

Boletim Técnico da Escola Politécnica da USP
Departamento de Engenharia de Computação e
Sistemas Digitais

ISSN 1413-215X

BT/PCS/0321

Alinhamento de Corpus
Bilíngües: Modelos e Aplicações

José Fontebasso Neto
Jorge Kinoshita

São Paulo - 2003

1376749

O presente trabalho é parte da dissertação de mestrado apresentada por José Fontebasso Neto, sob a orientação do Prof. Dr. Jorge Kinoshita.: "Alinhamento de Corpus Bilíngues: Modelos e Aplicações", defendida em 19/09/03, na EPUSP.

A íntegra da dissertação encontra-se à disposição com o autor e na Biblioteca de Engenharia Elétrica da Escola Politécnica da USP.

FICHA CATALOGRÁFICA

Fontebasso Neto, José

Alinhamento de corpus bilíngües : modelos e aplicações / José Fontebasso Neto, Jorge Kinoshita. -- São Paulo : EPUSP, 2003.

27 p. -- (Boletim técnico da Escola Politécnica da USP, Departamento de Engenharia de Computação e Sistemas Digitais ; BT/PCS/0321)

1. Linguagem natural (Processamento) 2. Modelos matemáticos
I. Kinoshita, Jorge II. Universidade de São Paulo. Escola Politécnica.
Departamento de Engenharia de Computação e Sistemas Digitais III.
Título IV. Série
ISSN 1413-215X

CDD 006.35
511.8

Alinhamento de Corpus Bilíngües: modelos e aplicações

José Fontebasso Neto - jose.fontebasso@poli.usp.br

Jorge Kinoshita - jorge.kinoshita@poli.usp.br

Depto de Engenharia de Computação e Sistemas Digitais

Escola Politécnica - Universidade de São Paulo

Resumo

Esta dissertação trata de modelos matemáticos para algumas aplicações em processamento de linguagem natural, cobrindo as etapas de identificação do limite das sentenças, alinhamento de sentenças e alinhamento de palavras, esse último através de aplicações.

Um objetivo é o desenvolvimento de um processo automático que, partindo de um par de textos planos, sem qualquer marcação específica, a não ser a informação de que são traduções mútuas, produza as marcações e informações necessárias à construção de bases de dados ao nível de palavras e suas traduções.

Outro objetivo é comparar o desempenho dos modelos quando aplicados aos idiomas português, inglês e espanhol e suas combinações, verificando se funcionam tão bem para o português como para o inglês e o espanhol.

Todos os programas necessários à implementação dos modelos foram desenvolvidos com ferramentas livres e gratuitas, disponíveis na internet, como: sistema operacional Linux e interpretadores PERL, que são executáveis em computadores PC comuns sem qualquer característica adicional.

1 Histórico

A idéia básica de *machine translation* é construir sistemas automáticos para fazer e auxiliar o trabalho de tradução entre idiomas. Tal idéia está intimamente ligada à história da computação digital e é tão antiga quanto ela.

Os esforços para construir os primeiros computadores digitais surgiram da necessidade de decifrar as mensagens transmitidas pelas linhas inimigas durante a Segunda Grande Guerra. Em 1940, Alan Turing e sua equipe construíram o primeiro computador digital e com ele, decifram as mensagens transmitidas pelo exército alemão [1]. Provavelmente foi o primeiro trabalho de *machine translation* da história.

Em 1949, Warren Weaver escreveu sobre tradução automática usando modelos estatísticos [19, 10] porém suas idéias ficaram esquecidas por cerca de 40 anos, até que na década de 90, com a disponibilização de grandes quantidades de textos em meio digital e de computadores com capacidade de processamento para atacar o problema, a abordagem estatística foi retomada para os trabalhos em processamento de linguagem natural, como:

- *Tagging*: atribuição das classes gramaticais às palavras em um texto (Brill [15], Ratnaparkhi [14]);
- *Parsing*: análise sintática da estrutura de um texto (Daelemans, Buchholz, Veenstra [18], Ratnaparkhi [14]);
- *Text Alignment*: busca e alinhamento de partes equivalentes entre traduções mútuas (Gale e Church [26], Kay e Röscheisen [27], Chen [28, 29] e Melamed [3, 4, 5])
- *Machine Translation*: tradução de textos em idiomas diferentes (Brown et alli [20] e Melamed [10]).

Em geral, os trabalhos de *processamento de linguagem natural* baseados em modelos estatísticos resultam em problemas que requerem uma extensa preparação dos *corpora*¹.

Neste trabalho, discutiremos alguns modelos para preparação dos corpora visando desenvolver um processo automático que partindo de um par de textos, os quais sabemos serem traduções mútuas, produza um conjunto de pares de palavras equivalentes.

¹Corpora são textos e suas traduções em diferentes idiomas, para treinamento e validação dos modelos.

Nos experimentos utilizaremos corpora, em inglês, espanhol e português, que construímos a partir das “*man pages*” do Linux² por uma questão de disponibilidade e facilidade, já que estão disponíveis gratuitamente, em vários idiomas, na internet.

2 A Preparação dos Corpora

Iniciamos as etapas da preparação dos corpora com disponibilização de textos em dois idiomas que, a priori, sabemos serem traduções mútuas e resultam em um processo separado em três etapas genéricas:

1. Separar os textos em sentenças;
2. Alinhar as sentenças dos textos;
3. Alinhar as palavras nas sentenças,

a partir daí, temos as entradas para nossas aplicações. A figura 1 dá uma visão do processo.

Fazemos a separação em três etapas para aumentar a eficiência, reduzindo o conjunto de possibilidades a ser pesquisado e aumentar a eficácia, fornecendo respostas com menor ambigüidade. Poderíamos pensar em combinar cada palavra do texto original com cada palavra do texto traduzido, colocar as contagens em ordem decrescente e produzir as entradas para as aplicações a partir das maiores contagens, mas além do número de operações, que cresce exponencialmente com a quantidade de palavras, há outros fenômenos, como a co-ocorrência, que pode fornecer resultados ambíguos.

Para ilustrar o desenvolvimento do processo, tomemos como exemplo a figura 2, onde estão dois trechos de texto que, a priori, são traduções mútuas, um em inglês e outro em português:

²“*man pages*” são documentos com informações sobre a descrição e uso dos diversos comandos do sistema operacional Linux.

