Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Investigação e desenvolvimento de modelos estatísticos para análise discursiva automática

Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-05-02

Janeiro, 2005

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório, são apresentados modelos estatísticos inéditos para a análise discursiva automática. Esses modelos são baseados no modelo *Noisy-Channel* (Shannon, 1948) e treinados com o método de Aprendizado de Máquina *Expectation-Maximization* (Dempster et al., 1977), produzindo resultados promissores. Apresenta-se, além disso, um modelo para aprendizado não supervisionado das estruturas argumentais dos verbos, sendo este um passo necessário para um dos modelos de análise discursiva anteriores. A avaliação deste novo modelo também é relatada, demonstrando resultados satisfatórios. O trabalho apresentado neste relatório faz parte do projeto de doutorado que visa à investigação e desenvolvimento de técnicas de análise discursiva automática e à produção de um analisador para o português do Brasil.

Este trabalho conta com o suporte das agências FAPESP, CAPES, CNPq e Comissão Fulbright.

ÍNDICE

1. Introdução

Apresentam-se, neste relatório, modelos estatísticos inéditos desenvolvidos para a realização de análise discursiva automática. Modelos são construtos que tentam explicar como eventos/processos do mundo real ocorrem. Por modelo estatístico, ou probabilístico, entende-se um modelo que faz uso de probabilidades para determinar como e que eventos/processos ocorrem (Manning and Schütze, 1999). Análise discursiva automática, por sua vez, diz respeito ao processo de reconhecimento automático do conhecimento funcional e estrutural subjacente ao texto, implícito em sua superfície. Modelo estatístico para análise discursiva significa, portanto, um modelo que simula o processo de análise discursiva por meio de probabilidades.

Os modelos desenvolvidos são baseados no modelo *Noisy-Channel* de Shannon (1948) e treinados por meio do algoritmo de Aprendizado de Máquina *Expectation-Maximization* – EM (Dempster et al., 1977). Esses modelos visam a aprender regras semânticas, modeladas por parâmetros probabilísticos, que permitam a realização da análise discursiva.

Essa pesquisa faz parte do projeto de doutorado que visa a investigar e desenvolver técnicas de análise discursiva automática e produzir um analisador para o português do Brasil. Em particular, a pesquisa relatada foi desenvolvida durante realização de doutorado sanduíche na *University of Southern California*/*Information Sciences Institute*, em Los Angeles-CA, sob a supervisão do Prof. Dr. Daniel Marcu. Para mais detalhes sobre este trabalho de doutorado, vide Pardo et al. (2004) e Pardo e Nunes (2004).

Na próxima seção, introduz-se a área de análise discursiva automática. Na Seção 3, os modelos desenvolvidos são descritos e avaliados. Conclusões e comentários finais são apresentados na Seção 4.

2. Análise discursiva

As pesquisas em Lingüística e Lingüística Computacional têm comprovado há tempos que um texto não é simplesmente uma seqüência de sentenças justapostas, mas que possui uma estrutura altamente elaborada. Essa estrutura não é só composta por componentes lingüísticos, mas por informações de natureza extralingüística que influenciam a própria escrita do texto e a forma como ele é interpretado pelos leitores, fatores estes que se encontram nos níveis mais abstratos da língua. Apesar de tal certeza, poucos sistemas computacionais são capazes de extrair tal conhecimento de um texto ou mesmo de manipulá-lo.

Sistemas de Processamento de Línguas Naturais (PLN), como tradutores automáticos, corretores ortográficos e gramaticais e sumarizadores de texto, fazem uso, em sua maioria, de recursos como analisadores fonético-fonológicos, morfológicos e sintáticos. Considerando toda a riqueza da língua, o conhecimento por eles fornecido é pouco comparado ao que pode ser feito. Entretanto, tratar outros níveis que não estes traz aos sistemas um grau de complexidade normalmente inadmissível, devido à dificuldade de manipulação deste conhecimento e à ambigüidade conceitual inerente a estes níveis. A relação entre abstração lingüística e tipos de conhecimento é mostrada na Figura 1.

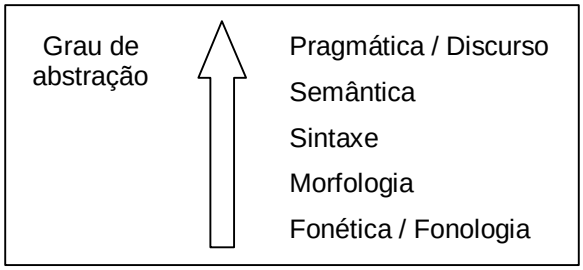


Figura 1 – Grau de abstração e níveis de conhecimento

Uma ferramenta capaz de extrair de um texto as informações dos níveis mais abstratos da língua é de grande utilidade para qualquer sistema de PLN. Os conhecimentos semântico e pragmático-discursivo possibilitam a tais sistemas terem maior controle sobre a tarefa a ser realizada. Por exemplo, com o conhecimento pragmático-discursivo, um tradutor automático é capaz de tratar questões de estruturação e organização textual dependentes de língua e gerar melhores resultados, enquanto um sumarizador automático pode saber com mais exatidão o que é supérfluo em um texto para ser excluído; com o conhecimento semântico, um corretor gramatical pode desambiguar as sentenças de um texto e gerar correções mais adequadas.

Analisadores discursivos, especificamente, são sistemas capazes de extrair o conhecimento discursivo (isto é, do nível da pragmática e do discurso) subjacente a um texto automaticamente. A próxima subseção introduz alguns trabalhos de análise discursiva automática.

2.1. Trabalhos sobre análise discursiva automática

Para a língua portuguesa, sabe-se de somente alguns trabalhos que abordam o tratamento discursivo como é comentado aqui. Rino (1996a), por exemplo, define uma modelagem profunda do discurso no contexto da sumarização automática que é aplicável para o português (Rino, 1996b). Ela sugere que a informação a ser selecionada para se formar um sumário deve ser deduzida com base semântica e intencional. Feltrim et. al. (2004) e Feltrim e Teufel (2004) investigam a estruturação esquemática de textos científicos em português com o intuito de construir uma ferramenta de auxílio à escrita. Soma-se a estes o DiZer[1] (Pardo et al., 2004), que é um analisador discursivo automático resultante deste trabalho de doutorado. Este sistema realiza a análise com base em informação morfossintática e sintático-semântica.

Para o inglês, muitos trabalhos se destacam na área de processamento de discurso. Em especial e de grande importância são os trabalhos de Marcu (1997a, 1997b, 1998a, 1998b, 1999a, 1999b, 2000a, 2000b), que desenvolveu o primeiro "*parser* discursivo" (conforme denominado por ele) para o inglês, de Corston-Oliver (1998a, 1998b, 1998c) e de Schielder (2002), sendo que cada um deles abordou de forma distinta o problema de análise discursiva: com base na ocorrência de marcadores discursivos, basicamente, o analisador discursivo de Marcu realiza a análise retórica de textos[2]; Corston-Oliver utiliza a sintaxe e aspectos da forma

[1] http://www.nilc.icmc.usp.br/~thiago/DiZer.html
[2] A retórica, segundo Hovy (1988), é a parte palpável da pragmática, através da qual se estabelece a coerência de um texto.

lógica das sentenças para realizar a análise retórica; Schielder, por sua vez, além dos marcadores discursivos, analisa características como topicalidade e posição das sentenças de um texto para produzir sua análise retórica.

Há diversas teorias discursivas que representam e abordam aspectos distintos do discurso, por exemplo, RST (*Rhetorical Structure Theory*) (Mann and Thompson, 1987), GSDT (*Grosz and Sidner Discourse Theory*) (Grosz and Sidner, 1986), DRT (*Discourse Representation Theory*) (Kamp, 1981), SDRT (*Segmented Discourse Representation Theory*) (Asher and Lascarides, 2003), *clausal relations* (Jordan, 1992), entre outras. Algumas outras teorias de discurso unificam características de várias outras, como é o caso do modelo de discurso de Rino (1996a) e de Moser e Moore (1996). Assim como na maioria dos trabalhos de análise discursiva, a RST é a teoria de discurso utilizada neste trabalho de doutorado. Ela é introduzida na subseção seguinte.

2.2. RST – *Rhetorical Structure Theory*

Segundo Reiter e Dale (2000), a RST é a teoria de discurso mais difundida atualmente. Ela tem sido usada para os mais diversos fins além de análise discursiva, como tradução automática (p.ex., Marcu et al., 2000), geração de língua natural (p.ex., Maybury, 1992; Moore e Paris, 1993), correção automática de textos (p.ex., Burstein et al., 2003), sumarização automática (p.ex., O'Donnel, 1997; Marcu, 2000a; Pardo, 2002), resolução anafórica (p.ex., Cristea et al., 1998; Schauer, 2000), etc.

