MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO DE PESQUISAS ESPACIAIS

AUTORIZAÇÃO PARA PUBLICAÇÃO
AUTHORIZATION FOR PUBLICATION

PALAVRAS CHAVES/KEY WORDS

AUTORES/AUTHORS

LINGUAGEM NATURAL
AQUISIÇÃO DE CONHECIMENTO
INTERFACES
TEXTO
COESÃO
COERÊNCIA

AUTORIZADA POR/AUTHORIZED BY

Marco Antonio Raupp
Diretor Geral

AUTOR RESPONSAVEL
RESPONSIBLE AUTHOR

Carlos Alberto de Oliveira

DISTRIBUIÇÃO/DISTRIBUTION

- [ ] INTERNA / INTERNAL
- [x] EXTERNA / EXTERNAL
- [ ] RESTRITA / RESTRICTED

REVISADA POR/REVISED BY

Valter Rodrigues

CDU/UDC

681.3.019

DATA/DATE

Novembro 1988

PUBLICAÇÃO Nº
PUBLICATION NO

INPE-4737-PRE/1412

ORIGEM
ORIGIN

LAC

PROJETO
PROJECT

INTAL

Nº DE PAG.
NO OF PAGES

19

ULTIMA PAG.
LAST PAGE

13

VERSÃO
VERSION

Nº DE MAPAS
NO OF MAPS

TÍTULO/TITLE

O TRATAMENTO AUTOMÁTICO DE LINGUAGEM NATURAL EM PROCESSOS DE AQUISIÇÃO DE CONHECIMENTO

AUTORES/AUTHORSHIP

Carlos Alberto de Oliveira

RESUMO - NOTAS/ABSTRACT - NOTES

A LN é a linguagem de todos os homens, sendo que estes ao usá-la transmitem subliminarmente seu conhecimento sobre o mundo e sobre a própria linguagem. No entanto, a adoção de apriorismos (lexicais, gramaticais, etc.), restringindo a capacidade de expressão dos usuários, impede a exploração desse conhecimento subjacente. Sugere-se neste trabalho, uma interação dinâmica com os usuários para que, explorando o conhecimento intuitivo da língua materna, possa-se suprir, durante a fase de aquisição de conhecimento de dado domínio de aplicação, o universo semântico necessário para a interpretação de frases em tal domínio. Tem-se, então, **uma delimitação semântico-pragmática sem que, para isso, restrinja-se a linguagem do usuário.**

OBSERVAÇÕES/REMARKS

Submetido para apresentação no 5º Simpósio Brasileiro de Inteligência Artificial, 7 a 11 de novembro de 1988, Natal - RN.

ABSTRACT

An approach for natural language processing based on textual knowledge is presented. In this way, any natural language interface must consult users' linguistic knowledge through a dynamic interactive process, in order to build the particular linguistic universe of an application domain.

SUMÁRIO

O TRATAMENTO AUTOMÁTICO DE LINGUAGEM NATURAL EM PROCESSOS DE AQUISIÇÃO DE CONHECIMENTO

Carlos Alberto de Oliveira
Ministério da Ciência e Tecnologia - MCT
Instituto de Pesquisas Espaciais - INPE
Laboratório de Computação e Matemática Aplicada - LAC
12201 - São José dos Campos - SP

RESUMO: A LN é a linguagem de todos os homens, sendo que estes ao usá-la transmitem subliminarmente seu conhecimento sobre o mundo e sobre a própria linguagem. No entanto, a adoção de apriorismos (lexicais, gramaticais, etc), restringindo a capacidade de expressão dos usuários, impede a exploração desse conhecimento subjacente. Sugere-se neste trabalho, uma interação dinâmica com os usuários para que, explorando o conhecimento intuitivo da língua materna, possa-se suprir, durante a fase de aquisição de conhecimento de dado domínio de aplicação, o universo semântico necessário para a interpretação de frases em tal domínio. Tem-se, então, **uma delimitação semântico-pragmática sem que, para isso, restrinja-se a linguagem do usuário**.