Figura 1: processo de preparação dos corpora

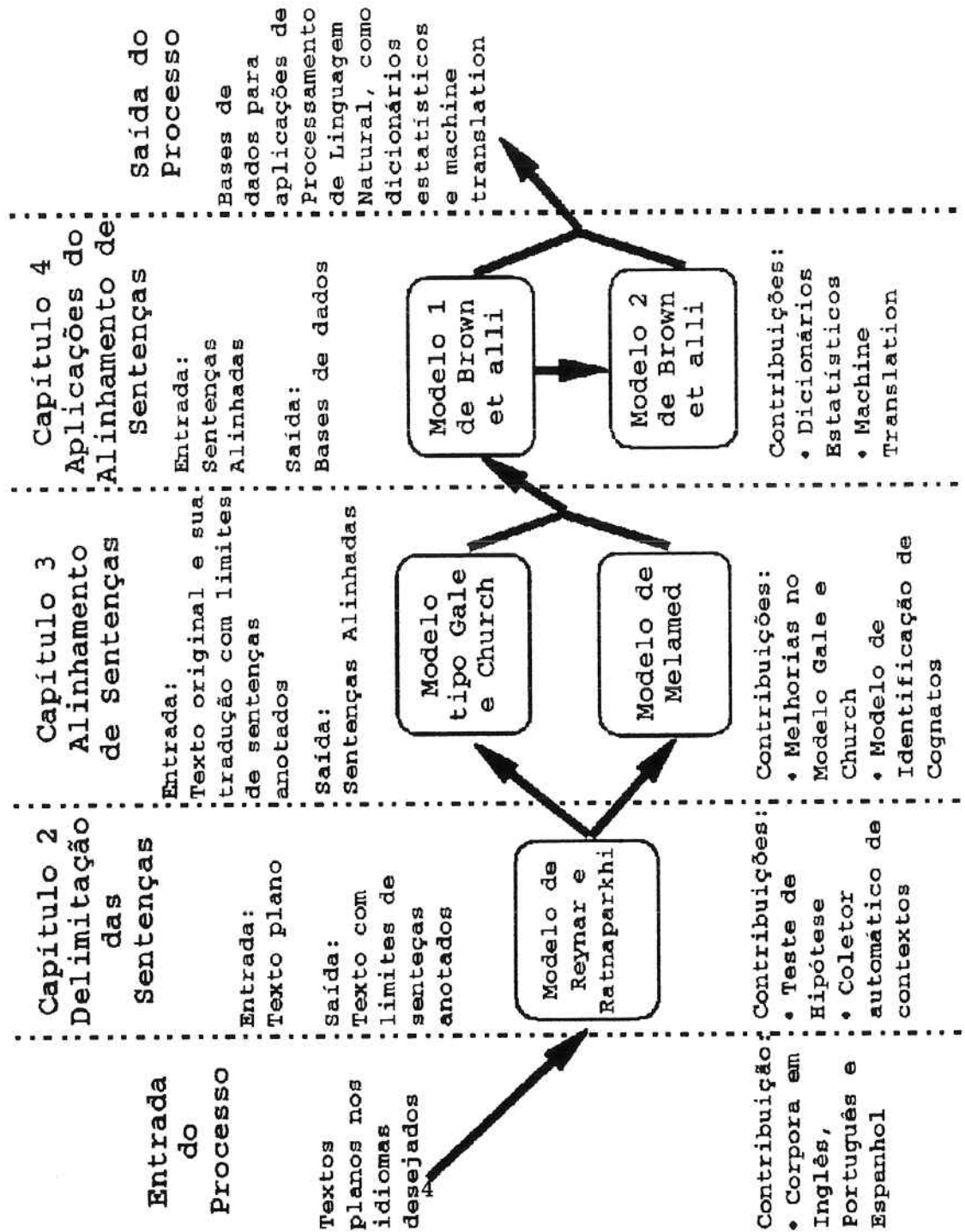


Figura 2: trechos de texto

Trecho Inglês	Trecho Português
For long-distance network links, a different type of equipment is frequently used, which is based on a standard named X.25. Many so-called Public Data Networks, like Tymnet in the U.S., or Datex-P in Germany, offer this service. X.25 requires special hardware, namely a Packet Assembler/Disassembler or PAD and it defines a set of networking protocols of its own right, but is nevertheless frequently used to connect networks running TCP/IP and other protocols. Since IP packets cannot simply be mapped onto X.25 (and vice versa), they are simply encapsulated in X.25 packets and sent over the network.	Para conexões de rede a longas distâncias, um diferente tipo de equipamento é utilizado, o qual é baseado em um padrão chamado X.25. Muitas das chamadas Redes Públicas de Dados, como a Tymnet nos Estados Unidos, ou a Datex-P na Alemanha, oferecem este serviço que requer um hardware especial, chamado de Montador/Desmontador de Pacotes ou PAD. O X.25 define um conjunto de protocolos de rede próprio, mas frequentemente é usado para conectar redes que executem TCP/IP ou outros protocolos. Considerando que pacotes IP não podem ser mapeados em X.25 e vice-versa, eles são simplesmente encapsulados em pacotes X.25 e enviados pela rede.

2.1 Separar Sentenças

Nossa primeira etapa, trata a identificação do limite das sentenças, pois os modelos de alinhamento, como os propostos por Gale e Church [26], Kay e Röscheisen [27], Melamed [4], necessitam que os textos estejam com suas sentenças separadas. Em Chen [29] há uma exceção interessante.

Aqui, *sentenças* são fragmentos de texto limitados entre dois dos seguintes sinais de pontuação: ponto final “.”, ponto de exclamação “!” e ponto de interrogação “?” assim, para identificar os limites das sentenças, basta encontrar e marcar os sinais de pontuação “.”, “!” e “?”.

Vejamos na figura 2 que há pontos finais ambíguos nos trechos ... *U.S.* ... e ... *X.25* ..., que não se referem a limites de sentenças, exigindo um refinamento na busca e marcação dos limites de sentenças.

Para tratar os sinais de pontuação ambíguos, implementaremos um modelo de identificação de limites de sentenças para encontrar e marcar apenas os sinais de pontuação “.”, “!” e “?” que realmente sejam os limites das sentenças, produzindo uma separação do texto como a da figura 3.

Figura 3: trechos separados em sentenças

Trecho Inglês	Trecho Português
For long-distance network links, a different type of equipment is frequently used, which is based on a standard named X.25. Many so-called Public Data Networks, like Tymnet in the U.S., or Datex-P in Germany, offer this service. X.25 requires special hardware, namely a Packet Assembler/Disassembler or PAD and it defines a set of networking protocols of its own right, but is nevertheless frequently used to connect networks running TCP/IP and other protocols. Since IP packets cannot simply be mapped onto X.25 (and vice versa), they are simply encapsulated in X.25 packets and sent over the network.	Para conexões de rede a longas distâncias, um diferente tipo de equipamento é utilizado, o qual é baseado em um padrão chamado X.25. Muitas das chamadas Redes Públicas de Dados, como a Tymnet nos Estados Unidos, ou a Datex-P na Alemanha, oferecem este serviço que requer um hardware especial, chamado de Montador/Desmontador de Pacotes ou PAD. O X.25 define um conjunto de protocolos de rede próprio, mas frequentemente é usado para conectar redes que executem TCP/IP ou outros protocolos. Considerando que pacotes IP não podem ser mapeados em X.25 e vice-versa, eles são simplesmente encapsulados em pacotes X.25 e enviados pela rede.

Nosso modelo, baseado no trabalho de Reynar e Ratnaparkhi [12], utiliza modelos de máxima entropia para analisar o *contexto* em torno do candidato e decidir pela condição desse. O contexto analisado envolve:

- Palavra³ à esquerda: é a palavra anterior e não adjacente ao candidato;
- Prefixo: é a palavra anterior e adjacente ao candidato;
- Sufixo: é a palavra posterior e adjacente ao candidato;
- Palavra à direita: é a palavra posterior e não adjacente ao candidato.

Em linhas gerais, na fase de treinamento, o modelo extrai o contexto a partir de condições bem definidas nos textos, como: finais de parágrafos, que são limites de sentenças e de URLs e siglas⁴, que não são limites de sentenças.

³Palavra é toda sequência de símbolos gráficos, como: letras, números e outros sinais, exceto espaços em branco. Por outro lado, uma sequência com um ou mais espaços em branco é também uma palavra.

⁴Siglas são sequências de sinais gráficos, sem espaços em branco, com um candidato inserido entre eles.

Na sequência, os contextos são submetidos a um teste de hipótese para atestar sua significância em relação à condição do candidato e têm suas probabilidades de ocorrência calculadas.

A última etapa do treinamento é a estimativa dos parâmetros que maximizam a entropia. Isso é feito através do algoritmo *Generalized Iterative Scaling* (Darroch e Ratcliff, 1972, apud Ratnaparkhi [14]).

A fase de aplicação consiste em percorrer o texto, coletar o contexto em torno dos candidatos, calcular suas probabilidades usando os parâmetros estimados no treinamento e verificar qual é a maior probabilidade para o candidato, ser, ou não, um limite de sentença.

2.2 Alinhar Sentenças

Com os textos separados em sentenças, nosso objetivo na segunda etapa é descobrir quais grupos de sentenças são traduções mútuas, o qual, como no caso dos limites de sentenças, também não é problema simples, vez que os tradutores humanos, por motivos linguísticos, estéticos e de clareza, nem sempre mantêm a distribuição de parágrafos e sentenças do texto original.

Para dar uma idéia do problema, vejamos, na figura 4, o nosso exemplo com seus grupos de sentenças alinhados. O primeiro grupo tem um par de sentenças, o segundo grupo contém dois pares de sentenças e o terceiro grupo, um par de sentenças.

Observamos que no segundo grupo, a segunda sentença do trecho inglês é correspondente a parte da primeira sentença e a totalidade da segunda sentença do trecho português, enquanto que a primeira sentença do trecho em português corresponde a totalidade da primeira sentença e a parte da segunda sentença do trecho inglês, de modo que o alinhamento só é correto quando os dois pares de sentenças são considerados.

Há vários modelos para alinhar sentenças, todos o fazem através de algum tipo de medida do comprimento das sentenças, isto é, medem o comprimento em caracteres ou em palavras e determinam, através de uma função estatística ou de alguma heurística, a probabilidade de que as sentenças se alinhem,

Figura 4: grupos de sentenças alinhados

Trecho Inglês	Trecho Português
For long-distance network links, a different type of equipment is frequently used, which is based on a standard named X.25.	Para conexões de rede a longas distâncias, um diferente tipo de equipamento é utilizado, o qual é baseado em um padrão chamado X.25.
Many so-called Public Data Networks, like Tymnet in the U.S., or Datex-P in Germany, offer this service. X.25 requires special hardware, namely a Packet Assembler/Disassembler or PAD and it defines a set of networking protocols of its own right, but is nevertheless frequently used to connect networks running TCP/IP and other protocols.	Muitas das chamadas Redes Públicas de Dados, como a Tymnet nos Estados Unidos, ou a Datex-P na Alemanha, oferecem este serviço que requer um hardware especial, chamado de Montador/Desmontador de Pacotes ou PAD. O X.25 define um conjunto de protocolos de rede próprio, mas frequentemente é usado para conectar redes que executem TCP/IP ou outros protocolos.
Since IP packets cannot simply be mapped onto X.25 (and vice versa), they are simply encapsulated in X.25 packets and sent over the network.	Considerando que pacotes IP não podem ser mapeados em X.25 e vice-versa, eles são simplesmente encapsulados em pacotes X.25 e enviados pela rede.

obtendo resultados similares ao da figura 4. Implementamos dois modelos com “filosofias” diferentes.

Nosso primeiro modelo é uma versão do trabalho de Gale e Church [26] incrementado com idéias de Kay e Röscheisen [27], o qual parte de um par de textos com os limites sentenças marcados e que são traduções mútuas e, através de scores de probabilidade e programação dinâmica, estima quais sentenças são traduções mútuas.

Assim como o trabalho de Gale e Church [26], nosso modelo faz o alinhamento de sentenças em duas etapas, alinhando os parágrafos inicialmente e, em seguida, as sentenças dentro desses.

Na fase de treinamento, nosso modelo aproveita as idéias de Kay e Röscheisen [27] para alinhar parágrafos, as quais sugerem que partes dos textos que se alinham estão em posições semelhantes nos textos e possuem palavras equivalentes (cognatos, nomes e termos técnicos). Essa idéias restringem as opções de procura, buscando por parágrafos com palavras cognatas e em posições semelhantes nos textos. Com os parágrafos alinhados, estimamos os parâmetros necessários à fase de aplicação.

O alinhamento das sentenças dentro dos parágrafos, ocorre na fase de aplicação e usa as premissas de Gale e Church [26], nas quais as sentenças longas são traduzidas para sentenças longas, enquanto que sentenças curtas são traduzidas para sentenças curtas e que as sentenças não são traduzidas necessariamente uma para uma, como o exemplo da figura 4.