A RST define relações retóricas (cerca de 25) que se estabelecem entre segmentos discursivos, também chamados de proposições, que são unidades mínimas de significado que, geralmente, são expressas por orações. O reconhecimento dessas relações em um texto constitui um passo importante para o entendimento do texto, uma vez que elas acabam por refletir o raciocínio que derivou o conteúdo veiculado pelo texto, e, portanto, sua própria coerência. Diz-se que um texto é coerente se relações retóricas podem ser estabelecidas entre suas proposições. A Figura 2 mostra um exemplo de texto estruturado retoricamente.

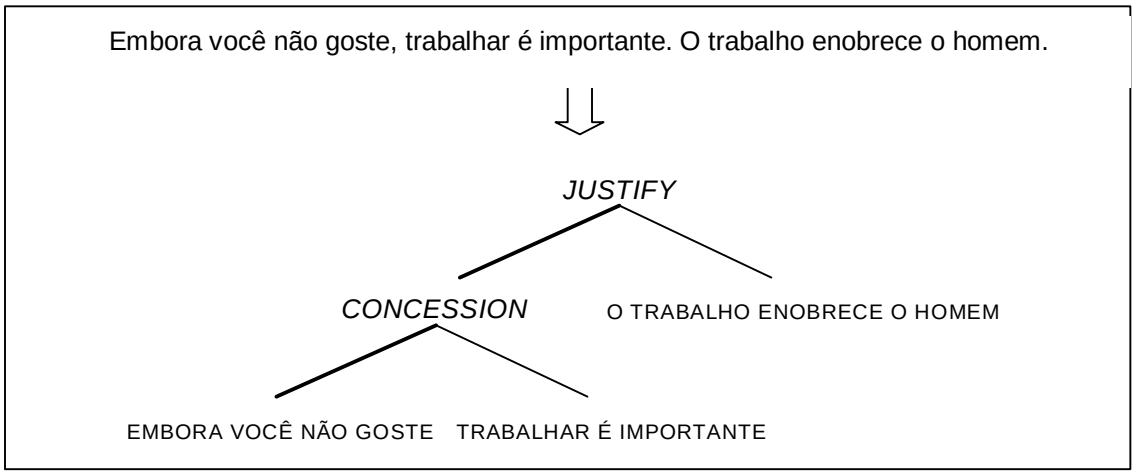


Figura 2 – Exemplo de uso da RST

Nesse texto, tem-se que:

- a relação *concession* entre as duas proposições expressas pelas orações da primeira sentença indica uma aparente incompatibilidade entre elas;

- a relação *justify* entre as proposições expressas pela primeira e pela segunda sentença justifica o porquê de se produzir a primeira.

As linhas mais escuras indicam a informação mais importante entre as proposições relacionadas, que, de acordo com a própria RST, é chamada de núcleo da relação. O trecho de texto não nuclear é chamado de satélite. É possível, portanto, entender o texto, visto que se pode reconhecer e estabelecer as devidas relações retóricas entre suas partes.

É importante notar que análises nesse nível são inerentemente ambíguas, ou seja, em muitos casos, não é possível determinar uma única relação retórica que se estabeleça entre duas proposições, pois a determinação da relação retórica é dependente da interpretação feita do texto (por um humano), a qual, por ser puramente subjetiva, pode variar. Um sistema de análise discursiva deve ser capaz, portanto, de reconhecer e tratar essa ambigüidade.

Para mais informação sobre algumas das principais teorias discursivas, vide Pardo e Nunes (2003).

3. Uma abordagem estatístico-semântica para a análise discursiva

Nesta seção, apresentam-se os modelos estatísticos desenvolvidos para a realização da análise discursiva automática, assim como a avaliação e validação destes.

Inicialmente, nas próximas duas subseções, o modelo *Noisy-Channel* e o método EM, nos quais os modelos estatísticos se baseiam, são introduzidos. A seguir, os modelos de análise discursiva são descritos.

3.1. Modelo *Noisy-Channel*

O modelo *Noisy-Channel* foi proposto por Shannon (1948) para a modelagem da capacidade de transmissão de dados em um canal. Originalmente, esse modelo foi aplicado para os sistemas de telefonia para tentar prever e corrigir os erros ocorridos durante a transmissão de mensagens. A Figura 3 mostra as componentes do modelo. Inicialmente, tem-se uma mensagem M1 produzida por uma fonte (*source*) com probabilidade P(M1); essa mensagem, ao ser transmitida por um canal com ruído (*noisy-channel*), é corrompida e erros são introduzidos, transformando-a em M2 com probabilidade P(M2|M1).

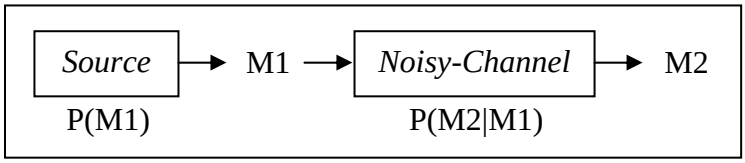


Figura 3 – Modelo *Noisy-Channel*

No caso da telefonia, a fonte normalmente é uma pessoa e o canal com ruído é a linha telefônica. Segundo esse modelo, sabendo-se as probabilidades P(M1) e P(M2|M1), é possível determinar M1 a partir de M2, por um processo conhecido como decodificação. Esse processo consiste em escolher a mensagem M1 que maximize as probabilidades P(M1) e P(M2|M1).

O modelo *Noisy-Channel* foi aplicado com sucesso na área de processamento de fala e, mais recentemente, começou a ser aplicado em tarefas de processamento de língua natural escrita. Em especial, esse modelo causou grandes avanços na área

de tradução automática estatística, produzindo os tradutores automáticos com melhores resultados na área, segundo as avaliações internacionais realizadas pelo NIST[3] (*National Institute of Standards and Technology*). Nesta área, a tradução de uma sentença em francês para o inglês, por exemplo, é modelada da seguinte forma: uma sentença em inglês E é produzida por uma pessoa (*source*) com probabilidade P(E) e, ao ser comunicada (por meio de um canal com ruído hipotético), transforma-se em uma sentença em francês F com probabilidade P(F|E). Fazer a tradução de F para E consiste, portanto, em um processo de decodificação, considerando-se que P(E) e P(F|E) são conhecidos ou podem ser aprendidos. Em geral, as probabilidades P(E) e P(F|E) são estimadas pela aplicação de uma técnica de Aprendizado de Máquina chamada *Expectation-Maximization* (EM) (Dempster et al., 1977), que será introduzida na próxima subseção.

Em PLN, o conjunto de probabilidades P(E) é chamado modelo lingüístico, pois informa a probabilidade de E ocorrer em uma língua; o conjunto de probabilidades P(F|E), por sua vez, é chamado modelo de tradução, pois indica como E se traduz/transforma em F. O processo de transformação de E em F é a parte principal do modelo *Noisy-Channel* e é chamado modelo ou história gerativa. No geral, pode-se afirmar que o sucesso do modelo *Noisy-Channel* na tarefa em que é aplicado depende da história gerativa que se assume para o problema tratado. Como ilustração de uma história gerativa, considere a história comumente adotada por modelos simples de tradução automática de inglês para português para a sentença em inglês *Mary did not slap the green witch*:

(1) Inicialmente, algumas palavras são replicadas um número determinado de vezes ou eliminadas (a palavra *slap* foi replicada duas vezes e a palavra *did* foi eliminada)

Mary not slap slap slap the green witch

(2) A seguir, cada palavra em inglês é substituída por uma palavra em português (na seqüência, *Mary* por “Maria”, *not* por “não”, primeira ocorrência de *slap* por “deu”, segunda ocorrência de *slap* por “um”, terceira ocorrência de *slap* por “tapa”, *the* por “na”, *green* por “verde”, *witch* por “bruxa”)

Maria não deu um tapa na verde bruxa

(3) Por fim, as palavras são reordenadas, produzindo a sentença em português

Maria não deu um tapa na bruxa verde

As decisões de replicar e eliminar palavras, substituir palavras em inglês por suas correspondentes em português e reordenar as palavras da sentença são feitas com base nas probabilidades que compõem a modelo de tradução, isto é, P(F|E). Essas probabilidades são chamadas parâmetros do modelo e, em geral, são aprendidas a partir de um corpus paralelo pelo uso do método EM, introduzido na subseção seguinte.

Para mais detalhes sobre o modelo *Noisy-Channel*, sugere-se a leitura da obra de referência ou Manning e Schütze (1999).

[3] http://www.nist.gov/

3.2. Método EM

O método EM (*Expectation-Maximization*) (Dempster et al., 1977) é utilizado para estimar parâmetros de modelos probabilísticos em que há variáveis não observadas. Se todos os dados/variáveis de um problema fossem observados, seria simples estimar os parâmetros de seu modelo probabilístico subjacente. Quando algum dado/variável não está disponível, o método EM costuma ser aplicado.