1 - CONSIDERAÇÕES INICIAIS

Qualquer tipo de abordagem de Linguagem Natural **(LN)** se faz tendo em vista, de maneira consciente ou inconsciente, hipóteses sobre **o que é linguagem natural**. Dessa forma, a LN pode ser tomada como um conjunto de palavras, como um rol de frases ou como um texto ("discourse"). Nos dois primeiros casos, a interpretação de frases está fundamentada no apriorismo sintático-lexical. No último caso, torna-se necessário o concurso da análise textual e a inserção do conhecimento de mundo do usuário para a consecução dos objetivos propostos.

Interfaces em LN têm seu grau de eficiência determinado pela escolha de uma daquelas hipóteses quando de sua construção. Sugere-se neste trabalho um tipo de abordagem fundamentada na terceira hipótese na tentativa de se eliminar o cerceamento comunicativo (restrições impostas às entradas em LN) e maximizar a integração do processo interpretativo.

2 - O QUE É TEXTO ?

A hipótese de trabalho da análise textual toma como unidade básica, ou seja, como objeto particular de investigação, **não a palavra ou a frase, mas sim, o texto por ser este a forma específica de manifestação da linguagem** (Fávero e Kock, 1983). Considera-se que, desde que cada enunciação pode ter uma multiplicidade de significações, visto que as intenções do falante ao produzir um enunciado podem ser as mais variadas, não se vê sentido na pretensão de lhes atribuir uma interpretação única e verdadeira (Kock, 1984). De uma forma geral, texto "consiste em qualquer passagem, falada ou escrita, que forma um todo significativo, independente de sua extensão. Trata-se pois de uma unidade de **sentido** , de um contínuo comunicativo contextual que se caracteriza pela coerência e pela coesão, conjunto de relações responsáveis pela tessitura do texto"(Fávero e Kock, 1983, p. 25). Indicam a coesão as maneiras como os elementos constituintes do universo textual estão ligados dentro da linearidade, isto é, numa espécie de semântica da Expressão, na qual se estudam as maneiras como são usados padrões formais na transmissão de conhecimentos e sentidos (Fávero, 1985, p. 148); já a coerência, resultado de processos cognitivos que se operam entre usuários, é indicada pelos conceitos e relações que subjazem a Expressão, isto é, situa-se no terreno do Conteúdo. Percebe-se, então, que **representar conhecimento** também se insere no terreno do textual.

Observe-se também que o **usuário** é um dos participantes na construção de um texto: nele estão centradas a intencionalidade, a informatividade, a situacionalidade e a intertextualidade. A

intencionalidade abrange todas as formas como os usuários usam textos para a consecução de seus objetivos; a informatividade designa em que medida os dados lingüísticos são ou não esperados, são ou não conhecidos por parte dos receptores; a situacionalidade refere-se à função dominante do texto (controle ou manejo da situação); a intertextualidade reporta-se às formas pelas quais a produção e recepção de um texto dependem do conhecimento de outros textos (Kock,1985). Diante disso, deve o usuário ser considerado como peça fundamental quando do tratamento automático de LN.

3 - ALGUNS ASPECTOS DA LINGUAGEM NATURAL

A linguagem natural permeia quase, senão, todos os campos do conhecimento humano, haja vista a maioria de nosso acervo cultural estar vazado em LN, mais especialmente na sua modalidade escrita. Dessa forma, o especialista ou o usuário final, ao interagir em dada LN com a máquina, estará sempre veiculando muito mais conhecimento do conscientemente pretende: sua gramática particular (estilo), sua intenção comunitativa, seu universo cultural e, conseqüentemente, o universo lingüístico que expressa o domínio no e sobre o qual discorre. Ademais, a LN já contém em seu bojo elementos que possibilitam inferir e pressupor interpretações possíveis para dada entrada em LN, conforme o universo de relações semânticas de dado domínio o permita. Além disso, quase toda a forma de representação de conhecimento pode ser "traduzida" para a linguagem natural. Neste caso, a estrutura lingüística contém elementos que possibilitam tal tradução, desde que convenientemente depreendidos e manipulados.