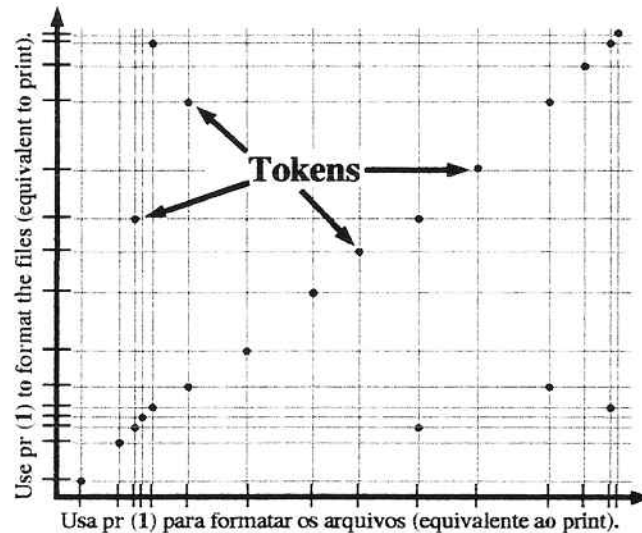
Assim, através de uma medida de distância, que leva em conta o comprimento das sentenças e sua variação, aplicada a um modelo de programação dinâmica, nosso modelo, tal e qual o de Gale e Church [26], verifica qual das seguintes situações de alinhamento produz o menor distância entre as sentenças dadas:

- Incluir uma sentença na tradução;
- Não traduzir uma sentença do original;
- Traduzir uma sentença do original para uma sentença na tradução;
- Traduzir uma sentença do original para duas sentenças na tradução;
- Traduzir duas sentenças no original para uma sentença na tradução, e;
- Traduzir duas sentenças no original para duas sentenças na tradução.

O outro modelo que implementamos é baseado no trabalho de Melamed [3, 4, 5], o qual usa o conceito de bitexto introduzido por Harris em 1988, apud Melamed [3, 4, 5].

A idéia do bitexto é transformar um texto original e sua tradução em um plano cartesiano. Para tanto, cada texto produz um eixo, onde as coordenadas indicam aproximadamente a posição central, em caracteres, de cada palavra. Os pontos no bitexto formados pelas coordenadas de palavras que são traduções mútuas são os tokens. A figura 5 traz um exemplo de bitexto. Contudo, nem todos os tokens indicam as traduções corretas para a coordenada original. Às traduções corretas chamamos TPCs (True Point of Correspondence).

Figura 5: exemplo de bitexto



Nosso modelo, assim como o de Melamed [3, 4, 5], usa o fato que os TPCs produzem funções bijetoras para conseguir um pré-alinhamento dos textos, explorando as características dessas funções:

- *Linearidade*: as sequências de TPCs, chamadas correntes, formam arranjos aproximadamente lineares, assim é possível interpolar uma reta de mínimos quadrados por elas com bastante precisão;
- *Inclinação constante*: a inclinação de uma corrente é muito próxima da inclinação do bitexto, definida pela reta que passa por seus dois extremos, e apresenta sempre o mesmo sentido, ou seja, correntes são monotônicas;
- *Bijetividade*: não há dois TPCs em uma corrente compartilhando as mesmas coordenadas.

Assim, para produzir o pré-alinhamento, é necessário encontrar os TPCs, o que se faz em duas etapas: encontrar os tokens e escolher os TPCs entre os tokens.

Para encontrar os tokens, assim como Melamed [3, 4, 5], marcamos uma área retangular, paralela aos eixos do bitexto, e procuramos por cognatos dentro dessa área, aproveitando o fato que cognatos são palavras com grafia⁵ semelhante em idiomas diferentes.

Uma vez que temos os tokens, formamos conjuntos com eles, através de uma heurística, e interpolamos retas de mínimos quadrados por esses conjuntos, escolhendo aquela que apresentar menor dispersão em seus pontos e menor desvio em relação à inclinação do bitexto. Os tokens da melhor reta são os TPCs que procuramos.

Definimos uma nova área retangular cujo vértice superior direito seja o TPC com maiores valores para as coordenadas e repetimos o processo de encontrar tokens e escolher os TPCs. Fazemos isso até o final dos eixos.

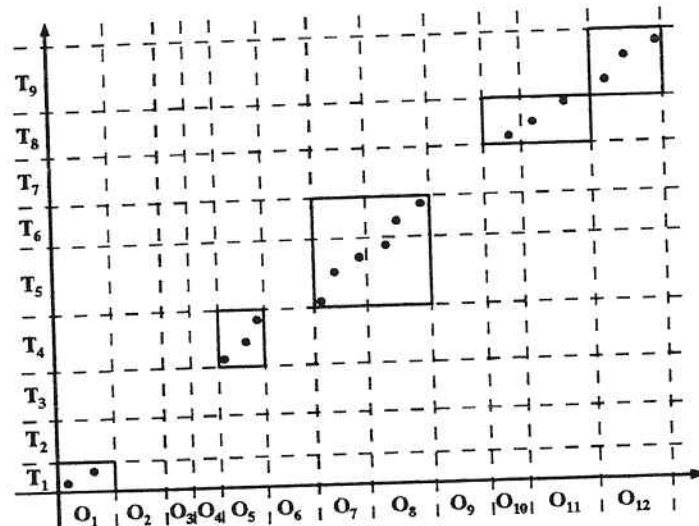
Ao final dos eixos, passamos à fase de alinhar as sentenças. Para tanto, temos textos com os limites de sentenças e TPCs identificados, formando algo similar à figura 6, que mostra as sentenças de um bitexto ($O_1, O_2, O_3, \dots, T_1, T_2, T_3, \dots$) com os TPCs (pontos escuros) e limites das sentenças (linhas tracejadas) anotados. O cruzamento desses limites forma células no espaço do bitexto que equivalem a intersecção de um par de sentenças.

Para alinhar as sentenças, cada vez que o modelo encontra um bloco que não forma um retângulo, faz o fechamento transitivo dos elementos do bloco de modo a conseguir um retângulo. Por exemplo, na figura 6 há TPCs em $\langle O_7, T_5 \rangle$, $\langle O_8, T_5 \rangle$ e $\langle O_8, T_6 \rangle$, assim, o modelo insere $\langle O_7, T_6 \rangle$.

Essa estratégia de “desenhar retângulos” produz segmentos de textos alinhados e não sobrepostos, mas não necessariamente com uma sentença cada, sendo que nesses casos, onde há blocos alinhados com mais de uma sentença cada, aplicamos o modelo de alinhamento de Gale e Church [26] para conseguir o alinhamento a nível de sentença, tal e qual Melamed [4].

⁵Na verdade cognatos são palavras com mesmo significado e grafia e/ou pronúncia semelhantes em idiomas diferentes. Os que apresentam grafia semelhantes são cognatos ortográficos, como paralelo (português) e parallel (inglês), enquanto que os outros, que apresentam pronúncia semelhante, são cognatos fonéticos, como /sys/tem/ (inglês) - /sys/te/mu (japonês).

Figura 6: modelo geométrico de alinhamento de sentenças



2.3 Alinhar Palavras

O último passo do processo é descobrir quais grupos de palavra no texto original se correspondem a quais grupos de palavras no texto traduzido.

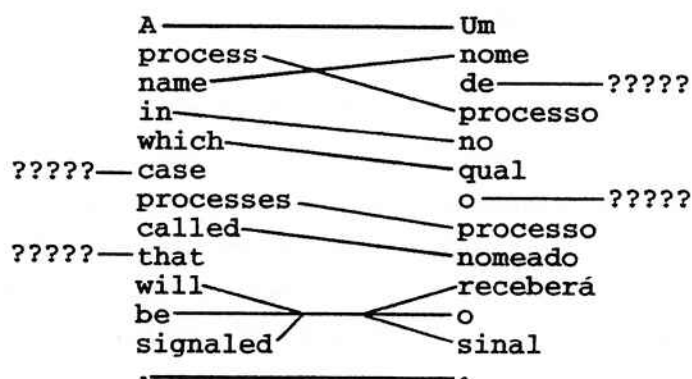
Há muitas maneiras de encontrar as correspondências entre as palavras, umas mais simples e diretas, outras mais elaboradas, tudo depende da aplicação. Assim, na última etapa, encontraremos as correspondências entre as palavras desenvolvendo uma aplicação em machine translation que usa sentenças alinhadas como entrada.

Os problemas para descobrir as correspondências entre as palavras são similares aos do alinhamento de sentenças. Os tradutores humanos não traduzem exatamente cada palavra do texto original, seja por motivos estéticos ou lingüísticos, com agravante de também não manterem a ordem das palavras.

A figura 7 apresenta uma sentença com suas correspondências entre as palavras assinaladas. Essas correspondências chamaremos de *alinhamento*.

Podemos ver na figura 7 alguns dos problemas inerentes ao alinhamento de palavras, como palavras sem a devida correspondência, indicadas por ?????,

Figura 7: alinhamento de palavras



mudança na ordem das palavras, como no par *process/processo*, e o caso onde um grupo de palavras em um dos textos se alinha com um grupo de palavras no outro texto.

Para mostrar uma aplicação onde temos alinhamentos de palavras, implementamos os modelo 1 e 2 desenvolvidos no trabalho de Brown et alli [20].

Em linhas gerais, tanto o modelo 1, como o modelo 2, de Brown et alli [20] compilam, a partir de um conjunto de pares de sentenças, uma lista contendo as palavras do texto original e suas traduções com a probabilidade de ocorrência.

A diferença básica entre os modelos 1 e 2 é que o último leva em consideração a posição das palavras nas sentenças, enquanto que o primeiro não.

A forma de funcionamento dos modelos para compilação das listas consiste em ler um par de sentenças e contabilizar as ocorrências de pares de palavras. Após a contabilização, os modelos calculam as probabilidades de ocorrência dos pares de palavras, e no caso do modelo 2, também das posições dessas.

