CORREFSUM: REVISÃO DE COESÃO REFERENCIAL EM SUMÁRIOS EXTRATIVOS

Mestrando: Patricia Nunes Gonçalves

Orientadora: Renata Vieira

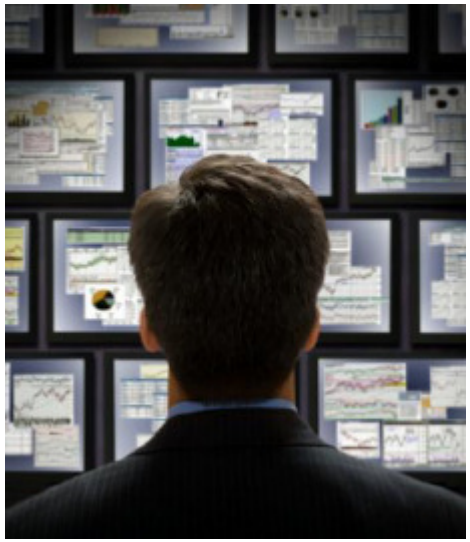
UNISINOS

SUMÁRIO

- Introdução
- Objetivos
- Ferramentas
- Exemplo
- Sistema CorrefSum
- Experimentos
- Avaliação Automatica
- Avaliação Sujetiva
- Considerações Finais

INTRODUÇÃO

- Atualmente convivemos com a sobrecarga de informação.
- Nesse contexto a área de sumarização automática tem se tornado uma área proeminente, contribuindo para filtrar e discernir informações de maior relevância



INTRODUÇÃO

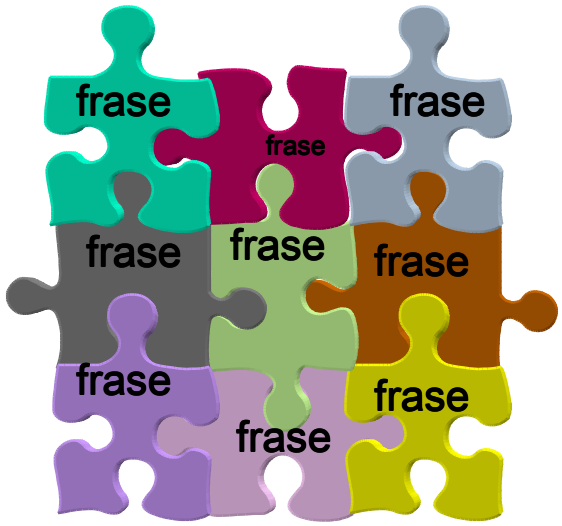
- A sumarização é o processo de seleção de informações mais importantes de um texto, com o objetivo de produzir uma versão resumida do mesmo. [Mani,2001]

INTRODUÇÃO

- O foco deste trabalho é na análise e na recuperação da coesão referencial nos sumários extrativos que utilizam a escolha de sentenças de maior relevância do texto para compor o sumário.

SUMARIZAÇÃO EXTRATIVA

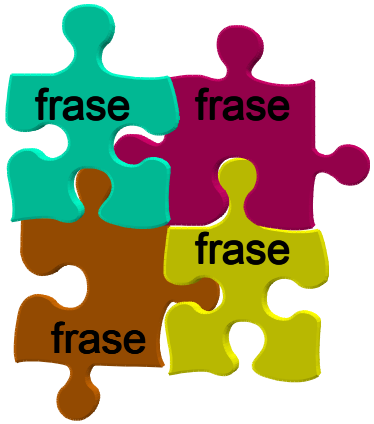
MÉTODO EXTRATIVO – SUMARIZAÇÃO AUTOMÁTICA



texto-fonte

freqüência
- stopwords

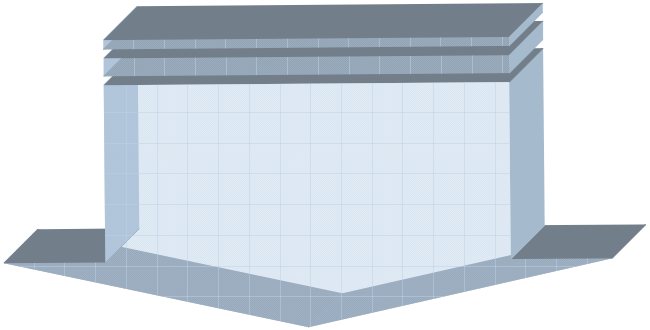
posição



sumário

CADEIAS DE CORREFERÊNCIA

Cadeias de Correferência



diferentes expressões invocando um mesmo referente

Exemplo de Cadeia:

- Antônio Calmon, diretor da Motorola
- Calmon
- O diretor da empresa
- Ele

OBJETIVOS

- Objetivo geral é enriquecer os sumários extrativos com a aplicação de resolução de correferência utilizando a recuperação de expressões referenciais mais completas nos textos-fonte.
  - Análise da coesão referencial dos sumários.
  - Verificar na cadeia de correferência do texto fonte buscando um antecedente textual mais completo.
  - Implementar uma ferramenta para manipular essa informação. Possibilitar a configuração automática ou manual do sistema.
  - Avaliação dos sumários.