Em tradução automática estatística, como ilustrado anteriormente, assume-se que uma sentença é tradução de outra se e somente se é possível alinhar suas palavras, isto é, determinar as palavras da sentença em português que correspondem às palavras da sentença em inglês. Nesse caso, o alinhamento entre as sentenças é a variável não observada. Note que há diversos alinhamentos possíveis entre duas sentenças, pois se devem considerar todas as possibilidades de correspondência entre as palavras (por exemplo, no exemplo anterior, *Mary* com "Maria", *Mary* com "não", *Mary* com "deu", ... , *witch* com "Maria", *witch* com "não", etc.). Se esses parâmetros de correspondência entre as palavras fossem conhecidos, seria possível determinar o melhor alinhamento entre duas sentenças; por outro lado, se o alinhamento entre duas sentenças fosse conhecido, seria possível estimar os parâmetros do modelo. Em casos como esse, em que não se tem nenhuma destas informações devido à presença da variável não observada, utiliza-se o método EM.

O método EM funciona da seguinte forma:

1. são atribuídas probabilidades iniciais para todos os parâmetros do modelo (assume-se, normalmente, a distribuição uniforme, em que todos os parâmetros têm igual probabilidade);
2. determinam-se as probabilidades de todos os alinhamentos possíveis para cada par de sentenças inglês-português do corpus de treinamento, em que a probabilidade de um alinhamento qualquer é dada pela multiplicação de todos os parâmetros que compõem o alinhamento (por exemplo, para o par de sentenças da subseção anterior, considerando-se somente a correspondência entre as palavras, a probabilidade do alinhamento considerado é dada por t(Maria|*Mary*) x t(não|*not*) x t(deu|*slap*) x t(um|*slap*) x t(tapa|*slap*) x t(na|*the*) x t(verde|*green*) x t(bruxa|*witch*), sendo que t é o parâmetro que indica a probabilidade de tradução de uma palavra em outra);
3. com base nas probabilidades dos alinhamentos, estimam-se novas probabilidades para os parâmetros;
4. com base nos novos parâmetros, estimam-se novas probabilidades para os alinhamentos;
5. e assim por diante.

O cálculo das probabilidades dos alinhamentos é o passo *expectation* do método EM; a estimativa dos parâmetros com base nos alinhamentos é o passo *maximization* do método. É garantido que, a cada iteração do método EM, as probabilidades dos parâmetros se aproximam do valor ideal. Encerra-se o método quando as probabilidades convergem e estabilizam.

A base do aprendizado do método EM consiste na repetição de padrões no corpus de treinamento. A cada padrão que se repete, o método tem mais evidências para incrementar a probabilidade de determinados parâmetros do modelo, penalizando os parâmetros que representam eventos improváveis ou inadequados. Por exemplo, em todas as sentenças do corpus de treinamento, é provável que existam muitos alinhamentos entre as palavras *witch* e "bruxa"; para cada padrão

deste encontrado, a probabilidade do parâmetro t(bruxa|*witch*) é incrementada, enquanto os outros parâmetros possíveis t(Maria|*witch*), t(deu|*witch*), ... , t(verde|*witch*) não são, sendo, portanto, penalizados.

Um dos problemas do método EM é que ele é de complexidade exponencial e, diante da quantidade de dados necessários para seu treinamento (principalmente em PLN), sua aplicação se torna, na maior parte dos casos, inviável. Há diversas formas de se lidar com esse problema. A maioria das pesquisas na área reduz o número de valores possíveis para as variáveis não observadas e a quantidade de dados para treinamento. Por exemplo, no caso da tradução automática, costuma-se utilizar somente os alinhamentos mais prováveis entre duas sentenças (em vez de todos) e limitar o tamanho das sentenças para um número determinado de palavras.

Para mais detalhes sobre o método EM, sugere-se a leitura da obra de referência ou Manning e Schütze (1999). Na próxima subseção, os modelos estatísticos desenvolvidos para análise discursiva são descritos.

3.3. Modelos de análise discursiva

Foram desenvolvidos três modelos estatísticos de análise discursiva segundo o modelo *Noisy-Channel*. Todos os modelos foram desenvolvidos, treinados e testados com base em um corpus de relações de discurso de causa-efeito para a língua inglesa. Apesar disso, os modelos são genéricos e independentes de língua, e, portanto, podem ser aplicados a qualquer relação discursiva ou língua. A escolha da relação causa-efeito para treino e teste dos modelos foi devido à importância desta relação e ao fato de ela ser comum a todas as teorias discursivas.

O modelo *Noisy-Channel* no qual os modelos de análise discursiva se baseiam é mostrado na Figura 4. Inicialmente, um evento de causa C é observado (produzido) com probabilidade P(C); esse evento é então corrompido e transforma-se no efeito E com probabilidade P(E|C). Nos três modelos de análise discursiva propostos, considera-se que a distribuição de P(C) é uniforme, isto é, todas os possíveis eventos C possuem a mesma probabilidade de serem observados. Com isso, P(E|C) é a principal componente do modelo, responsável por explicar como a causa C se transforma no efeito E. É nesse ponto em que os três modelos propostos divergem, ou seja, em como se supõe que um evento se transforma no outro.

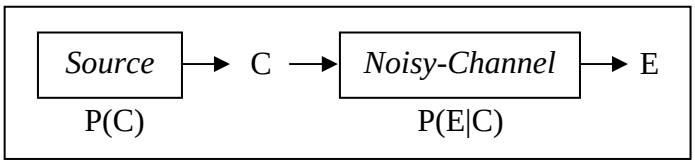


Figura 4 – Modelo *Noisy-Channel* para análise discursiva

Uma modificação que foi feita no modelo *Noisy-Channel* ao adaptá-lo para o problema da análise discursiva foi considerar uma probabilidade conjunta P(C,E) em vez da probabilidade condicional P(E|C). Isso foi feito porque, na probabilidade conjunta, não se assume como conhecida a dependência que existe entre os eventos, isto é, não se afirma que E é dependente de C (como ocorre na probabilidade condicional). Em relações causa-efeito, é evidente que o efeito deve-se à ocorrência da causa e, portanto, que o efeito é condicionado à causa. Entretanto, não se pode afirmar isso para outras relações do discurso, nas quais a direção da dependência não é clara ou não existe, como nas relações *contrast* e *sequence*, por exemplo. Usar a

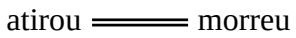
probabilidade conjunta torna o modelo flexível o bastante para ser aplicado para qualquer uma das relações consideradas (no caso, as relações retóricas da RST).

Nas subseções seguintes, os modelos desenvolvidos são detalhados.

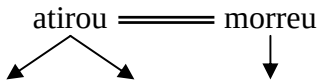
3.3.1. Um modelo baseado em palavras

O primeiro modelo desenvolvido é baseado no relacionamento entre as palavras dos eventos. Sua história gerativa é delineada abaixo, exemplificada para as sentenças "Ele atirou em Maria." e "Ela morreu.", entre as quais há uma relação causa-efeito. Em cada passo da história gerativa, a produção da estrutura de causa-efeito subjacente às sentenças é exibida.

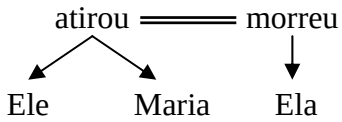
1) produz-se a relação causa-efeito entre "atirou" e "morreu" com probabilidade ce(atirou,morreu)



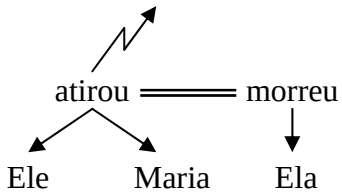
2) determina-se que "atirou" possui 2 argumentos (quem atirou e quem foi baleado) com probabilidade narg(2|atirou) e que "morreu" possui 1 argumento (quem morreu) com probabilidade narg(1|morreu)



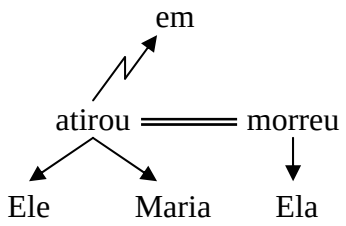
3) escolhem-se os argumentos dos eventos: "atirou" requer os argumentos "Ele" e "Maria" com probabilidades arg(Ele|atirou) e arg(Maria|atirou), respectivamente; "morreu" requer o argumento "Ela" com probabilidade arg(Ela|morreu)



4) determina-se que "atirou" produz uma palavra extra (isto é, que não é argumento) com probabilidade phi(1|atirou) e que "morreu" não produz palavras extras com probabilidade phi(0|morreu)



5) a palavra extra “em” é produzida com probabilidade ew(em)



Portanto, segundo essa história gerativa, a probabilidade P(C,E) é calculada pela seguinte fórmula:

$$
\begin{aligned}
P(C,E) = \quad & ce(causa,efeito) \times \\
& narg(M|causa) \times narg(N|efeito) \times \\
& \prod_{i=1}^{M} arg(w_i|causa) \times \prod_{i=1}^{N} arg(w_i|efeito) \times \\
& phi(|C|-M|causa) \times phi(|E|-N|efeito) \times \\
& \prod_{i=1}^{|C|-M} ew(w_i) \times \prod_{i=1}^{|E|-N} ew(w_i)
\end{aligned}
$$

em que M e N indicam o número de palavras da causa e do efeito, respectivamente; |C| e |E| são o tamanho (número de palavras) da causa C e do efeito E, respectivamente; a letra w representa uma palavra.