Como todo o cálculo das probabilidades é iterativo, o processo é repetido

até que as diferenças entre as probabilidades da interação anterior e atual esteja menor que o erro desejado.

3 Experimentos

Elaboramos experimentos visando estimar os desempenhos geral e com relação ao idioma português para cada um dos modelos.

Basicamente, os experimentos para estimar o desempenho geral verificam o comportamento do modelo quando variamos a quantidade de texto para treinamento, assim iniciamos o experimento treinando o modelo com uma certa quantidade de texto, na sequência, submetemos um texto amostra ao modelo e, por fim, coletamos e medimos os resultados. Repetimos o experimento reduzindo a quantidade de texto para treinamento à metade e submetendo o mesmo texto amostra.

Os experimentos para estimar o desempenho com relação ao idioma português, envolvem treinar o modelo separadamente para cada par de idiomas⁶ (inglês-espanhol, inglês-português e português-espanhol) usando textos de características semelhantes no treinamento, submeter ao modelo um texto amostra no respectivo par de idiomas e, por fim, coletar e comparar os resultados.

4 Conclusões

A figura 1 traz um diagrama do processo automático de tratamento de textos indicando a etapa do processo, os modelos envolvidos, suas entradas e saídas e as contribuições que acrescentamos.

Nosso primeiro objetivo, elaborar um processo automático para processamento de linguagem natural, foi atingido através da elaboração de um processo em três etapas.

⁶No caso do modelo de identificação de limites de sentenças, usamos os idiomas individualmente.

Na etapa de identificação dos limites das sentenças. Entramos com um texto plano, como o da figura 8, sem qualquer anotação e saímos com um texto com os limites das sentenças identificados e etiquetados, como o da figura 9, onde os limites de sentenças estão marcados com “YYY”.

Mostramos que nossa implementação para identificação de limites de sentenças tem desempenho semelhante ao modelo original de Reynar e Ratnaparkhi [12], e que funciona tão bem para o idioma português como para os idiomas inglês e espanhol.

Na etapa de alinhamento das sentenças. Entramos com um par de textos, sendo o primeiro texto o original e o segundo sua tradução, ambos com os limites de sentenças anotados, como na figura 9. Dessa etapa saímos com um corpus bilíngue formado por pares de sentenças, cuja primeira sentença pertence ao texto original e a segunda à tradução, conforme a figura 10.

Mostramos que os dois modelos de alinhamento de sentenças implementados têm desempenho semelhante e independente do par de idiomas utilizado, funcionando tão bem para os corpora com o idioma português como para o corpus inglês-espanhol.

Por fim, na etapa do alinhamento das palavras. Entramos com um corpus bilíngue, como o da figura 10 e saímos com algum resultado aplicado ao processamento de linguagem natural como dicionários estatísticos e bases de dados para machine translation, como o dicionário da figura 11.

Nessa etapa, mostramos que os pares de idiomas empregados não influenciam no desempenho dos modelos, ou seja, os modelos funcionam para corpus inglês-português tão bem como para o corpus inglês-espanhol. Mostramos, também, que os modelos 1 e 2 de Brown et alli [20] têm desempenho semelhante.

Mostramos também que o uso de corpora formados por traduções do mesmo original apresenta resultados inferiores aos corpora formados pelo original e sua tradução.

Além disso, trouxemos algumas contribuições aos modelos implementados como :

- Inclusão de um teste estatístico para verificar a significância dos con-

Figura 8: trecho de texto plano

Original

...
Description
chdir changes the current directory to that specified in path.
fchdir is identical to chdir, only that the directory is given as an open file descriptor.
Return Value
On success, zero is returned. On error, -1 is returned, and errno is set appropriately.
Errors
Depending on the file system, other errors can be returned. The more general errors for chdir are listed below:
EFAULT path points outside your accessible address space.
ENAMETOOLONG
path is too long.
ENOENT The file does not exist.
ENOMEM Insufficient kernel memory was available.
ENOTDIR A component of path is not a directory.
...

Tradução

...
Descrição
O chdir muda o diretório corrente para o especificado na rota.
O fchdir é idêntico ao chdir, diferindo somente que o diretório é dado como um descritor de arquivo aberto.
Valor Retornado
Quando bem sucedido, o valor retornado é zero. Quando ocorre erro, o -1 é retornado e o erro é definido adequadamente.
Erros
Dependendo do sistema de arquivos, outros erros podem ser retornados. Outros erros gerais para o chdir estão listados abaixo:
EFAULT A rota aponta para um espaço fora do seu endereço acessível.
ENAMETOOLONG
A rota é muito comprida.
ENOENT O arquivo não existe.
ENOMEM A memória do cerne (kernel) foi insuficiente.
ENOTDIR Um componente da rota não é um diretório.
...

Figura 9: marcações dos limites de sentenças

Original

...
 Description
 chdir changes the current directory to that specified in path.YYY
 fchdir is identical to chdir, only that the directory is given as an open file descriptor.YYY
 Return Value
 On success, zero is returned.YYY On error, -1 is returned, and errno is set appropriately.YYY
 Errors
 Depending on the file system, other errors can be returned.YYY The more general errors for
 chdir are listed below:
 EFAULT path points outside your accessible address space.YYY
 ENAMETOOLONG
 path is too long.YYY
 ENOENT The file does not exist.YYY
 ENOMEM Insufficient kernel memory was available.YYY
 ENOTDIR A component of path is not a directory.YYY
 ...

Tradução

...
 Descrição
 O chdir muda o diretório corrente para o especificado na rota.YYY
 O fchdir é idêntico ao chdir, diferindo somente que o diretório é dado como um descritor de
 arquivo aberto.YYY
 Valor Retornado
 Quando bem sucedido, o valor retornado é zero.YYY Quando ocorre erro, o -1 é retornado e o
 errno é definido adequadamente.YYY
 Erros
 Dependendo do sistema de arquivos, outros erros podem ser retornados.YYY Outros erros gerais
 para o chdir estão listados abaixo:
 EFAULT A rota aponta para um espaço fora do seu endereço acessível.YYY
 ENAMETOOLONG
 A rota é muito comprida.YYY
 ENOENT O arquivo não existe.YYY
 ENOMEM A memória do kernel (kernel) foi insuficiente.YYY
 ENOTDIR Um componente da rota não é um diretório.YYY
 ...

Figura 10: trecho de texto alinhado

```

...
Description
Descrição

chdir changes the current directory to that specified in path.
O chdir muda o diretório corrente para o especificado na rcta.

fchdir is identical to chdir, only that the directory is given as an open file descriptor.
O fchdir idêntico ao chdir, diferindo somente que o diretório é dado como um descritor de arquivo aberto.

Return Value
Valor Retornado

On success, zero is returned. On error, -1 is returned, and errno is set appropriately.
Quando bem sucedido, o valor retornado é zero. Quando ocorre erro, o -1 é retornado e o errno é definido
adequadamente.

Errors
Erros

Depending on the file system, other errors can be returned. The more general errors for chdir are listed
below.
Dependendo do sistema de arquivos, outros erros podem ser retornados. Outros erros gerais para o chdir
estão listados abaixo:

EFAULT path points outside your accessible address space.
EFAULT A rota aponta para um espaço fora do seu endereço acessível.

ENAMETOOLONG
ENAMETOOLONG

path is too long.
A rota é muito comprida.

ENOENT The file does not exist.
ENOENT O arquivo não existe.

ENOMEM Insufficient kernel memory was available.
ENOMEM A memória do kernel (kernel) foi insuficiente.

ENOTDIR A component of path is not a directory.
ENOTDIR Um componente da rota não é um diretório.
...

```

Figura 11: trecho do dicionário estatístico inglês-português

original	: tradução	probabilidade indiv acum
successfully	: sucesso	0.16 0.16
successfully	: com	0.07 0.24
successfully	: problemas	0.05 0.29
successfully	: sem	0.05 0.34
successfully	: uma	0.04 0.38
successfully	: normalmente	0.04 0.42
successfully	: na	0.04 0.46
successfully	: help	0.04 0.50
successfully	: conclui	0.04 0.54
successfully	: -	0.03 0.57
successfully	: tradução	0.03 0.60
successfully	: portuguesa	0.03 0.63
successfully	: língua	0.03 0.66
successfully	: revisão	0.03 0.68
successfully	: mensagem	0.03 0.71
such	: por	0.25 0.25
such	: tais	0.19 0.44
such	: indicador	0.19 0.63
such	: estejam	0.19 0.82
suf	: suf	0.73 0.73
sufficient	: suficientes	0.14 0.14
sufficient	: ferramentas	0.14 0.29
sufficient	: fornecer	0.14 0.43
sufficient	: estas	0.14 0.57
sufficient	: necessário	0.14 0.71
suffix	: sufixo	0.44 0.44
suffix	: -	0.28 0.72
suffixes	: sufixos	0.91 0.91

textos em relação ao sinal de pontuação no modelo de identificação de limites de sentenças proposto por Reynar e Ratnaparkhi [12];

- Automatização do modelo de identificação de limites de sentenças pelo acréscimo de uma heurística de coleta de contextos que dispensa anotações nos textos;
- Elaboração de um modelo de alinhamento de sentenças mais robusto, baseado no modelo proposto por Gale e Church [26], através da inclusão das idéias de Kay e Röscheisen [27] e de um modelo de identificação de cognatos;
- Substituição, no modelo de alinhamento de sentenças proposto por Melamed [4], do modelo de identificação de cognatos original por outro modelo de maior robustez por nós desenvolvido.