3.1 - A problemática dos apriorismos

"Palavras" podem estabelecer diferentes relações e, por isso, assumirem diferentes significações conforme o domínio em que se inserem: o elemento **sol** no domínio "fenômemos meteorológicos", por exemplo, pode estabelecer uma relação de antonímia com **chuva**, ou seja, ~**fazer sol (-) chover**, o que necessariamente nem sempre é verdade no mundo real; no domínio "astronomia", a relação

estabelece-se de maneira indireta com **lua** , ou seja, ambos são particularizações de uma classe de objetos; no domínio "divisão do espaço de tempo=24 horas", as relações estabelecidas são de antonímia com **lua** e de sinonímia com **dia**. Pode-se notar conforme mudam-se os domínios, mudam-se também as relações que dão significação a dada "palavra".

Por tal grau de dificuldade no estabelecimento do conteúdo semântico adequado para dada "palavra", o recurso mais prático comumente usado, no que concerne à elaboração de interfaces, é o de desconhecer a existência da intertextualidade, limitando e predefinindo a coesão e a coerência do domínio através de, por exemplo, bases lexicais e semânticas apriorísticas. Tal proceder, implicitamente, também restringe a capacidade de expressão dos usuários. Isto se explica pelo fato de que o ser humano pode verbalizar a informação de muitas formas diferentes. Se houver restrição à estruturação sintática ou ao léxico, como no caso, o poder de argumentação e de comunicação da linguagem também se restringe. Ou melhor, ela deixa de ser "natural". Nesse contexto, o usuário tem que forçosamente se adaptar a um sistema que use tais apriorísmos sob pena de não se comunicar com ele. Assim, o problema apenas muda de foco: deixa-se de se preocupar com o aprendizado e uso de uma linguagem formal específica para o fazê-lo com o aprendizado das restrições impostas às frases em LN.

3.2 - A LN e as representações de conhecimento

A coerência textual da LN (aspecto conceitual) pode sempre ser representada pelos mesmos esquemas de representação amiudemente usados e conhecidos na área de Inteligência Artificial: frames, scripts, esquemas, etc. Logo, a tradução dos dados contidos numa entrada em LN para qualquer representação especificamente adotada pode ser naturalmente feita.

Exemplificando: "crenças" associadas a dados morfológicos, lexicais, sintáticos e semânticos permitem depreender da frase **"se as condições da zona frontal são favoráveis"** que : **a)** pela taxe

do **SE**, ele deve introduzir uma frase condicional; **b)** a relação entre **AS CONDIÇÕES** e **A ZONA FRONTAL** intermediada pela preposição **DE** permite inferir que a segunda expressão é um complemento da primeira; as "terminações" **-ÇÕES** e **-AL** permitem inferir que as palavras as quais estão agregadas podem ser, respectivamente, a substantivação de um verbo e um adjetivo .Isto possibilita a transformação dessa frase em outra :" se a zona frontal se condiciona favoravelmente". Torna possível também saber que o verbo CONDICIONAR pode se caracterizar como um Atributo da frase,que ZONA FRONTAL é o Objeto da frase, e que FAVORÁVEL é o Valor do Atributo frasal.

Cabe ao "construtor" da interface indicar um conjunto inicial de "crenças" em dadas regras lingüísticas, conforme o uso comum (pragmática) da língua em análise, no molde de sistemas especialistas. No entanto, somente a dinâmica interativa é que poderá reavaliar e/ou atualizar os valores de tais "crenças" aplicáveis a dado domínio **[1]**. No Apêndice está um exemplo sucinto de determinação de algumas "crenças".

Desse modo, pode-se analisar várias formas de expressão do mesmo conteúdo (estruturas de superfície), reduzindo-as sempre para uma estrutura profunda única e, passo contínuo,traduzi-la para uma regra de decisão, por exemplo. Isto dentro de outras interpretações possíveis e para qualquer outro esquema de representação de conhecimento adotado.