As probabilidades ce, narg, arg, phi e ew são os parâmetros do modelo e são estimadas pela aplicação do método EM a partir de um corpus de sentenças relacionadas por relações causa-efeito (vide subseção 3.4 para descrição do corpus). Para aplicação do método EM, nos modelos de análise discursiva propostos, as estruturas de causa-efeito são as variáveis não observadas. Idealmente, espera-se que o método EM aprenda os parâmetros que representem mais adequadamente as estruturas de causa-efeito subjacentes às sentenças. Note que, para cada sentença, há diversas estruturas de causa-efeito possíveis (considerando diferentes eventos de causa e efeito, diferentes números de argumentos e diferentes argumentos, diferentes números de palavras extras e diferentes palavras extras). Para as sentenças “Ele atirou em Maria” e “Ela morreu”, espera-se que o método aprenda, por exemplo, que: a probabilidade ce(atirou,morreu) seja maior do que ce(Ele,morreu), ce(Maria,morreu), ce(Ele,Ela) e ce(Maria,Ela); a probabilidade narg(2|atirou) seja maior do que narg(3|atirou), narg(1|atirou) e narg(0|atirou); a probabilidade arg(Ele|atirou) seja maior do que arg(em|atirou); etc. Os parâmetros aprendidos codificam o conhecimento semântico que se deseja aprender para a realização da análise discursiva automática.

Sabendo-se o valor dos parâmetros e, portanto, como calcular P(C,E), é possível identificar sentenças relacionadas por relações causa-efeito: se P(C,E) das sentenças é alta (isto é, há pelo menos uma estrutura de causa-efeito provável subjacente às sentenças), então há uma relação causa-efeito entre as sentenças; caso contrário, não há uma relação causa-efeito.

Para tornar a história gerativa proposta mais informada e o método EM mais eficiente, algumas restrições foram impostas, a saber:

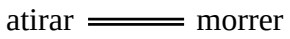
1) somente substantivos e verbos podem ser eventos de causa e efeito: para determinar as classes gramaticais das palavras, utiliza-se um *tagger*, isto é, um etiquetador morfossintático, segundo o modelo de Ratnaparkhi (1996), que possui uma precisão de 97%, a maior relatada na área até o momento;
2) somente substantivos (e pronomes), verbos, adjetivos e advérbios, ou seja, as palavras de classe aberta, podem ser argumentos de eventos de causa e efeito, e, assim, as palavras de classe fechada serão sempre consideradas palavras extras do modelo.

3.3.2. Um modelo baseado em conceitos

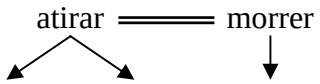
O segundo modelo proposto não se baseia apenas em palavras, mas nos conceitos correspondentes às palavras também. Com isso, espera-se que, em vez de se aprender que os argumentos de “atirou” é “Ele” e “Maria”, aprenda-se que “atirou” tem argumentos do tipo “pessoa”. Isso tornaria o modelo mais genérico e o conhecimento extraído mais intuitivo.

A história gerativa é ilustrada abaixo. Da mesma forma, é acompanhada de um exemplo e da estrutura de causa-efeito sendo construída. Note que, diferentemente da história gerativa anterior, manipula-se, agora, conceitos que, ao final, serão mapeados nas palavras das sentenças (conforme o novo parâmetro t).

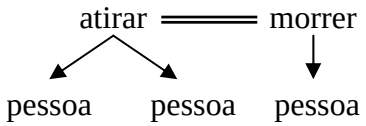
1) produz-se a relação causa-efeito entre os conceitos “atirar” e “morrer” com probabilidade ce(atirar,morrer)



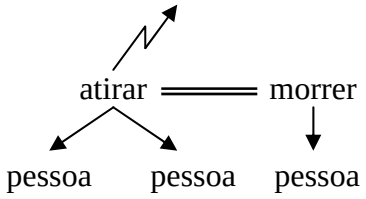
2) determina-se que “atirar” possui 2 argumentos com probabilidade narg(2|atirar) e que “morrer” possui 1 argumento com probabilidade narg(1|morrer)



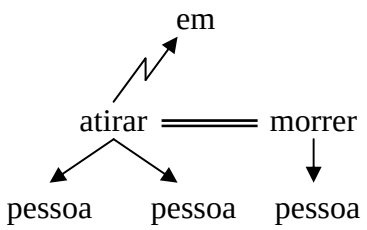
3) escolhem-se os argumentos dos eventos: “atirar” requer argumentos “pessoa” com probabilidades arg(pessoa|atirar); “morrer” também requer o argumento “pessoa” com probabilidade arg(pessoa|morrer)



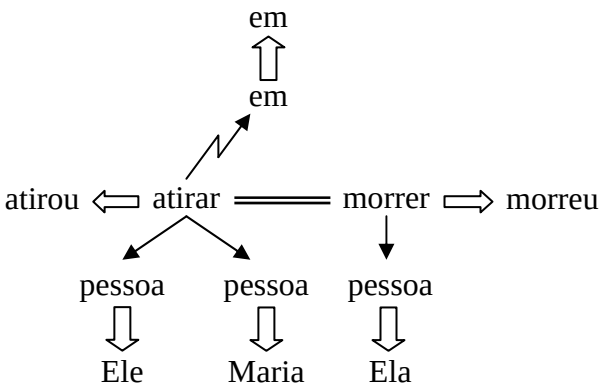
4) determina-se que “atirar” produz uma palavra extra (isto é, que não é argumento) com probabilidade phi(1|atirar) e que “morrer” não produz palavras extras com probabilidade phi(0|morrer)



5) o conceito extra “em” é produzido com probabilidade ew(em)



6) os conceitos são mapeados nas respectivas palavras com probabilidades t(atirou|atirar), t(morreu|morrer), t(Ele|pessoa), t(Maria|pessoa), t(Ela|pessoa) e t(em|em).



Segundo essa história gerativa, a probabilidade P(C,E) é calculada pela seguinte fórmula:

$$
\begin{aligned}
P(C,E) = \quad & ce(causa,efeito) \times t(w_{causa}|causa) \times t(w_{efeito}|efeito) \\
& narg(M|causa) \times narg(N|efeito) \times \\
& \prod_{i=1}^{M} (arg(c_i|causa) \times t(w_i|c_i)) \times \prod_{i=1}^{N} (arg(c_i|efeito) \times t(w_i|c_i)) \times \\
& phi(|C|-M|causa) \times phi(|E|-N|efeito) \times \\
& \prod_{i=1}^{|C|-M} (ew(c_i) \times t(w_i|c_i)) \times \prod_{i=1}^{|E|-N} (ew(c_i) \times t(w_i|c_i))
\end{aligned}
$$

Durante o treinamento do modelo, para se obter os conceitos correspondentes às palavras, utilizou-se a WordNet[4] (Fellbaum, 1999). Para cada palavra, os três hiperônimos mais freqüentes foram buscados automaticamente.

De forma similar ao modelo anterior, assume-se que somente substantivos e verbos podem ser eventos de causa e efeito e que somente palavras de classe aberta podem ser argumentos destes eventos. Os parâmetros do modelo também são aprendidos por meio do método EM.

[4] http://www.cogsci.princeton.edu/~wn/

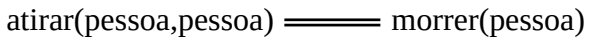
3.3.3. Um modelo baseado na estrutura argumental dos verbos

Os dois modelos anteriores fazem a suposição simplista de que a relação causa-efeito se estabelece entre palavras/conceitos das sentenças. Por exemplo, segundo esses modelos, o evento de atirar causa o evento de morrer, ou seja, ce(atirar,morrer).

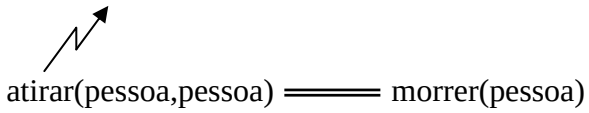
O novo modelo proposto considera que a relação causa-efeito se estabelece entre o fato de uma pessoa atirar em outra pessoa e o fato dessa pessoa morrer, isto é, ce(atirar(pessoa$_1$,pessoa$_2$),morrer(pessoa$_2$)), em que atirar(pessoa,pessoa) e morrer(pessoa) são as estruturas argumentais dos verbos atirar e morrer, respectivamente. A estrutura argumental de um verbo indica quantos são e quais são os possíveis argumentos que o verbo exige.

A nova história gerativa é delineada a seguir.