5 Trabalhos Futuros

Durante o desenvolvimento do processo de processamento de linguagem natural, deparamo-nos com diversas soluções para os problemas que tínhamos em mãos, mas que por questões de prazo e espaço precisamos abandonar. O mesmo ocorreu com as implementações para os modelos que escolhemos. Sempre há espaço para melhorias.

No caso da primeira etapa do processo, identificação dos limites de sentenças, há melhorias que efetuaremos quanto às características, pois alguns erros que o modelo comete, como em nomes com abreviações (por exemplo: Bruce F. Wayne), apontando um sinal de pontuação onde não há, ou vice-versa, tem sua solução através da implementação de novas características.

Por outro lado, há outros modelos, como os desenvolvidos por Daelermans[17], e os modelos de Palmer e Hearst [11], muito interessantes e que despertam o interesse para sua implementação.

Quanto a segunda etapa, o alinhamento de sentenças, há inúmeros desafios nos modelos implementados, como melhorar o modelo de identificação de

cognatos para reduzir o tempo quadrático de execução e a sensibilidade do modelo GSA à segmentação de sentenças.

Além desses, há o modelo proposto por Chen [29], que parece bastante robusto a omissões nos textos e por isso também merecerá atenção.

A terceira etapa, o alinhamento de palavras, traz como desafios futuros a implementação dos outros modelos propostos por Brown et alli [20], pois estes outros modelos são mais completos e permitem tratar palavras que traduzem ou são traduzidas para mais de uma palavra.

Outros trabalhos relacionados à etapa e que despertam interesse são os modelos propostos por Fung [30], Melamed [10] e Nirenburg, Beale e Domashnev [34]. Todos tratam o alinhamento de palavras de forma diferente e com resultados interessantes.

Por fim, continuaremos nossas pesquisas sobre a extração e montagem de corpora, vez que os corpora que produzimos para este trabalho são pequenos e apresentam alguns erros.

Referências

- [1] Russel, S.; Norvig, P. *Artificial Intelligence : a modern approach*. New Jersey: Prentice-Hall, Inc., 1995.
- [2] Melamed, I. Dan *Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons*. Proceedings of the Third Workshop on Very Large Corpora, Boston, MA, 1995. Disponível em: <<http://citeseer.nj.nec.com/melamed95automatic.html>>. Acesso em: 10/03/2002.
- [3] Melamed, I. Dan *Automatic Detection of Omission in Translations*. Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Dinamarca, 1996. Disponível em: <<http://citeseer.nj.nec.com/149949.html>>. Acesso em: 10/03/2002.
- [4] Melamed, I. Dan *A Geometric Approach to Mapping Bilingual Correspondence*. Proceedings of the First Conference on Empirical Methods

- in Natural Language Processing, Philadelphia, PA, 1996. Disponível em: <<http://citeseer.nj.nec.com/182431.html>>. Acesso em: 10/03/2002.
- [5] Melamed, I. Dan *Porting SIMR to new language pairs*. Institute for Research in Cognitive Science Technical Report 96-26. University of Pennsylvania. 1996. Disponível em: <<http://citeseer.nj.nec.com/melamed96porting.html>>. Acesso em: 10/03/2002.
- [6] Melamed, I. Dan *A portable algorithm for mapping bitext correspondence*. In Proceedings of the 35th Conference of the Association for Computational Linguistics. 1997. Disponível em: <<http://citeseer.nj.nec.com/melamed97portable.html>>. Acesso em: 10/03/2002.
- [7] Melamed, I. Dan *A word-to-word model of translational equivalence*. In 35th Conference of the Association for Computational Linguistics (ACL'97), pages 490-497, Madrid, 1997. Disponível em: <<http://citeseer.nj.nec.com/article/melamed97wordtoword.html>>. Acesso em: 10/03/2002.
- [8] Melamed, I. Dan *Models of Co-occurrence*. Institute for Research in Cognitive Science Technical Report #98-05. 1998. Disponível em: <<http://citeseer.nj.nec.com/melamed98models.html>>. Acesso em: 10/03/2002.
- [9] Melamed, I. Dan *Empirical Methods for MT Lexicon Development*. AMTA. p18-30. 1998. Disponível em : <<http://citeseer.nj.nec.com/152353>>. Acesso em : 13/03/2002.
- [10] Melamed, I. Dan *Models of Translational Equivalence among Words*. Computational Linguistics, v. 26 n. 2 p.221-249. 2000.
- [11] Palmer, D.; Hearst, M. *Adaptive Multilingual Sentence Boundary Disambiguation*. Computational Linguistics, v.23 n.2 p.241-269. 1997. Disponível em : <<http://citeseer.nj.nec.com/palmer97adaptive.html>>. Acesso em : 10/03/2002.

- [12] Reynar, J.; Ratnaparkhi, A. *A Maximum Entropy Approach to Identifying Sentence Boundaries*. In Proceedings of the Fifth Conference on Applied Natural Language Processing, pages 16–19, Washington D.C. 1997. Disponível em : <<http://citeseer.nj.nec.com/reynar97maximum.html>>. Acesso em : 10/03/2002.
- [13] Ratnaparkhi, A. *A simple introduction to maximum entropy models for natural language processing*. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania. 1997. Disponível em : <<http://citeseer.nj.nec.com/128751.html>>. Acesso em : 10/03/2002.
- [14] Ratnaparkhi, A. *Maximum Entropy Models For Natural Language Ambiguity Resolution*. 1998. 147p. Tese (Doutorado) - University of Pennsylvania, Philadelphia. Disponível em : <<http://citeseer.nj.nec.com/ratnaparkhi98maximum.html>>. Acesso em : 10/03/2002.
- [15] Brill, E. *A Report of Recent Progress in Transformation-Based Error-Driven Learning*. Proceedings of the Twelfth National Conference on Artificial Intelligence, pages 722–727, 1994. Disponível em : <<http://citeseer.nj.nec.com/brill94report.html>>. Acesso em : 10/03/2002.
- [16] Daelemans, W. et alli *MBT: A Memory-Based Part of Speech Tagger-Generator*. In Ejerhed, Eva and Dagan, Ido, editor, Proceedings of the Fourth Workshop on Very Large Corpora. p.14– 27. 1996. Disponível em : <<http://citeseer.nj.nec.com/daelemans96mbt.html>>. Acesso em : 10/03/2002.
- [17] Daelemans, W.; van den Bosch, A. e Weijters, T. *IGTree: Using Trees for Compression and Classification in Lazy Learning Algorithms*. Artificial Intelligence Review 11, 407–423. 1997. Disponível em : <<http://citeseer.nj.nec.com/daelemans97igtree.html>>. Acesso em : 10/03/2002.

- [18] Daelemans, W.; Buchholz, S.; Veenstra, J. *Memory-Based Shallow Parsing*. In proceedings of CoNLL, Bergen, N. submitted. 1999. Disponível em : <<http://citeseer.nj.nec.com/daelemans99memorybased.html>>. Acesso em : 10/03/2002.
- [19] Brown, P. et alli *A Statistical Approach to Machine Translation*. Computational Linguistics, v.16 n.2 p.79-85. 1990.
- [20] Brown, P. et alli *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, v.19 n.2 p.263-311. 1993.
- [21] Dunning, T. *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics, v.19 n.1 p.60-74. 1993.
- [22] Macklovitch, E.; Hannan, M. *Line 'Em Up: Advances in Alignment Technology and Their Impact On Translation Support Tools*. 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada, 1996. Disponível em: <<http://citeseer.nj.nec.com/macklovitch96line.html>>. Acesso em: 10/03/2002.
- [23] Simard, M.; Plamondon, P. *Bilingual Sentence Alignment: Balancing Robustness and Accuracy*. In Proc. Conference of the Association for Machine Translation in the Americas (AMTA). 1996. Disponível em : <<http://citeseer.nj.nec.com/michel96bilingual.html>>. Acesso em : 10/03/2002.
- [24] Church, K.; Mercer, R. *Introduction to the Special Issue on Computational Linguistics Using Large Corpora*. Computational Linguistics, v.19 n.1 p.01-25. 1993.
- [25] Gale, W.; Church, K. *Identifying word correspondance in parallel text*. Proceedings of the DARPA NLP Workshop. 1991. Disponível em : <<http://citeseer.nj.nec.com/gale91identifying.html>>. Acesso em : 10/03/2002.

- [26] Gale, W.; Church, K. *A Program for Aligning Sentences in Bilingual Corpora*. Computational Linguistics, v.19 n.1 p.75-102. 1993.
- [27] Kay, M.; Röscheisen, M. *Text-Translation Alignment*. Computational Linguistics, v.19 n.1 p.121-142. 1993.
- [28] Chen, S. *Aligning sentences in bilingual corpora using lexical information*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL). 1993. Disponível em : <<http://citeseer.nj.nec.com/stanley93aligning.html>>. Acesso em : 10/03/2002.
- [29] Chen, S. *Building Probabilistic Models for Natural Language*. 1996. 151p. Tese (Doutorado) - Harvard University, Cambridge, Massachusetts. Disponível em : <<http://citeseer.nj.nec.com/chen96building.html>>. Acesso em : 10/03/2002.
- [30] Fung, P. *Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus*. Proceedings of the Third Workshop on Very Large Corpora, Boston, MA, 1995. Disponível
- [31] Kinoshita, J. *Aspectos de Implementação de uma Ferramenta de Auxílio à Tradução Inglês-Português*. 1997. 163p. Tese (Doutorado) - Universidade de São Paulo, São Paulo, São Paulo.
- [32] Smadja, F.; McKeown, K. e Hatzivassiloglou, V. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Computational Linguistics, v.22 n.1 p.01-38. 1996.
- [33] Utsuro, T. et alli *Bilingual Text Matching using Bilingual Dictionary and Statistics*. In 15th COLING, pp. 1076-1082. 1994. Disponível em : <<http://citeseer.nj.nec.com/utsuro94bilingual.html>>. Acesso em : 10/03/2002.
- [34] Nirenburg, S.; Beale, S. e Domashnev, C. *A Full-Text Experiment in Example-Based Machine Translation*. School of Computer Science. In