FERRAMENTAS E RECURSOS

- Palavras [Bick,2002]
- MMAX[Muller and Strube, 2000]
- GistSumm[Pardo,2005]
- Supor-2[Leite,2007]
- Corpus Summ-it[Collovini et al, 2007]
- Rouge[Lin, 2004]

EXEMPLO

[S1]A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia.
[S2]A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina).
[S3]Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.
[S4]Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare.
[S5]Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra.
[S6]O presidente da Embrapa (Empresa Brasileira de Pesquisa Agropecuária), Alberto Portugal, salientou que a empresa busca soluções para os problemas da agricultura nacional.
[S7]Ele citou o exemplo de pesquisas que, por meio de engenharia genética, buscam obter mamão livre de vírus e feijão também resistente a vírus, culturas de interesse para exportação e consumo interno.
[S8]Portugal disse que os agronegócios correspondem a 25% do PIB brasileiro e que a biotecnologia é fundamental para manter a competitividade da agricultura.

CIENCIA_2000_6389

EXEMPLO

[S1]A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia.
[S2]A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina).
[S3]Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.
[S4]Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare.
[S5]Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra.
[S6]O presidente da Embrapa (Empresa Brasileira de Pesquisa Agropecuária), Alberto Portugal, salientou que a empresa busca soluções para os problemas da agricultura nacional.
[S7]Ele citou o exemplo de pesquisas que, por meio de engenharia genética, buscam obter mamão livre de vírus e feijão também resistente a vírus, culturas de interesse para exportação e consumo interno.
[S8]Portugal disse que os agronegócios correspondem a 25% do PIB brasileiro e que a biotecnologia é fundamental para manter a competitividade da agricultura.

CIENCIA_2000_6389

EXEMPLO

[S1]A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia.
[S2]A opinião é d**o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)**.
[S3]Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.
[S4]Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare.
[S5]Para **o agrônomo**, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra.
[S6]O presidente da Embrapa (Empresa Brasileira de Pesquisa Agropecuária), Alberto Portugal, salientou que a empresa busca soluções para os problemas da agricultura nacional.
[S7]Ele citou o exemplo de pesquisas que, por meio de engenharia genética, buscam obter mamão livre de vírus e feijão também resistente a vírus, culturas de interesse para exportação e consumo interno.
[S8]Portugal disse que os agronegócios correspondem a 25% do PIB brasileiro e que a biotecnologia é fundamental para manter a competitividade da agricultura.

CIENCIA_2000_6389

RESULTADO

- Sumário Automático

  Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

- Sumário Corrigido

  O agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina) citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