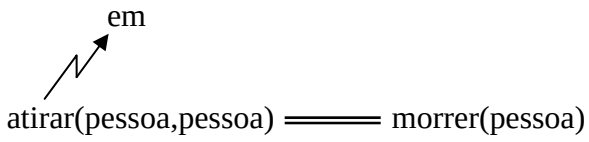
1) produz-se a relação causa-efeito entre o fato de uma pessoa atirar em outra pessoa e o fato dessa pessoa morrer com probabilidade ce(atirar(pessoa,pessoa),morrer(pessoa))



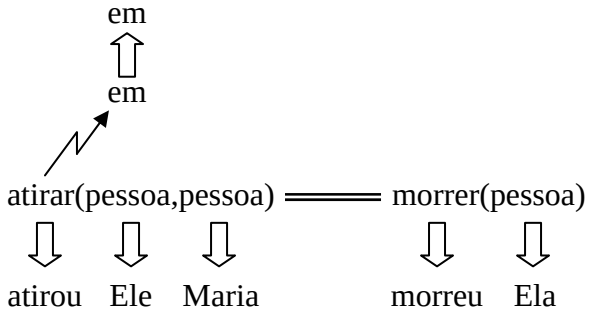
2) determina-se que “atirar” produz uma palavra extra (isto é, que não é argumento) com probabilidade phi(1|atirar) e que “morrer” não produz palavras extras com probabilidade phi(0|morrer)



3) o conceito extra “em” é produzido com probabilidade ew(em)



4) os conceitos são mapeados nas respectivas palavras com probabilidades t(atirou|atirar), t(morreu|morrer), t(Ele|pessoa), t(Maria|pessoa), t(Ela|pessoa) e t(em|em).



Para essa história gerativa, a probabilidade P(C,E) é calculada pela seguinte fórmula:

$$
\begin{aligned}
P(C,E) = \quad & ce(v_{causa}(Args_{causa}),v_{efeito}(Args_{efeito})) \times \\
& t(w_{causa}|v_{causa}) \times t(w_{efeito}|v_{efeito}) \times \\
& \overset{\text{number of args in cause}}{\prod_{i=1}} t(w_i|arg_i) \quad \times \quad \overset{\text{number of args in effect}}{\prod_{i=1}} t(w_i|arg_i) \times \\
& phi(M|causa) \times phi(N|efeito) \times \\
& \prod_{i=1}^{M} (ew(c_i) \times t(w_i|c_i)) \times \prod_{i=1}^{N} (ew(c_i) \times t(w_i|c_i))
\end{aligned}
$$

Para se determinar as possíveis estruturas argumentais em uma sentença, um *parser* estatístico foi usado. O *parser* segue o modelo de Collins (1999) e tem precisão de aproximadamente 91%, sendo o *parser* de maior precisão para a língua inglesa. Com o uso do *parser*, o modelo é capaz de lidar com as diferenças entre argumentos e adjuntos dos verbos (argumentos são obrigatórios, enquanto adjuntos não) e de identificar as orações encaixadas, que, em princípio, não devem ser consideradas na produção da estrutura argumental de uma sentença.

Considerou-se que somente substantivos e verbos podem ser predicadores em uma sentença e que todas as palavras de classe aberta podem ser argumentos em uma estrutura argumental. Entretanto, restringiu-se o número de argumentos de uma estrutura argumental para, no máximo, 3. Essa restrição foi sintaticamente motivada, pois, na maior parte das teorias sintáticas, um verbo não possui mais do que 3 argumentos.

Neste modelo, o método EM deve aprender tanto as melhores estruturas argumentais para as sentenças relacionadas (por exemplo, que a melhor estrutura argumental para a sentença "Ele atirou em Maria" é atirou(Ele,Maria)) quanto os melhores pares de causa-efeito (por exemplo, ce(atirou(Ele,Maria),morreu(Ela))).

3.4. Corpus

Para compor um corpus de relações causa-efeito, foram selecionadas 400.000 sentenças em inglês de diversos corpora jornalísticos (*Reuters*, *New York Times*, *Wall Street Journal* e coleção TREC[5], entre outros) que contivessem a palavra *because* em seu interior, totalizando 11 milhões de palavras. Optou-se por selecionar as sentenças com a palavra *because* porque essa palavra é um dos marcadores discursivos mais fortes que sinalizam a relação causa-efeito na língua inglesa.

As sentenças coletadas foram anotadas com as ferramentas necessárias (*tagger*, *parser* e WordNet) para a execução de cada um dos modelos propostos e, a seguir, segmentadas em partes que correspondessem à causa (o trecho da sentença que está depois da palavra *because*) e ao efeito (o trecho de texto que está antes da palavra *because*) dentro da sentença. Por exemplo, na sentença *They prefer women commanders because they are less strict*, o seguinte par de causa-efeito é produzido:

they are less strict – they prefer women commanders

[5] http://trec.nist.gov/

Foram selecionadas sentenças com, no máximo, 35 palavras. Essa limitação se faz necessária devido ao algoritmo EM, que é de complexidade exponencial, como já discutido.

De forma similar, foram coletadas 900.000 sentenças com a palavra *but* em seu interior, para formar um corpus de relações contraste. Parte desse corpus foi usada para testar os modelos treinados com os pares de causa-efeito. Esses modelos devem (a) reconhecer os pares de causa-efeito como sendo realmente de causa-efeito e (b) rejeitar os pares de contraste. A seguir, a avaliação dos modelos é apresentada.

3.5. Avaliação e validação dos modelos de análise discursiva

Os três modelos propostos foram treinados com o corpus de relações causa-efeito e testados com esse mesmo corpus e com o corpus de relações contraste, produzindo os resultados mostrados na Tabela 1. O modelo baseado em palavras obteve uma taxa de acerto de 59% ao ser aplicado ao corpus de teste de relações causa-efeito (isto é, em 59% dos casos, o modelo detectou a relação causa-efeito entre as sentenças) e uma taxa de acerto de 61% ao ser aplicado ao corpus de teste de relações contraste (isto é, em 61% dos casos, o modelo detectou que aquelas relações não eram de causa-efeito); o modelo baseado em conceitos obteve taxas de acerto de 71% e 43% ao ser aplicado aos corpora de relações causa-efeito e contraste, respectivamente; e o modelo baseado nas estruturas argumentais obteve taxas de acerto de 61% e 50% ao ser aplicado aos corpora de relações causa-efeito e contraste, respectivamente.

Para a realização desta avaliação, determinou-se que duas sentenças estariam relacionadas por uma relação causa-efeito se a probabilidade P(C,E) fosse maior do que um *threshold*, o qual foi calculado com um corpus de relações causa-efeito (disjunto do corpus de treino) de 100 sentenças. De fato, em vez de se usar a probabilidade P(C,E), utilizou-se a medida de perplexidade, que é, basicamente, a probabilidade P(C,E) normalizada em relação ao tamanho das sentenças relacionadas. Diferentemente da probabilidade P(C,E), a perplexidade é independente do tamanho das sentenças, tornando a avaliação mais robusta.

Tabela 1 – Taxa de acerto dos modelos de análise de discursiva

	causa-efeito	**contraste**
Modelo baseado em palavras	59%	61%
Modelo baseado em conceitos	71%	43%
Modelo baseado em estruturas argumentais	61%	50%

Algumas observações que puderam ser feitas sobre cada um dos modelos são:

- há um vocabulário comum entre as palavras das relações causa-efeito e contraste, o que não permite que o modelo baseado em palavras tenha uma taxa de acerto maior;
- a utilização de conceitos melhorou a identificação das relações causa-efeito, mas piorou o desempenho do modelo na identificação de relações que não são de causa-efeito;
- no modelo baseado nas estruturas argumentais, o método de aprendizado de máquina não foi capaz de aprender boas estruturas argumentais e bons pares de causa-efeito ao mesmo tempo.

Os dois primeiros modelos não tiveram uma taxa de acerto maior devido, principalmente, às suposições simplistas que fazem, como já foi discutido anteriormente. Em relação ao terceiro modelo, por sua vez, o problema reside no fato dos dados serem esparsos para que o aprendizado seja eficaz. Dados esparsos apresentam pouca redundância, que é a condição necessária para que técnicas de Aprendizado de Máquina tenham um bom desempenho. Para resolver essa questão, a análise discursiva como tratada pelo modelo baseado em estruturas argumentais foi dividida em dois subproblemas: (I) determinar as possíveis estruturas argumentais e (II) determinar os pares de causa-efeito. Assim, em vez de um único modelo *noisy-channel* para a análise discursiva, são necessários, agora, dois modelos, um para cada subproblema. O novo modelo proposto para a derivação das estruturas argumentais é apresentado na subseção 3.6.

O modelo baseado em estruturas argumentais foi, ainda, validado. Apesar de seu desempenho mediano para análise discursiva, ele se mostrou muito interessante na questão de predição de efeitos dado um evento de causa ou vice-versa. Foi desenvolvido um decodificador que, a partir do conjunto de parâmetros de causa-efeito aprendido pelo método EM durante o treinamento do modelo baseado em estruturas argumentais (isto é, os parâmetros ce), procura pelos efeitos mais prováveis dada uma causa ou vice-versa. Por exemplo, dada a estrutura argumental de causa *was*(*he,insane*), o decodificador sugere como possível efeito *argue*(*lawyers,innocence*).