4 - A LN EM PROCESSOS DE AQUISIÇÃO DE CONHECIMENTO

Durante a fase de aquisição de conhecimento, ou seja, o processo de inserção de dados que possibilitam dado sistema tornar-se aplicativo (sistemas especialistas, por exemplo), sugere-se que a interação em LN seja adotada para possibilitar, ao

---

[1] Como sugerido e especificado em Oliveira (1987a; 1987b) para os níveis "fonológico" (divisão silábica)/morfológico e morfológico/sintático.

mesmo tempo que se inserem os dados do domínio de aplicação desejados, apreender e delimitar as particularidades lingüísticas válidas para tal domínio [2]. Estas permitirão interações futuras mais livres no diz respeito aos usuários e mais abertas com relação à intencionalidade comunicativa.

A solução aqui sugerida compreende a inserção de um "interlocutor" que possa administrar o processo dialógico e, junto com o usuário, possa construir, delimitar e tornar explícito o universo do diálogo. Este interlocutor deve carregar em seu bojo um "apreendedor" de conhecimento lingüístico que possa absorver e **criar um léxico, uma semântica e uma pragmática pertinentes ao domínio e aos usuários** que manipulam o conhecimento deste. Tal proposição fundamenta-se em: **a)** a coerência(conteúdo) de um domínio de aplicação somente pode ser construída pelos usuários, desde que são eles em última análise que conhecem as relações que ali vigem ; **b)** a expressão lingüística pode variar de usuário para usuário, porém, sendo as relações conceituais sempre as mesmas, é factível o tradução expressão do usuário/semântica do domínio de aplicação e vice-versa ; **c)** dado o fenômeno da intertextualidade, o usuário normalmente transmite mais conhecimento do que transparece e esse conhecimento pode permitir o estabelecimento dos parâmetros em que ele se baseia para determinar a coesão textual.

Assim, nesse contexto, por exemplo, diante de uma frase do tipo **"se existem nuvens no céu"**, introduzida numa sessão de aquisição de conhecimento de dado domínio de aplicação, ao usuário será inqüirido sobre a flexão morfológica da "nova palavra" nuvens, até a flexão verdadeira seja detectada:

---

[2] Sistemas há que já adotam um procedimento similar, no mínimo no que se refere à interação com o usuário. Cite-se o sistema IRUS (Bates, 1984) para banco de dados, onde pode-se aumentar o domínio de aplicação. Os processos estão descritos em Stallard (1984) e Moser (1984).

a nuvem as nuvens
ESTÁ CORRETO ? (s)im/ (n)ão/ não se(i)

Será guardado, então, o lexema (nuve-) correspondente à palavra em análise junto com a regra morfológica que pode gerar todas as formas possíveis e futuras de tal lexema (nuvem, nuvenzinha, etc). Isto impedirá que sobre essas palavras novas questões morfológicas sejam levantadas. A título de exemplo, um pequeno número de regras, baseadas no conhecimento lingüístico e sucintamente descritas a seguir, dão uma mostra de como esse conhecimento lingüístico pertinente à estrutura da LN aliado ao conhecimento semântico-pragmático do usuário podem encetar um processo integrado de análise/interpretação lingüística. Tais regras permitem ao sistema inferir ou questionar sobre um dado que auxilie nesse processo, concomitante e integradamente adquirindo o conhecimento do domínio de aplicação e do subconjunto particular da língua vigente para tal domínio. Assim:

1. as palavras morfologicamente flexionáveis no masculino e feminino podem com ser consideradas como: **a)** pertencentes a classe dos seres vivos, subclasse dos animais , candidatando-se, por isso, com um "crença" maior , ao caso agente (ex.: porco **[3]**, cão,etc); **b)** pertencentes a classe dos seres inaminados abstratos, subclasse dos atributos nominais (adjetivos), candidatando-se como casos estado ou propriedade de outros seres (ex.: bom, nervoso, grande,etc).

2. as flexionáveis em um só gênero podem ser questionados como pertencentes: **a)** a classe dos seres vivos, animais, agentes (ex.: cobra, João **[4]**, Maria, etc) ou vegetais/microorganismos, não-agentes (ex.: bactéria, árvore) ou **b)** seres inanimadoss,

---

[3] Considerem-se flexionáveis em dois gêneros as formas supletivas: homem/mulher, touro/boi/vaca, bode/cabra, etc.

