunderland.

[Lip87] Lippmann, Richard P. "An Introduction to Computing with Neural Nets". IEEE ASSP Magazine, April 1987, pp. 4-22

[McR86] McClelland, J.L., Rumelhart D.E. and the PDP Research Group. Parallel Distributed Processing Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models. A Bradford Book, The MIT Press, 1986.

[Obe87] Obermeier, Klaus K. "Natural-Language Processing". Byte, pages 225-232. December 1987.

[Roc90] Rocha, A.F. "Symbolic Reasoning: A Natural Affair for K-Neural Nets". IFLSI Conference, Tizuka, Japan, July 1990.

[She74] Shepherd, Gordon M. The synaptic organization of the brain. An Introduction NY, Oxford University Press, 1974.

[Sow84] Sowa, John F. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Publishing Company, 1984.

[Vem88] Vemuri, V. "Artificial Neural Networks: An Introduction" Artificial Neural Networks: Theoretical Concepts V. Vemuri (ed.). The Computer Society of the IEEE, 1988, p. 1-12.

[WaP88] Waltaz, David L. & Pollack, Jordan B. "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation". Connectionist Models and their Inplications: Readings from Cognitive Science. Edited by David Waltz and Jerome A. Feldman Chapter 7, pp. 181-204. Ablex Publishing Corporation. Norwood, New Jersey, 1988.

[Zei87] Zeidenberg, Matthew. Modeling the Brain". Byte, pages 237-246. December 1987.