Em uma avaliação preliminar do modelo baseado em estruturas argumentais nessa aplicação, os seguintes resultados foram obtidos: em 40% dos casos, as predições faziam sentido, consideradas passíveis de ocorrência no mundo real; em 25% dos casos, não foi possível determinar com certeza se as predições eram plausíveis; em 35% dos casos, os efeitos previstos não faziam sentido. Nos casos em que não era possível determinar se os efeitos eram plausíveis ou não, dois problemas foram identificados nas estruturas argumentais sugeridas: presença de nomes próprios e presença de palavras estranhas ou desconhecidas, como ocorrem nas estruturas *killed*(*capano,fahey*) e *shut*(*camp,unher*).

Até onde se sabe, este é o primeiro trabalho que tenta mapear causas em possíveis efeitos e vice-versa automaticamente. Melhorando-se a qualidade dos pares de causa-efeito aprendidos pelo modelo baseado em estruturas argumentais, os resultados da predição devem melhorar também, o que consiste no próximo passo dessa pesquisa, baseado no que é descrito a seguir.

A subseção seguinte apresenta o modelo proposto para o aprendizado das estruturas argumentais, como discutido anteriormente (subproblema I).

3.6. Um modelo estatístico para o aprendizado das estruturas argumentais

Foi proposto um modelo estatístico baseado no modelo *noisy-channel* para resolver o subproblema I descrito anteriormente, isto é, o problema de se determinar as possíveis estruturas argumentais dos verbos. Considerou-se, nesse caso, somente a classe dos verbos como predicadora.

Conhecendo-se as estruturas argumentais mais prováveis das sentenças, basta que, no modelo de análise discursiva baseado em estruturas argumentais, o método EM aprenda apenas quais os melhores pares de causa-efeito. Com isso, o problema dos dados serem esparsos é significativamente reduzido.

Há diversas propostas na literatura para a aquisição das estruturas argumentais dos verbos. Esses trabalhos foram estudados e avaliados para se

determinar suas vantagens e desvantagens para a tarefa em questão, assim como para se entender a problemática envolvida na tarefa. Os trabalhos mais relevantes são brevemente descritos na subseção seguinte.

3.6.1. Trabalhos correlatos

Há alguns projetos para o desenvolvimento em larga escala de repositórios de informação semântica de verbos em inglês, visando a codificar suas estruturas argumentais possíveis, suas estruturas de subcategorização e exemplos reais de uso dos verbos. Os projetos mais conhecidos são a FrameNet (Baker et al., 1998), a VerbNet (Kipper et al., 2000) e o PropBank (Kingsbury and Palmer, 2002). Como exemplo, as Figuras 5, 6 e 7 mostram partes das anotações da FrameNet, da VerbNet e do PropBank para o verbo *buy*, respectivamente. A FrameNet mostra o padrão no qual o verbo ocorre e exemplos; a VerbNet mostra os papéis temáticos dos argumentos que o verbo requer e seus traços semânticos, as possíveis estruturas de subcategorização e exemplos para cada uma; o PropBank exibe os papéis dos argumentos, as possíveis estruturas de subcategorização e exemplos para cada uma. Note que o PropBank também distingue os complementos obrigatórios (os argumentos, propriamente ditos) dos opcionais (os adjuntos). O complemento opcional na Figura 7 é o argumento identificado por ArgM-MNR (isto é, argumento de modo – do inglês, *manner*).

Typical pattern:

BUYER buys GOODS from SELLER for MONEY

Example:

Abby bought a car from Robin for $5,000.

Figura 5 – Anotação da FrameNet para o verbo *buy*

Thematic Roles:
Agent[+animate OR +organization]
Asset[-location -region]
Beneficiary[+animate OR +organization]
Source[+concrete]
Theme[]

Frames:

Basic Transitive:

"Carmen bought a dress"
Agent V Theme

Benefactive Alternation (double object):

"Carmen bought Mary a dress"
Agent V Beneficiary Theme

Figura 6 – Anotação da VerbNet para o verbo *buy*

Roles:
Arg0:buyer
Arg1:thing bought
Arg2:seller
Arg3:price paid
Arg4:benefactive

Examples:

Intransitive:

Consumers who buy at this level are more educated than they were.

Arg0: Consumers
REL: buy
ArgM-MNR: at this level

Basic transitive:

They bought $2.4 billion in Fannie Mae bonds
Arg0: They
REL: bought
Arg1: $2.4 billion in Fannie Mae bonds

Figura 7 – Anotação do PropBank para o verbo *buy*

Estes repositórios têm sido construídos manualmente, dada a dificuldade e subjetividade da tarefa. Os problemas dessa abordagem de construção são: individualmente, ela é custosa, pois precisa de humanos treinados, consome muito tempo e está sujeita à inserção de dados errados, inconsistentes ou incompletos; unificando-se as informações dos repositórios, podem-se notar diferentes níveis de abstração na anotação (por exemplo, para descrever o objeto comprado – vide Figuras 5, 6 e 7, tem-se *goods* na FrameNet, *theme* na VerbNet e *thing bought* no PropBank) e diferentes esquemas de anotação (por exemplo, o PropBank distingue os argumentos dos adjuntos em suas estruturas, enquanto a FrameNet e a VerbNet não).

Alguns trabalhos propuseram formas automáticas (por exemplo, Brent, 1991; Resnik, 1992; Grishman e Sterling, 1992; Manning, 1993; Framis, 1994; Briscoe e Carroll, 1997) e semi-automáticas (por exemplo, Green et al., 2004; Gomez, 2004) para a derivação das estruturas argumentais dos verbos. No geral, estes trabalhos fazem uso de *parsers* e/ou dicionários de subcategorização, para identificar os argumentos de um verbo em uma sentença, ou assumem como conhecidos os tipos da estrutura que um verbo pode possuir (em termos de número e ordem de argumentos). Alguns trabalhos (por exemplo, Grishman e Sterling, 1994; Framis, 1994; Gomez, 2004) também tentam fazer generalizações nas estruturas aprendidas, calculando a similaridade entre as palavras de diferentes estruturas ou usando recursos lexicais como a WordNet. A maioria deles também possui algum passo de filtragem dos resultados, no qual algumas estruturas aprendidas são descartadas manualmente ou automaticamente (por meio de medidas baseadas em freqüência). Abordagens mais recentes (por exemplo, Rooth et al., 1999; McCarthy, 2000;

Gildea, 2002) são baseadas em modelos probabilísticos e tentam agrupar verbos de comportamento similar no que diz respeito aos argumentos que estes requerem. Estes modelos também baseiam-se no uso de *parsers* e não são capazes de produzir por si só as estruturas argumentais completas dos verbos.

As desvantagens destas abordagens são claras: apenas algumas línguas possuem bons *parsers*, WordNets ou dicionários de subcategorização disponíveis; é necessário que se saibam os tipos de estrutura que um verbo possui; qualquer análise manual é custosa nesse cenário. Diante disto, foi proposto, neste trabalho, um modelo estatístico para aprendizado das estruturas argumentais dos verbos, o qual supera as dificuldades apontadas, produzindo resultados compatíveis com o estado da arte.

3.6.2. Um modelo para aprendizado não supervisionado de estruturas argumentais

Na Figura 8, o modelo proposto é esquematizado.

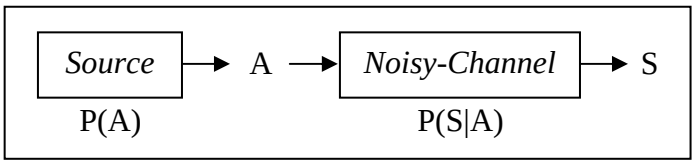


Figura 8 – Modelo *Noisy-Channel* para aprendizado das estruturas argumentais dos verbos

Nesse modelo, assume-se que uma estrutura argumental A qualquer é produzida por uma fonte com probabilidade P(A) e que, após ser corrompida em um canal com ruído, A se transforma na sentença S com probabilidade P(S|A). Nesse modelo, explica-se, portanto, como uma estrutura argumental é realizada superficialmente em uma sentença S. A história gerativa para isso é mostrada abaixo, exemplificada para a derivação da sentença “O menino comprou um brinquedo”.