[4] No caso de nomes próprios, estes se candidatam preponderantemente como particularizações dos casos agente ou locativo

concretos (ex.: pedra, água, etc) ou abstratos (ex.: gente, pessoa, dor, etc). Incluído aqui o exemplo anteriormente citado sobre "nuvens", o usuário será questionado, se dúvidas houver, sobre a classificação semântica mais provável de tal lexema:

NUVEM É _____
1. um ser com forma/cor/ volume, etc
2. outro tipo de coisa

3. as flexionáveis em um só gênero e um só número podem ser inanimados concretos (ex.: óculos) ou abstratos (ex.: féria, férias).

4. os verbos podem ser classificados, num processo similar à dependência conceitual schankiana, conforme certos primitivos da língua portuguesa. Assim, por exemplo: chover/ "fazer" [objeto:chuva], trabalhar/ "fazer" [objeto:trabalho], chorar/ "fazer" [objeto: cair [objeto: lágrima]], pensar/ "fazer" [objeto: pensamento].

Ao responder tais questões, o usuário estará validando e hierarquizando, conforme aumento de certas "crença", dadas regras que permitirão aumentar gradativamente a eficiência do sistema. No exemplo em discussão, uma relação entre nuvem/céu (conteúdo/continente, por hipótese) será estabelecida por regras sintáticas (coesão), facilitando inferências futuras sobre sinonímias, tais como, altos níveis/céu. Outro exemplo: se num banco de conhecimento de dado domínio (fenômenos meteorológicos, por hipótese) houver o objeto chuva e um usuário genérico questionar **"vai fazer sol amanhã?"**, a relação **chover/ fazer [objeto:chuva]** e o conhecimento de que **chuva é objeto inanimado** já

adquiridos permitem pressupor a relação **fazer sol (*ensolarar[5])/ fazer chuva (chover)**, dentre outras, é claro. Logo a palavra "sol" deverá obedecer a mesma regra de flexão de "nuvem" (um só gênero): se confirmada tal pressupossição pelo usuário, questiona-se qual a relação que há entre os objetos "sol" e "nuvem". Nesse proceder, não se recusa uma entrada em LN pelo fato de não existir no léxico a palavra em análise, mas sim, busca-se interpretar primeiro o que o usuário quis dizer, isto é, sua intencionalidade: no nosso exemplo, possivelmente, **"vai fazer sol amanhã? = poderá chover no espaço de 24 horas?"**.

5 - CONSIDERAÇÕES FINAIS

Sintetizando: sugere-se no tratamento automático de LNs conjugar o que o usuário pode e deve saber (conhecimento de mundo) com o que a interface sabe (regras) e faz (inferências), integrando-se os níveis de análise lingüística. Estes fornecerão entre si maiores ou menores "crenças" em determinadas conclusões: o que puder ser inferido o será e, em caso de duas ou mais interpretações possíveis, caberá ao usuário decidir e informar o que mais se adequa ao contexto. À medida que as frases forem sendo fornecidas e as questões do sistema sendo respondidas, este incorporará as relações abstraídas das respostas, aumentando o poder de interpretação da interface. Serão estabelecidas "crenças" lingüísticas que funcionarão como "guias" de análise, construindo-se um subconjunto específico de regras para a tarefa de interpretar frases num domínio particular de aplicação.

[5] O asterisco antes da palavra denota aqui a não-verbalização ou o pouco uso, sendo porém possível dentro do sistema lingüístico do usuário.Do mesmo modo que, embora potencialmente tenham a mesma significação, pode-se ter pensamento, pensação, pensência, priviligiando-se pelo uso a primeira forma.