- Proceedings of the International Conference on New Methods in Language Processing, Manchester, England, pp. 78–87. 1994. Disponível em : <http://citeseer.nj.nec.com/nirenburg94fulltext.html>. Acesso em : 10/03/2002.
- [35] Wu, D. *Aligning a parallel English-Chinese corpus statistically with lexical criteria*. In Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, 80–87, Las Cruces, New Mexico. 1994. Disponível em : <http://citeseer.nj.nec.com/wu94aligning.html>. Acesso em : 10/03/2002.
- [36] Wu, D. *Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora*. In IJCAI'95, volume 2, pages 1328#1335, Montr#eal (Canada), Aug. 1995. Morgan Kaufmann. 1995. Disponível em : <http://citeseer.nj.nec.com/wu95stochastic.html>. Acesso em : 10/03/2002.
- [37] Wu, D. *Trainable coarse bilingual grammars for parallel text bracketing*. 1995. Disponível em : <http://citeseer.nj.nec.com/wu95trainable.html>. Acesso em : 10/03/2002.
- [38] Sang, E. *Aligning the Scania Corpus*. Technical Report, Department of Linguistics, University of Uppsala, 1996. Disponível em : <http://citeseer.nj.nec.com/kim96aligning.html>. Acesso em : 10/03/2002.
- [39] Covington, M. An Algorithm to Align Words for Historical Comparison. *Computational Linguistics*, v.22 n.4 p.481-496. 1996.
- [40] Knight, K.; Graehl, J. Machine Transliteration. *Computational Linguistics*, v.24 n.4 p.599-612. 1998.
- [41] Mason, R.; Gunst, R.; Hess, J. *Statistical Design and Analysis of Experiments: with applications to engineering and science*. New York. John Wiley & Sons, Inc. 1989.

- [42] Proakis, J.; Salehi, M. *Communication System Enginnering*. New Jersey. Prentice Hall Inc. 1994.
- [43] Mood, A.; Graybill, F.; Boes, D. *Introduction to the Theory of Statistics*. Singapore. McGraw-Hill Book Company. 1974.

BOLETINS TÉCNICOS - TEXTOS PUBLICADOS

- BT/PCS/9301 - Interligação de Processadores através de Chaves Ômicron - GERALDO LINO DE CAMPOS, DEMI GETSCHKO
- BT/PCS/9302 - Implementação de Transparência em Sistema Distribuído - LUÍSA YUMIKO AKAO, JOÃO JOSÉ NETO
- BT/PCS/9303 - Desenvolvimento de Sistemas Especificados em SDL - SIDNEI H. TANO, SELMA S. S. MELNIKOFF
- BT/PCS/9304 - Um Modelo Formal para Sistemas Digitais à Nível de Transferência de Registradores - JOSÉ EDUARDO MOREIRA, WILSON VICENTE RUGGIERO
- BT/PCS/9305 - Uma Ferramenta para o Desenvolvimento de Protótipos de Programas Concorrentes - JORGE KINOSHITA, JOÃO JOSÉ NETO
- BT/PCS/9306 - Uma Ferramenta de Monitoração para um Núcleo de Resolução Distribuída de Problemas Orientado a Objetos - JAIME SIMÃO SICHMAN, ELERI CARDOSO
- BT/PCS/9307 - Uma Análise das Técnicas Reversíveis de Compressão de Dados - MÁRIO CESAR GOMES SEGURA, EDIT GRASSIANI LINO DE CAMPOS
- BT/PCS/9308 - Proposta de Rede Digital de Sistemas Integrados para Navio - CESAR DE ALVARENGA JACOBY, MOACYR MARTUCCI JR.
- BT/PCS/9309 - Sistemas UNIX para Tempo Real - PAULO CESAR CORIGLIANO, JOÃO JOSÉ NETO
- BT/PCS/9310 - Projeto de uma Unidade de Matching Store baseada em Memória Paginada para uma Máquina Fluxo de Dados Distribuído - EDUARDO MARQUES, CLAUDIO KIRNER
- BT/PCS/9401 - Implementação de Arquiteturas Abertas: Uma Aplicação na Automação da Manufatura - JORGE LUIS RISCO BECERRA, MOACYR MARTUCCI JR.
- BT/PCS/9402 - Modelamento Geométrico usando do Operadores Topológicos de Euler - GERALDO MACIEL DA FONSECA, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/9403 - Segmentação de Imagens aplicada a Reconhecimento Automático de Alvos - LEONCIO CLARO DE BARROS NETO, ANTONIO MARCOS DE AGUIRRA MASSOLA
- BT/PCS/9404 - Metodologia e Ambiente para Reutilização de Software Baseado em Composição - LEONARDO PUJATTI, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/9405 - Desenvolvimento de uma Solução para a Supervisão e Integração de Células de Manufatura Discreta - JOSÉ BENEDITO DE ALMEIDA, JOSÉ SIDNEI COLOMBO MARTINI
- BT/PCS/9406 - Método de Teste de Sincronização para Programas em ADA - EDUARDO T. MATSUDA, SELMA SHIN SHIMIZU MELNIKOFF
- BT/PCS/9407 - Um Compilador Paralelizante com Detecção de Paralelismo na Linguagem Intermediária - HSUEH TSUNG HSIANG, LÍRIA MATSUMOTO SAITO
- BT/PCS/9408 - Modelamento de Sistemas com Redes de Petri Interpretadas - CARLOS ALBERTO SANGIORGIO, WILSON V. RUGGIERO
- BT/PCS/9501 - Síntese de Voz com Qualidade - EVANDRO BACCI GOUVÊA, GERALDO LINO DE CAMPOS
- BT/PCS/9502 - Um Simulador de Arquiteturas de Computadores "A Computer Architecture Simulator" - CLAUDIO A. PRADO, WILSON V. RUGGIERO
- BT/PCS/9503 - Simulador para Avaliação da Confiabilidade de Sistemas Redundantes com Reparo - ANDRÉA LUCIA BRAGA, FRANCISCO JOSÉ DE OLIVEIRA DIAS
- BT/PCS/9504 - Projeto Conceitual e Projeto Básico do Nível de Coordenação de um Sistema Aberto de Automação, Utilizando Conceitos de Orientação a Objetos - NELSON TANOMARU, MOACYR MARTUCCI JUNIOR
- BT/PCS/9505 - Uma Experiência no Gerenciamento da Produção de Software - RICARDO LUIS DE AZEVEDO DA ROCHA, JOÃO JOSÉ NETO
- BT/PCS/9506 - Método OO - Método de Desenvolvimento de Sistemas Orientado a Objetos: Uma Abordagem Integrada à Análise Estruturada e Redes de Petri - KECHI HIRAMA, SELMA SHIN SHIMIZU MELNIKOFF
- BT/PCS/9601 - MOOPP: Uma Metodologia Orientada a Objetos para Desenvolvimento de Software para Processamento Paralelo - ELISA HATSUE MORIYA HUZITA, LÍRIA MATSUMOTO SATO
- BT/PCS/9602 - Estudo do Espalhamento Brillouin Estimulado em Fibras Ópticas Monomodo - LUIS MEREGE SANCHES, CHARLES ARTUR SANTOS DE OLIVEIRA
- BT/PCS/9603 - Programação Paralela com Variáveis Compartilhadas para Sistemas Distribuídos - LUCIANA BEZERRA ARANTES, LÍRIA MATSUMOTO SATO
- BT/PCS/9604 - Uma Metodologia de Projeto de Redes Locais - TEREZA CRISTINA MELO DE BRITO CARVALHO, WILSON VICENTE RUGGIERO