1) produz-se a estrutura argumental comprou(menino,brinquedo) com probabilidade arg(comprou(menino,brinquedo))

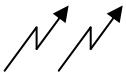
comprou(menino,brinquedo)

2) determina-se que “comprou” produzirá duas palavras extras (isto é, adjuntos) com probabilidade phi(2|comprou)

comprou(menino,brinquedo)

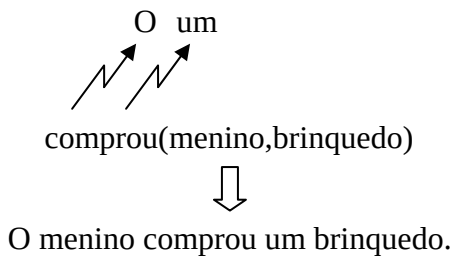
3) escolhem-se as palavras extras “O” e “um” com probabilidades ew(O) e ew(um), respectivamente

O um



comprou(menino,brinquedo)

4) as palavras são reordenadas para produzir a sentença com distribuição uniforme



Portanto, os parâmetros deste modelo que devem ser aprendidos são três: arg, phi e ew. O que se objetiva conseguir, de fato, são os parâmetros arg mais prováveis para cada verbo da língua. Note que os parâmetros phi e ew são essenciais para o modelo, pois é por meio deles que o modelo é capaz de diferenciar os argumentos dos adjuntos. Entretanto, concluído o aprendizado, os parâmetros phi e ew são dispensáveis para a tarefa em foco, pois os parâmetros arg codificam toda a informação desejada.

O modelo é treinado por meio do método EM, já explicado anteriormente. Neste contexto, todas as possíveis estruturas argumentais para uma sentença são consideradas. A Figura 9 mostra todas as estruturas possíveis para a sentença "Ele comprou presentes para ela". As palavras apontadas pelas setas são os argumentos da estrutura argumental; as palavras não apontadas são as palavras extras.

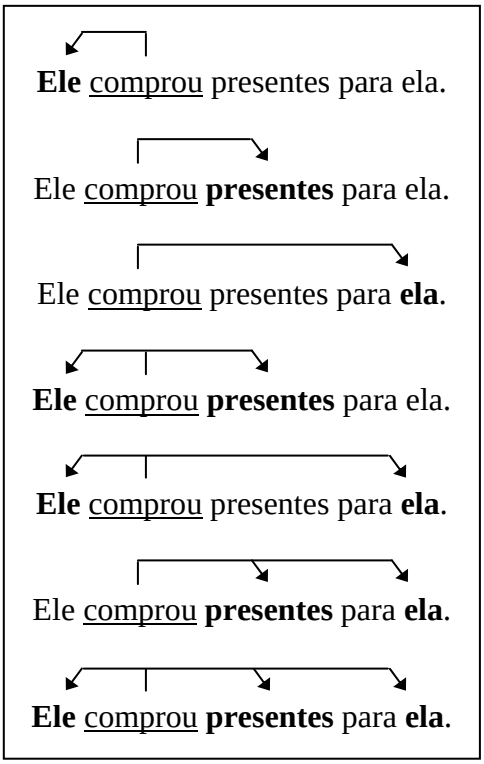


Figura 9 – Estruturas argumentais possíveis para a sentença "Ele comprou presentes para ela."

Para tornar o aprendizado mais eficiente, as seguintes restrições foram adotadas (como pode ser observado na figura): uma estrutura argumental pode ter, no

máximo, 3 argumentos; os argumentos podem ser somente palavras de classe aberta, isto é, substantivos (incluindo pronomes), adjetivos, verbos e advérbios. Para se identificar as palavras de classe aberta, um *tagger* é utilizado, como explicado na subseção seguinte, a qual descreve o corpus sobre o qual o modelo foi treinado.

3.6.3. Corpus

Para formar o corpus de treinamento do modelo, foram coletadas todas as sentenças da coleção TREC'2002 para 20 verbos selecionados aleatoriamente. Devido à complexidade exponencial do método EM, foram selecionadas sentenças de tamanho máximo de 10 palavras. As duas primeiras colunas da Tabela 2 exibem os verbos selecionados e o número de sentenças obtido para cada verbo. Em média, para cada verbo, foram coletadas 2.402 sentenças.

Para identificação das palavras de classe aberta, as únicas que podem ser argumentos (segundo a restrição adotada no modelo proposto), as sentenças foram etiquetadas por um *tagger* (Ratnaparki, 1996). Além disso, para tornar o modelo apto a fazer generalizações, optou-se por dados anotados por um reconhecedor de entidades mencionadas (REM), isto é, um sistema que identifica, em uma sentença, as palavras que representam entidades das classes "organização", "data", "lugar", "pessoa", etc. Foi por esta razão que as sentenças do corpus de treinamento foram coletadas da coleção TREC'2002, pois esta já havia sido anotada pelo REM *BBN IdentiFinder* (Bikel et al., 1999). Outras modificações que foram feitas nos dados são:

- todos os numerais foram substituídos pela classe genérica *number*;
- com exceção dos pronomes *it*, *they* e *them*, os pronomes foram substituídos pela classe genérica *person*; os pronomes *it*, *they* e *them* foram considerados como podendo ser tanto da classe *person*, quanto da classe *thing* (isto é, qualquer coisa que não seja *person*), já que eles podem se referir a qualquer entidade.

A Figura 10 exibe uma amostra dos dados de treinamento do modelo, com as etiquetas morfossintáticas após a barra e as entidades mencionadas em negrito.

about/IN ***money***/NN *home*/NN *water*/NN *heaters*/NNS *are*/VBP *bought*/VBN *each*/DT *year*/NN

person/PRP *bought*/VBD ***thing***/PRP ***number***/CD *years*/NNS *ago*/RB

organization/NNP *bought*/VBD ***organization***/NN *from*/IN ***organization***/NN last/JJ *year*/NN

thing/PRP *bought*/VBD *the*/DT *outstanding*/JJ *shares*/NNS *on*/IN ***date***/NNP

the/DT *cafeteria*/NN *bought*/VBD *extra*/JJ *plates*/NNS

Figura 10 – Amostra dos dados de treinamento do modelo de aprendizado de estruturas argumentais

Algumas coisas interessantes de se notar são: algumas sentenças são completamente lexicalizadas, sem entidades mencionadas (por exemplo, última sentença da figura); algumas sentenças possuem várias entidades mencionadas (por exemplo, segunda

sentença da figura); na primeira sentença da figura, o termo *8 million* foi classificado erroneamente pelo REM como sendo da classe *money*. Esses erros não interferem no aprendizado, pois, por serem pouco freqüentes, são naturalmente descartados pelo método EM.

Na próxima subseção, relata-se a avaliação do modelo proposto.

3.6.4. Avaliação e discussão

Para avaliar o modelo, as medidas clássicas de precisão e cobertura foram calculadas. Precisão, neste caso, indica quantas das estruturas aprendidas pelo modelo são corretas; cobertura indica quantas das estruturas que deviam ser aprendidas foram aprendidas, de fato, pelo modelo.

Para o cálculo da precisão, um juiz humano julgou a plausibilidade das estruturas aprendidas. Nos casos em que tinha dúvidas sobre a possibilidade de ocorrência de alguma estrutura, o juiz podia consultar as sentenças do corpus de treinamento que deram origem às estruturas. Para o cálculo da cobertura, o repositório utilizado como referência foi o PropBank. Neste caso, procurou-se por estruturas aprendidas similares às estruturas previstas pelo PropBank. Em ambas as medidas, consideraram-se somente as estruturas com probabilidade maior do que o *threshold* de 10^{-6}. Os resultados destas avaliações são apresentados na última coluna da Tabela 2. A terceira coluna da tabela mostra o número de estruturas avaliado para cada verbo. Na média, o modelo atingiu precisão de 76% e cobertura de 86%.

Tabela 2 – Desempenho do modelo de aprendizado de estruturas argumentais

Verbo	**Num. de sentenças**	**Num. de estruturas argumentais**	**Precisão e Cobertura (%)**
ask	3179	212	P=83, C=100
begin	4042	166	P=81, C=50
believe	910	45	P=87, C=100
buy	1106	75	P=79, C=80
cause	957	23	P=74, C=100
change	2648	152	P=88, C=100
die	4154	257	P=72, C=100
earn	952	72	P=83, C=100
expect	6039	422	P=73, C=75
find	4679	210	P=82, C=100
give	3581	249	P=67, C=100
help	1663	53	P=75, C=40
kill	3253	240	P=54, C=100
like	968	74	P=69, C=100
move	2118	162	P=71, C=60
offer	1599	59	P=85, C=67
pay	2076	142	P=76, C=88
raise	1268	79	P=71, C=50
sell	1538	98	P=81, C=100
spend	1322	62	P=61, C=100
Média	**2402**	**142**	**P=76, C=86**

Calcularam-se, também, as medidas de precisão e cobertura para as 10 estruturas mais prováveis (Top 10) e para as 20 estruturas mais prováveis (Top 20). A Tabela 3 mostra os resultados dessa avaliação. Como esperado, quanto mais estruturas são consideradas na avaliação, menor é a precisão (pois mais estruturas com menor probabilidade são consideradas) e maior é a cobertura.

Tabela 3 – Desempenho do modelo para as Top 10 e Top 20 sentenças

	Precisão (%)	**Cobertura (%)**
Top 10	93	36
Top 20	89	46
All	76	86

Para validar a avaliação feita, pediu-se a três juízes que avaliassem a precisão para uma amostra de três verbos (selecionados aleatoriamente), a fim de se verificar a concordância entre eles e, portanto, a própria viabilidade da tarefa. A Tabela 4 mostra os resultados dessa avaliação para as 10 e 20 sentenças mais prováveis. A medida de concordância Kappa (Carletta, 1996) foi calculada em 0.77. Um valor entre 0.6 e 0.8 indica uma boa concordância.