REFERÊNCIAS BIBLIOGRÁFICAS

BATES, M. Accessing a database with a transportable natural language interface. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATIONS, 1, Denver, CO., 1984, **Proceedings**, Los Angeles, CA., IEEE Computer Society, 1984, p. 9-12

FÁVERO, L.; KOCK, I. G. V. **Linglüística textual: uma introdução.** São Paulo, Cortez, 1983

KOCK, I. G. V. **Argumentação e linguagem**. S. Paulo, Cortez, 1984

_____ Mesa Redonda: Linglüística Textual. In: GRUPOS DE ESTUDOS LINGÜÍSTICOS, 10, Bauru, SP, 1985, **Anais**, Bauru, Faculdade Sagrado Coração, 1985, p. 153-161

MOSER, M. G. Domain dependent semantic acquisition. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATIONS, 1, Denver, CO., 1984, **Proceedings**, Los Angeles, CA., IEEE Computer Society, 1984, p. 13-18

OLIVEIRA, C. A. A divisão silábica e a morfologia: um enfoque integrado baseado no conhecimento linglüístico. São José dos Campos- SP,INPE, 1987 (INPE-4259-PRE/1132)

_____. A morfologia e a sintaxe: um enfoque integrado baseado no conhecimento linglüístico. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 1, Uberlândia, MG, 1987, **Anais**, Uberlândia, MG, UFUb, 1987, p. 187-196

STALLARD, D. Data modeling for natural language acess. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATIONS, 1, Denver, CO., 1984, **Proceedings**, Los Angeles, CA., IEEE Computer Society, 1984, p. 19-24

APÊNDICE (*)

	Condição	"Ação-Crença"
Sintática		
o(s)	1.frase (início)	determinante(.7);anáfora(.2);...
	2.infinit./nomes (antes).	determinante(.9) anáfora(.1).
Morfologia		
-a(s)	1.radical verbal.	atributo frasal (.6) [ex.:fuga];...
	2.radical nominal	masc-sing(.7);...
Léxico		
(nome próprio)	preposição "em" (anteposta)	caso locatico (.7);...
Semântica		
lexema	1.classe:animal	agente (.6); objeto (.3);...
	2.classe:ação	atributo frasal (.8);...

(*) Reticências denotam outras opções aqui não apresentadas.

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO DE PESQUISAS ESPACIAIS

PROPOSTA PARA PUBLICAÇÃO

- ☐ DISSERTAÇÃO
- ☐ TESE
- ☐ RELATÓRIO
- ☒ OUTROS

IDENTIFICAÇÃO

TÍTULO: O tratamento automático de linguagem natural em processos de aquisição de conhecimento

AUTOR(ES): C. A. Oliveira

ORIENTADOR:

CO-ORIENTADOR:

DISS. OU TESE

LIMITE: __/__/__ | DEFESA: __/__/__ | CURSO: | ORGÃO:

DIVULGAÇÃO

☒ EXTERNA ☐ INTERNA ☐ RESTRITA

EVENTO/MEIO

☒ CONGRESSO ☐ REVISTA ☐ OUTROS

REV. TÉCNICA

NOME DO REVISOR: Valter Rodrigues

RECEBIDO: __/__/__ | DEVOLVIDO: __/__/__ | ASSINATURA:

APROVAÇÃO

NOME DO RESPONSÁVEL: F. E. C. Vidal

APROVADO: ☐ SIM ☐ NÃO | DATA: __/__/__ | ASSINATURA:

REV. LINGUAGEM

Nº | PRIOR. | RECEBIDO: __/__/__ | NOME DO REVISOR:

PÁG. | DEVOLVIDO: __/__/__ | ASSINATURA:

DATILOGRAFIA

OS AUTORES DEVEM MENCIONAR NO VERSO INSTRUÇÕES ESPECÍFICAS, ANEXANDO NORMAS, SE HOUVER

RECEBIDO: | DEVOLVIDO: __/__/__ | NOME DA DATILÓGRAFA: Mª Helena

DIRETOR

Nº DA PUBLICAÇÃO: | PÁG.:

CÓPIAS: | Nº DISCO: | LOCAL:

AUTORIZO A PUBLICAÇÃO

☐ SIM

☐ NÃO __/__/__

OBSERVAÇÕES E NOTAS

Solicita dispensa de revisor de linguagem

Submetido pl 6º SBIA

APÓS A PUBLICAÇÃO, ESTA FICHA SERÁ ARQUIVADA JUNTAMENTE COM A VERSÃO ORIGINAL

INPE 106