- BT/PCS/9605 - Desenvolvimento de Sistema para Conversão de Textos em Fonemas no Idioma Português - DIMAS TREVIAN CHBANE, GERALDO LINO DE CAMPOS
- BT/PCS/9606 - Sincronização de Fluxos Multimídia em um Sistema de Videoconferência - EDUARDO S. C. TAKAHASHI, STEFANIA STIUBIENER
- BT/PCS/9607 - A importância da Completeza na Especificação de Sistemas de Segurança - JOÃO BATISTA CAMARGO JÚNIOR, BENÍCIO JOSÉ DE SOUZA
- BT/PCS/9608 - Uma Abordagem Paraconsistente Baseada em Lógica Evidencial para Tratar Exceções em Sistemas de Frames com Múltipla Herança - BRÁULIO COELHO ÁVILA, MÁRCIO RILLO
- BT/PCS/9609 - Implementação de Engenharia Simultânea - MARCIO MOREIRA DA SILVA, MOACYR MARTUCCI JÚNIOR
- BT/PCS/9610 - Statecharts Adaptativos - Um Exemplo de Aplicação do STAD - JORGE RADY DE ALMEIDA JUNIOR, JOÃO JOSÉ NETO
- BT/PCS/9611 - Um Meta-Editor Dirigido por Sintaxe - MARGARETE KEIKO IWAI, JOÃO JOSÉ NETO
- BT/PCS/9612 - Reutilização em Software Orientado a Objetos: Um Estudo Empírico para Analisar a Dificuldade de Localização e Entendimento de Classes - SELMA SHIN SHIMIZU MELNIKOFF, PEDRO ALEXANDRE DE OLIVEIRA GIOVANI
- BT/PCS/9613 - Representação de Estruturas de Conhecimento em Sistemas de Banco de Dados - JUDITH PAVÓN MENDONZA, EDIT GRASSIANI LINO DE CAMPOS
- BT/PCS/9701 - Uma Experiência na Construção de um Tradutor Inglês - Português - JORGE KINOSHITA, JOÃO JOSÉ NETO
- BT/PCS/9702 - Combinando Análise de "Wavelet" e Análise Entrópica para Avaliar os Fenômenos de Difusão e Correlação - RUI CHUO HUEI CHIOU, MARIA ALICE G. V. FERREIRA
- BT/PCS/9703 - Um Método para Desenvolvimento de Sistemas de Computacionais de Apoio a Projetos de Engenharia - JOSÉ EDUARDO ZINDEL DEBONI, JOSÉ SIDNEI COLOMBO MARTINI
- BT/PCS/9704 - O Sistema de Posicionamento Global (GPS) e suas Aplicações - SÉRGIO MIRANDA PAZ, CARLOS EDUARDO CUGNASCA
- BT/PCS/9705 - METAMBI-OO - Um Ambiente de Apoio ao Aprendizado da Técnica Orientada a Objetos - JOÃO UMBERTO FURQUIM DE SOUZA, SELMA S. S. MELNIKOFF
- BT/PCS/9706 - Um Ambiente Interativo para Visualização do Comportamento Dinâmico de Algoritmos - IZAURA CRISTINA ARAÚJO, JOÃO JOSÉ NETO
- BT/PCS/9707 - Metodologia Orientada a Objetos e sua Aplicação em Sistemas de CAD Baseado em "Features" - CARLOS CÉSAR TANAKA, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/9708 - Um Tutor Inteligente para Análise Orientada a Objetos - MARIA EMÍLIA GOMES SOBRAL, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/9709 - Metodologia para Seleção de Solução de Sistema de Aquisição de Dados para Aplicações de Pequeno Porte - MARCELO FINGUERMAN, JOSÉ SIDNEI COLOMBO MARTINI
- BT/PCS/9801 - Conexões Virtuais em Redes ATM e Escalabilidade de Sistemas de Transmissão de Dados sem Conexão - WAGNER LUIZ ZUCCHI, WILSON VICENTE RUGGIERO
- BT/PCS/9802 - Estudo Comparativo dos Sistemas da Qualidade - EDISON SPINA, MOACYR MARTUCCI JR.
- BT/PCS/9803 - The VIBRA Multi-Agent Architecture: Integrating Purposive Vision With Deliberative and Reactive Planning - REINALDO A. C. BIANCHI, ANNA H. REALI C. RILLO, LELIANE N. BARROS
- BT/PCS/9901 - Metodologia ODP para o Desenvolvimento de Sistemas Abertos de Automação - JORGE LUIS RISCO BECCERRA, MOACYR MARTUCCI JUNIOR
- BT/PCS/9902 - Especificação de Um Modelo de Dados Bitemporal Orientado a Objetos - SOLANGE NICE ALVES DE SOUZA, EDIT GRASSIANI LINO DE CAMPOS
- BT/PCS/9903 - Implementação Paralela Distribuída da Dissecção Cartesiana Aninhada - HILTON GARCIA FERNANDES, LIRIA MATSUMOTO SATO
- BT/PCS/9904 - Metodologia para Especificação e Implementação de Solução de Gerenciamento - SERGIO CLEMENTE, TEREZA CRISTINA MELO DE BRITO CARVALHO
- BT/PCS/9905 - Modelagem de Ferramenta Hipermídia Aberta para a Produção de Tutoriais Interativos - LEILA HYODO, ROMERO TORI
- BT/PCS/9906 - Métodos de Aplicações da Lógica Paraconsistente Anotada de Anotação com Dois Valores-LPA2v com Construção de Algoritmo e Implementação de Circuitos Eletrônicos - JOÃO I. DA SILVA FILHO, JAIR MINORO ABE
- BT/PCS/9907 - Modelo Nebuloso de Confiabilidade Baseado no Modelo de Markov - PAULO SÉRGIO CUGNASCA, MARCO TÚLIO CARVALHO DE ANDRADE
- BT/PCS/9908 - Uma Análise Comparativa do Fluxo de Mensagens entre os Modelos da Rede Contractual (RC) e Colisões Baseada em Dependências (CBD) - MÁRCIA ITO, JAIME SIMÃO SICHMAN
- BT/PCS/9909 - Otimização de Processo de Inserção Automática de Componentes Eletrônicos Empregando a Técnica de Times Assíncronos - CESAR SCARPINI RABAK, JAIME SIMÃO SICHMAN

- BT/PCS/9910 – MIISA – Uma Metodologia para Integração da Informação em Sistemas Abertos – HILDA CARVALHO DE OLIVEIRA, SELMA S. S. MELNIKOFF
- BT/PCS/9911 – Metodologia para Utilização de Componentes de Software: um estudo de Caso – KAZUTOSI TAKATA, SELMA S. S. MELNIKOFF
- BT/PCS/0001 – Método para Engenharia de Requisitos Norteado por Necessidades de Informação – ARISTIDES NOVELLI FILHO, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0002 – Um Método de Escolha Automática de Soluções Usando Tecnologia Adaptativa – RICARDO LUIS DE AZEVEDO DA ROCHA, JOÃO JOSÉ NETO
- BT/PCS/0101 – Gerenciamento Hierárquico de Falhas – JAMIL KALIL NAUFAL JR., JOÃO BATISTA CAMARGO JR.
- BT/PCS/0102 – Um Método para a Construção de Analisadores Morfológicos, Aplicado à Língua Portuguesa, Baseado em Autômatos Adaptativos – CARLOS EDUARDO DANTAS DE MENEZES, JOÃO JOSÉ NETO
- BT/PCS/0103 – Educação pela Web: Metodologia e Ferramenta de Elaboração de Cursos com Navegação Dinâmica – LUISA ALEYDA GARCIA GONZÁLEZ, WILSON VICENTE RUGGIERO
- BT/PCS/0104 – O Desenvolvimento de Sistemas Baseados em Componentes a Partir da Visão de Objetos – RENATA EVANGELISTA ROMARIZ RECCO, JOÃO BATISTA CAMARGO JÚNIOR
- BT/PCS/0105 – Introdução às Gramáticas Adaptativas – MARGARETE KEIKO IWAI, JOÃO JOSÉ NETO
- BT/PCS/0106 – Automação dos Processos de Controle de Qualidade da Água e Esgoto em Laboratório de Controle Sanitário – JOSÉ BENEDITO DE ALMEIDA, JOSÉ SIDNEI COLOMBO MARTINI
- BT/PCS/0107 – Um Mecanismo para Distribuição Segura de Vídeo MPEG – CÍNTIA BORGES MARGI, GRAÇA BESSAN, WILSON VICENTE RUGGIERO
- BT/PCS/0108 – A Dependence-Based Model for Social Reasoning in Multi-Agent Systems – JAIME SIMÃO SICHMAN
- BT/PCS/0109 – Ambiente Multilinguagem de Programação – Aspectos do Projeto e Implementação – APARECIDO VALDEMIR DE FREITAS, JOÃO JOSÉ NETO
- BT/PCS/0110 – LETAC: Técnica para Análise de Tarefas e Especificação de Fluxo de Trabalho Cooperativo – MARCOS ROBERTO GREINER, LUCIA VILELA LEITE FILGUEIRAS
- BT/PCS/0111 – Modelagem ODP para o Planejamento de Sistemas de Potência – ANIRIO SALLES FILHO, JOSÉ SIDNEI COLOMBO MARTINI
- BT/PCS/0112 – Técnica para Ajuste dos Coeficientes de Quantização do Padrão MPEG em Tempo Real – REGINA M. SILVEIRA, WILSON V. RUGGIERO
- BT/PCS/0113 – Segmentação de Imagens por Classificação de Cores: Uma Abordagem Neural – ALEXANDRE S. SIMÕES, ANNA REALI COSTA
- BT/PCS/0114 – Uma Avaliação do Sistema DSM Nautilus – MARIO DONATO MARINO, GERALDO LINO DE CAMPOS
- BT/PCS/0115 – Utilização de Redes Neurais Artificiais para Construção de Imagem em Câmara de Cintilação – LUIZ SÉRGIO DE SOUZA, EDITH RANZINI
- BT/PCS/0116 – Simulação de Redes ATM – HSU CHIH WANG CHANG, WILSON VICENTE RUGGIERO
- BT/PCS/0117 – Application of Monoprocessed Architecture for Safety Critical Control Systems – JOSÉ ANTONIO FONSECA, JORGE RADY DE ALMEIDA JR.
- BT/PCS/0118 – WebBee – Um Sistema de Informação via WEB para Pesquisa de Abelhas sem Ferrão – RENATO SOUSA DA CUNHA, ANTONIO MOURA SARAIVA
- BT/PCS/0119 – Parallel Processing Applied to Robot Manipulator Trajectory Planning – DENIS HAMILTON NOMIYAMA, LÍRIA MATSUMOTO SATO, ANDRÉ RIYUITI HIRAKAWA
- BT/PCS/0120 – Utilização de Padrão de Arquitetura de Software para a Fase de Projeto Orientado a Objetos – CRISITINA MARIA FERREIRA DA SILVA, SELMA SHIN SHIMIZU MELNIKOFF
- BT/PCS/0121 – Agilizando Aprendizagem por Reforço Através do uso de Conhecimento sobre o Domínio – RENÊ PEGORARO, ANNA H. REALI COSTA
- BT/PCS/0122 – Modelo de Segurança da Linguagem Java Problemas e Soluções – CLAUDIO MASSANORI MATAYOSHI, WILSON VICENTE RUGGIERO
- BT/PCS/0123 – Proposta de um Agente CNM para o Gerenciamento Web de um Backbone ATM – FERNANDO FROTA REDÍGOLO, TEREZA CRISTINA MELO DE BRITO CARVALHO
- BT/PCS/0124 – Um Método de Teste de software Baseado em Casos Teste – SÉRGIO RICARDO ROTTA, KECHI HIRAMA
- BT/PCS/0201 – A Teoria Nebulosa Aplicada a uma Bicicleta Ergométrica para Fisioterapia – MARCO ANTONIO GARMS, MARCO TÚLIO CARVALHO DE ANDRADE
- BT/PCS/0202 – Synchronization Constraints in a Concurrent Object Oriented Programming Model – LAÍS DO NASCIMENTO SALVADOR, LÍRIA MATSUMOTO SATO
- BT/PCS/0203 – Construção de um Ambiente de Dados sobre um Sistema de Arquivos Paralelos – JOSÉ CRAVEIRO DA COSTA NETO, LÍRIA MATSUMOTO SATO