CIENCIA_2000_6389

RESULTADO

Utilizando 1º Aposto

- Sumário Automático

  Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

- Sumário Corrigido

  O agrônomo Miguel Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

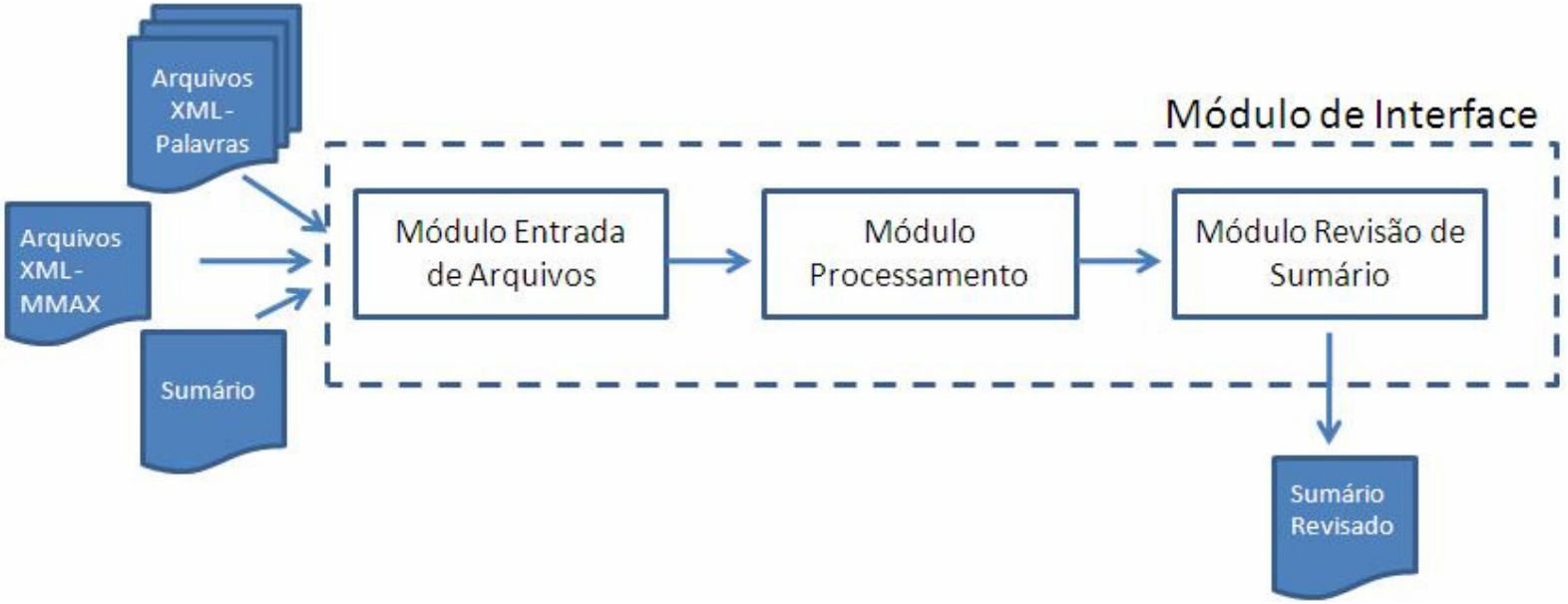
CIENCIA_2000_6389

MÉTODOS

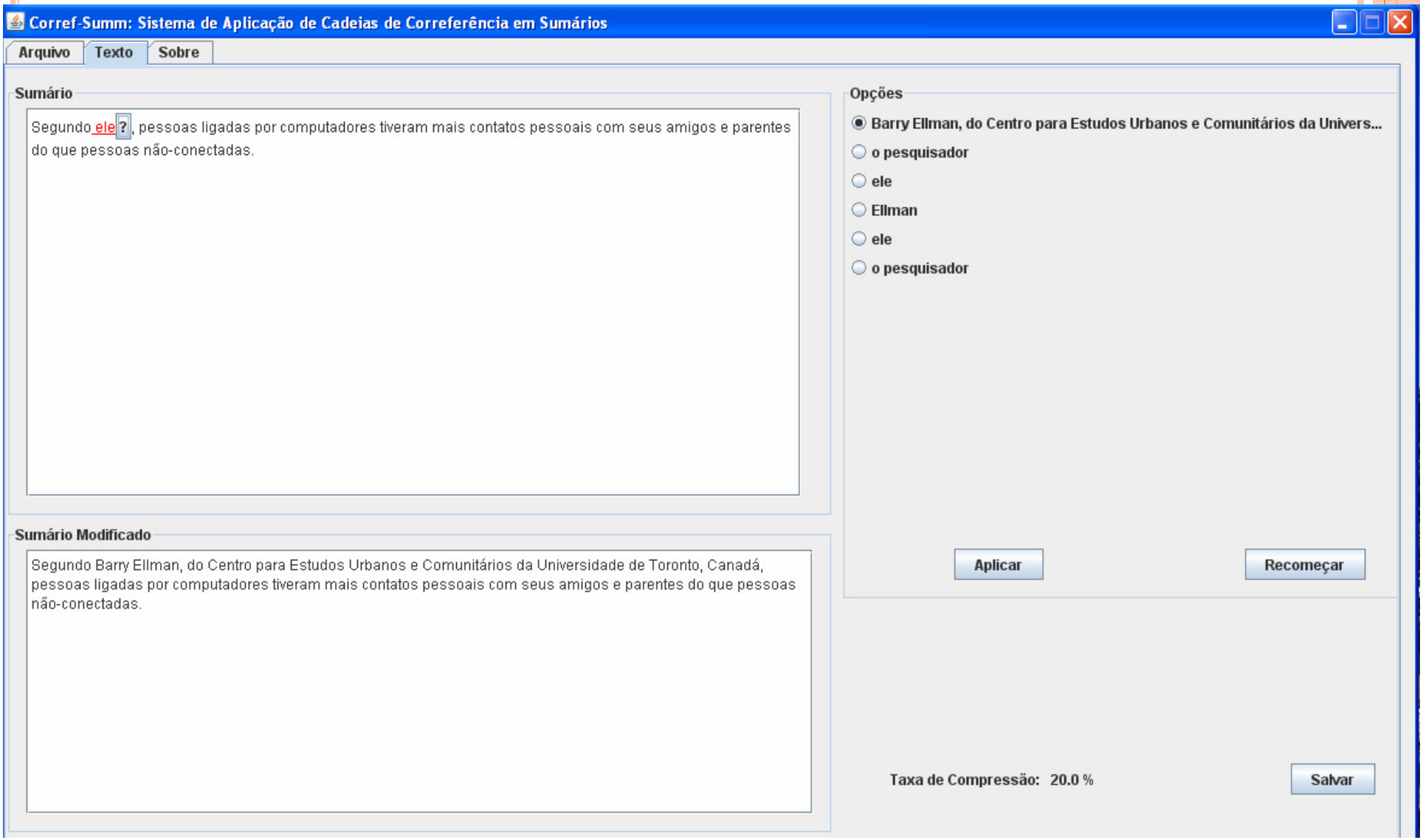
- Utiliza sistema de pontuação.
- Métodos implementados para pontuação:
  - Maior Sintagma - ***Sars (sigla em inglês para síndrome respiratória aguda grave)***
  - Primeiro Sintagma da Cadeia - ***a proteína beta-amilóide.... a proteína..... ela...***
  - Possui Aposto - ***a gripe de 1918, a chamada gripe espanhola***
  - Possui Proprio – ***Jonas Perales do Laboratório de Toxinologia***