- BT/PCS/0204 – Maestro: Um Middleware para Suporte a Aplicações Distribuídas Baseadas em Componentes de Software – CLÁUDIO LUÍS PEREIRA FERREIRA, JORGE LUÍS RISCO BECERRA
- BT/PCS/0205 – Sistemas de Automação dos Transportes (ITS) Descritos Através das Técnicas de Modelagem RM-OPD (ITU-T) e UML (OMG) – CLÁUDIO LUIZ MARTE, JORGE LUÍS RISCO BECERRA, JOSÉ SIDNEI COLOMBO
- BT/PCS/0206 – Comparação de Perfis de Usuários Coletados Através do Agente de Interface PersonalSearcher – GUSTAVO A. GIMÉNEZ LUGO, ANALÍA AMANDI, JAIME SIMÃO SICHMAN
- BT/PCS/0207 – Arquitetura Reutilizáveis para a Criação de Sistemas de Tutorização Inteligentes – MARCO ANTONIO FURLAN DE SOUZA, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0208 – Análise e Predição de Desempenho de Programas Paralelos em Redes de Estações de Trabalho – LIN KUAN CHING, LIRIA MATSUMOTO SATO
- BT/PCS/0209 – Previsões Financeiras Através de Sistemas Neuronebulosos – DANIEL DE SOUZA GOMES, MARCO TÚLIO CARVALHO DE ANDRADE
- BT/PCS/0210 – Proposta de Arquitetura Aberta de Central de Atendimento – ANA PAULA GONÇALVES SERRA, MOACYR MARTUCCI JÚNIOR
- BT/PCS/0211 – Alternativas de Implementação de Sistemas Nebulosos em Hardware – MARCOS ALVES PREDEBON, MARCO TÚLIO CARVALHO DE ANDRADE
- BT/PCS/0212 – Registro de Imagens de Documentos Antigos – VALGUIMA VICTORIA VIANA ODAKURA MARTINEZ, GERALDO LINO DE CAMPOS
- BT/PCS/0213 – Um Modelo de Dados Multidimensional – PEDRO WILLEMSSENS, JORGE RADY DE ALMEIDA JUNIOR
- BT/PCS/0214 – Autômatos Adaptativos no Tratamento Sintático de Linguagem Natural – CÉLIA YUMI OKANO TANIWAKI, JOÃO JOSÉ NETO
- BT/PCS/0215 – Fatores e Subfatores para Avaliação da Segurança em Software de Sistemas Críticos – JOÃO EDUARDO PROENÇA PÁSCOA, JOÃO BATISTA CAMARGO JÚNIOR
- BT/PCS/0216 – Derivando um Modelo de Projeto a Partir de um Modelo de Análise, com Base em Design Patterns J2EE – SERGIO MARTINS FERNANDES, SELMA SHIN SHIMIZU MELNIKOFF
- BT/PCS/0217 – Domínios Virtuais para Redes Móveis Ad Hoc: Um Mecanismo de Segurança – LEONARDO AUGUSTO MARTUCCI, TEREZA CRISTINA DE MELO BRITO CARVALHO
- BT/PCS/0218 – Uma Ferramenta para a Formulação de Consultas Baseadas em Entidades e Papéis – ANDRÉ ROBERTO DORETO SANTOS, EDIT GRASSIANI LINO CAMPOS
- BT/PCS/0219 – Avaliação de Performance de Arquiteturas para Computação de Alto Desempenho – KARIN STRAUSS, WILSON VICENTE RUGGIERO
- BT/PCS/0220 – BGLsim: Simulador de Sistema Completo para o Blue Gene/L – LUÍS HENRIQUE DE BARROS CEZE, WILSON VICENTE RUGGIERO
- BT/PCS/0221 – μ P: Uma Solução de Micropagamentos – PEDRO ANCONA LOPEZ MINDLIN, TEREZA CRISTINA MELO DE BRITO CARVALHO
- BT/PCS/0222 – Modelamento de Roteadores IP para Análise de Atraso – MARCELO BLANES, GRAÇA BRESSAN
- BT/PCS/0223 – Uma Biblioteca de Classes Utilizando Java 3D para o Desenvolvimento de Ambientes Virtuais Multi-Usuários – RICARDO NAKAMURA, ROMERO TORI
- BT/PCS/0224 – Interactive 3D Physics Experiments Through the Internet – ALEXANDRE CARDOSO, ROMERO TORI
- BT/PCS/0225 – Avaliação do Desempenho de Aplicações Distribuídas sob Duas Velocidades de Rede – AMILCAR ROSA PEREIRA, GERALDO LINO DE CAMPOS
- BT/PCS/0226 – Acompanhamento do Aprendizado do Aluno em Cursos a Distância através da WEB: Metodologias e Ferramentas – LUCIANA APARECIDA MARTINEZ ZAINA, GRAÇA BRESSAN
- BT/PCS/0227 – Um Ambiente Colaborativo para Simulação de Redes de Computadores – OSCAR DANTAS VILCACHAGUA, GRAÇA BRESSAN
- BT/PCS/0301 – Diretrizes para o Projeto de Base de Dados Distribuídas – PEDRO LUIZ PIZZIGATTI CORRÊA, JORGE RADY DE ALMEIDA JR.
- BT/PCS/0302 – Análise e Predição de Desempenho de Programas MPI em Redes de Estações de Trabalho – JEAN MARCOS LAINE, EDSON T. MIDORIKAWA
- BT/PCS/0303 – Padrões de Software para Tutores Inteligentes Cooperativos em Engenharia de Requisitos – MARIA EMILIA GOMES SOBRAL, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0304 – Performance Analysis and Prediction of Some MPI Communication Primitives – HÉLIO MARCI DE OLIVEIRA, EDSON TOSHIMI MIDORIKAWA
- BT/PCS/0305 – RM-ODP para Expressar o Licenciamento Nuclear – EDILSON DE ANDRADE BARBOSA, MOACYR MARTUCCI JÚNIOR
- BT/PCS/0306 – Modelo de Avaliação para Métricas de Software – VINICIUS DA SILVA ALMENDRA, KECHI HIRAMA

- BT/PCS/0307 – Análise de Confiabilidade de Sistemas Redundantes de Armazenamento em Discos Magnéticos – ENDERSON FERREIRA, JORGE RADY DE ALMEIDA JUNIOR
- BT/PCS/0308 – Utilizando Realidade Virtual e Objetos Distribuídos na Construção de uma Ferramenta de Aprendizagem Colaborativa – O Projeto Piaget – ISMAR FRANGO SILVEIRA, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0309 – Construção de Base de Conhecimento em Prolog a partir de Páginas HTML – WAGNER TOSCANO, EDSON SATOSHI GOMI
- BT/PCS/0310 – Verificação de Segurança em Confluência de Trajetórias de Aeronaves Utilizando Autômatos Híbridos – ÍTALO ROMANI DE OLIVEIRA, PAULO SÉRGIO CUGNASCA
- BT/PCS/0311 – Sistemas de Reconhecimento Biométrico Aplicados à Segurança de Sistemas de Informação – VILMAR DE SOUZA MACHADO, JORGE RADY DE ALMEIDA JUNIOR
- BT/PCS/0312 – Análise Comparativa de Arquiteturas Híbridas Intserv-Diffserv Utilizadas para Obtenção de QoS Fim-a-Fim em Redes IP – CARLOS A. A. BENITES, GRAÇA BRESSAN
- BT/PCS/0313 – Proposta para Otimização de Desempenho do Protocolo TCP em Redes Wireless 802.11 – ANDRÉ AGUIAR SANTANA, TEREZA CRISTINA DE MELO BRITO CARVALHO
- BT/PCS/0314 – Using the Moise + Model for a Cooperative Framework of MAS Reorganization – JOMI FRED HUBENER, JAIME SIMÃO SICHMAN
- BT/PCS/0315 – Ferramenta para Acompanhamento da Participação do Aluno em Sessões de Fórum Aplicada no Ensino a Distância via Web – GUSTAVO BIANCHI CINELLI, GRAÇA BRESSAN
- BT/PCS/0316 – Uma Infra-Estrutura para Agentes Arrematantes em Múltiplos Leilões Simultâneos – PAULO ANDRÉ LIMA DE CASTRO, JAIME SIMÃO SICHMAN
- BT/PCS/0317 – Reutilização de Software Através de Geração de Código e de Desenvolvimento de Componentes – Estudo de Caso – FÁBIO FÚRIA SILVA, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0318 – Detecção Automática das Transições de Corte e Fades – IZAURA CRISTINA ARAÚJO, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0319 – Educação a Distância e a Web Semântica: Modelagem Ontológica de Materiais e Objetos de Aprendizagem para Plataforma CoL – MOYSÉS DE ARAÚJO, MARIA ALICE GRIGAS VARELLA FERREIRA
- BT/PCS/0320 – Análise da Aplicação dos Padrões TMN no Gerenciamento de Sistemas de CRM – SANDRO ANTÔNIO VICENTE, MOACYR MARTUCCI JR