SISTEMA CORREFSUM

- Visão Geral do Sistema



SISTEMA CORREFSUM



EXPERIMENTOS

- GistSumm

	QTDE CADEIAS NO TEXTO	QTDE CADEIAS NO SUMARIO	QTDE. TROCAS	TAXA COMPRESSÃO ANTES (%)	TAXA COMPRESSÃO DEPOIS(%)
SOMA	586	330	89	-	-
MÉDIA	11,72	6,60	1,78	26,16	29,10

- Supor-2

	QTDE CADEIAS NO TEXTO	QTDE CADEIAS NO SUMARIO	QTDE. TROCAS	TAXA COMPRESSÃO ANTES (%)	TAXA COMPRESSÃO DEPOIS(%)
SOMA	586	402	67	-	-
MÉDIA	11,72	8,04	1,34	33,24	34,96

AVALIAÇÃO AUTOMÁTICA - ROUGE

- GistSumm

	GISTSUMM-ORIGINAL			GISTSUMM-CORRIGIDO		
	Cobertura	Precisão	F-Measure	Cobertura	Precisão	F-Measure
MÉDIA	48,41	56,6	51,83	53,15	56,03	54,25

- Supor-2

	SUPOR-2-ORIGINAL			SUPOR-2-CORRIGIDO		
	Cobertura	Precisão	F-Measure	Cobertura	Precisão	F-Measure
MÉDIA	63,6	59,34	60,94	64,7	57,15	60,26

AVALIAÇÃO SUBJETIVA -INFORMATIVIDADE

GistSumm

	JUIZ 1		JUIZ 2	
	Original	Corrigido	Original	Corrigido
CIENCIA_2000_6389		x		x
CIENCIA_2000_6391		x	x	
CIENCIA_2000_17109		x		x
CIENCIA_2000_17112		x		x
CIENCIA_2001_6410		x		x
CIENCIA_2001_6416		x		x
CIENCIA_2002_22005		x		x
CIENCIA_2004_6494		x	x	x
CIENCIA_2005_28747		x	x	
CIENCIA_2005_28766		x		x
SOMA	0	10	3	8

Supor-2

	JUIZ 1		JUIZ 2	
	Original	Corrigido	Original	Corrigido
CIENCIA_2000_17088		x		x
CIENCIA_2000_17112	x			x
CIENCIA_2000_17113	x		x	
CIENCIA_2001_6410	x			x
CIENCIA_2002_22005		x		x
CIENCIA_2003_24212		x		x
CIENCIA_2003_24219		x	x	
CIENCIA_2004_6480		x		x
CIENCIA_2004_26415		x		x
CIENCIA_2004_26423		x		x
SOMA	3	7	2	8

CONSIDERAÇÕES FINAIS

- Enriquecimento dos sumários extrativos
- Manutenção da coesão referencial dos sumários
- Primeira abordagem para língua portuguesa
- Construção de uma interface para usuários

BIBLIOGRAFIA

- [Bick, 2000] Bick, E. (2000). The Parsing System "PALAVRAS- Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Department of Linguistics, University of Århus, DK.
- [Collovini et al., 2007] Collovini, S., Carbonel, T., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In 5o Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007), Rio de Janeiro, RJ. Proceedings of the SBC.
- [Lin, 2004] Lin, C.-Y. (2004). Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In Proceedings of 4th Workshop NTCIR.
- [Mani, 2001] Mani, I. (2001). Automatic Summarization. John Benjamins Publishing Co.
- [Müller and Strube, 2001] Müller, C. and Strube, M. (2001). Mmax: A tool for the annotation of multi-modal corpora. In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, pages 45-50, Seattle, Washington.
- [Pardo, 2005] Pardo, T. (2005). Gistsumm - gist summarizer: Extensões e novas funcionalidades. Technical report, NILC-TR-05-05. São Carlos-SP.