Tabela 4 – Avaliação conjunta de alguns verbos

Verbo	**Juiz 1**		**Juiz 2**		**Juiz 3**	
	Top 10 (%)	Top 20 (%)	Top 10 (%)	Top 20 (%)	Top 10 (%)	Top 20 (%)
ask	89	89	89	89	89	95
buy	90	95	100	100	90	89
cause	89	87	89	87	100	94

Na geral, os erros detectados nas estruturas argumentais durante o cálculo da precisão foram causados pelos seguintes problemas:

- presença de advérbios nas estruturas argumentais: idealmente, advérbios não deveriam estar presentes nas estruturas argumentais dos verbos, pois são considerados adjuntos; entretanto, em sentenças como *He asked rhetorically* e *He asked incredulously*, os advérbios *rhetorically* e *incredulously* parecem essenciais para o significado das sentenças e, além disso, ocorrem tão freqüentemente com esse verbo, que foram considerados argumentos durante o aprendizado
- ocorrência de *phrasal verbs* no corpus de treinamento (isto é, verbos que, associados a algumas partículas, adquirem sentidos diversos): o método de aprendizado não é capaz de distinguir esse tipo de fenômeno lingüístico dos verbos comuns, acarretando o aprendizado errôneo de estruturas como *gave*(*he*,*up*) e *gave*(*he*) para a sentença *He gave up*, por exemplo, sendo que ambas são consideradas estruturas inadequadas

Em relação ao cálculo da cobertura, os seguintes problemas foram notados:

- algumas das palavras encontradas nas estruturas previstas no PropBank não possuíam palavras e/ou entidades correspondentes nas estruturas aprendidas automaticamente

- em alguns casos, as estruturas no PropBank contêm mais do que 3 argumentos, o que ocorre pela inclusão de adjuntos nas estruturas (vale lembrar que, no modelo proposto neste trabalho, consideram-se, no máximo, 3 argumentos)

Como ilustração, a Figura 11 mostra as 10 estruturas mais prováveis aprendidas pelo modelo proposto para o verbo *buy*.

1	*buy*(*organization*,*organization*)	1.20e-01
2	*buy*(*person*,*number*)	8.44e-02
3	*buy*(*person*,*thing*)	7.10e-02
4	*buy*(*organization*,*thing*)	5.63e-02
5	*buy*(*person*,*organization*)	4.28e-02
6	*buy*(*organization*,*person*)	3.51e-02
7	*buy*(*person*,*house*)	1.54e-02
8	*buy*(*person*,*thing*,*anyway*)	1.54e-02
9	*buy*(*money*,*money*)	1.40e-02
10	*buy*(*organization*,*organization*,*date*)	8.63e-03

Figura 11 – Estruturas argumentais mais prováveis aprendidas para o verbo *buy*

Algumas coisas interessantes de se notar nessas estruturas são:

- as estruturas 5 e 6 são similares: na primeira, uma pessoa compra uma organização (voz ativa); na segunda, uma organização é comprada por uma pessoa (voz passiva)
- a estrutura 7 possui um item lexicalizado (*house*)
- na estrutura 8, há um erro causado pela inclusão de um advérbio como argumento (*anyway*) (porque ele co-ocorreu muito freqüentemente com o verbo no corpus de treinamento)
- na estrutura 9, há um erro causado pelo *phrasal verb buy down* (como em *the dollar bought down the yen*) (como já discutido, o modelo não é capaz de distinguir os *phrasal verbs*)

Para muitos verbos, o modelo foi capaz de aprender sentidos não previstos no PropBank. Por exemplo, para o verbo *raise*, foram aprendidas estruturas para seu sentido de "crescer" (como em *Peter was raised in a big city*). Também se aprenderam diversas variações do uso dos verbos não listadas no PropBank. Por exemplo, para o verbo *die*, tem-se:
(a) In *date*, *person* died.
(b) *Person* died in *date*.
(c) *Person* died in *date* in *location*.
(d) *Person* died in *location* in *date*.

Além destas vantagens, o modelo proposto se destaca pelos seguintes pontos: (a) seu aprendizado é completamente automático, (b) por se basear em corpus, evidências lingüísticas são fornecidas para cada estrutura argumental aprendida, (c) o nível de abstração mais apropriado (lexemas vs. entidades) é determinado automaticamente de forma consistente, (d) não se faz necessário o uso de ferramentas sofisticadas de processamento de língua natural e (e) além das estruturas argumentais, tem-se probabilidades associadas a elas. Devido a (b), repositórios de informações semânticas para verbos (como FrameNet, VerbNet e PropBank) podem ser

produzidos automaticamente para línguas que não os têm ou podem ser complementados para as línguas que já os tem. Por causa de (e), é possível utilizar o conhecimento aprendido em diversas aplicações, por exemplo: em tradução automática, as probabilidades das estruturas argumentais podem ser associadas às sentenças decodificadas, melhorando o resultado do processo; em *parsers*, pode-se selecionar as análises mais prováveis com base nas probabilidades das estruturas argumentais.

4. Comentários finais

Neste relatório, foram relatadas as atividades de pesquisa desenvolvidas durante o doutorado sanduíche realizado no *Information Sciences Institute*/*University of Southern Califórnia*, sob a supervisão do Prof. Dr. Daniel Marcu. Essas atividades fizeram parte do projeto de doutorado que visa à investigação e desenvolvimento de técnicas de análise discursiva automática e produção de um analisador inédito para o português do Brasil.

Os modelos estatísticos apresentados neste relatório são o resultado de uma série de testes com uma grande variedade de modelos em busca dos modelos mais apropriados para as tarefas que se objetivou tratar. Os próximos passos desta pesquisa são: (a) aprendizado automático das estruturas argumentais dos verbos mais freqüentes do inglês; (b) treinamento e teste do novo modelo de análise discursiva baseado em estruturas argumentais, com base nas estruturas argumentais aprendidas no item anterior; (c) replicação dos experimentos relatados para o português do Brasil. Em relação ao item (a), os 2.000 verbos mais freqüentes do inglês já foram selecionados e as sentenças correspondentes já coletadas da coleção TREC'2002.

Para o português, os dados de treinamento dos modelos serão coletados do corpus do NILC (Pinheiro e Aluísio, 2003). Como resultado do aprendizado das estruturas argumentais dos verbos para essa língua, será produzido, também, o primeiro repositório em larga escala de informação semântica dos verbos.

Para mais detalhes sobre este projeto de doutorado, vide http://www.nilc.icmc.usp.br/~thiago/DiZer.html

Referências

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Baker, C.F.; Fillmore, C.J.; Lowe, J.B. (1998). The Berkeley FrameNet project. In the *Proceedings of COLING/ACL*, pp. 86-90, Montreal.

Bikel, D.M.; Schwartz, R.; Weischedel, R.M. (1999). An Algorithm that Learns What's in a Name. *Machine Learning* (Special Issue on NLP).

Brent, M.R. (1991). Automatic acquisition of subcategorization frames from untagged text. In the *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 209-214, Berkeley, CA.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In the *Proceedings of the 5th ANLP Conference*, pp. 356-363, Washington, D.C.

Burstein, J.; Marcu, D.; Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, pp. 32-39.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 22, N. 2, pp. 249–254.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD Thesis, University of Pennsylvania.

Corston-Oliver, S.H. (1998a). *Computing Representations of the Structure of Written Discourse*. Ph.D. Thesis, University of California, Santa Barbara.

Corston-Oliver, S.H. (1998b). Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. In E. Hovy and D. Radev (eds.), *Papers from the 1998 Spring Symposium*, pp. 9-15.

Corston-Oliver, S.H. (1998c). Identifying the linguistic correlates of rhetorical relations. In M. Stede, L. Wanner and E. Hovy (eds.), *Discourse Relations and Discourse Markers: Proceedings of the Workshop*, pp. 8-14.

Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory. An Approach to Global Cohesion and Coherence. In the *Proceedings of COLING*.

Dempster, A.P.; Laird, N.M.; Rugin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, Vol. 39, pp. 1-38.

Fellbaum, C. (1999). *Wordnet: an electronic lexical database*. The MIT Press. Cambridge.

Feltrim, V.D.; Pelizzoni, J.M.; Teufel, S.; Nunes, M.G.V.; Aluísio, S.M. (2004). Applying Argumentative Zoning in an Automatic Critiquer of Academic Writing. In *Proceedings of SBIA2004*.

Feltrim, V.D. and Teufel, S. (2004). Automatic Critiquing of Novices' Scientific Writing Using Argumentative Zoning. In *Proceedings of AAAI-EAAT 2004*, AAAI Press Technical Report, SS-04-07.

Framis, F.R. (1994). An experiment on learning appropriate selection restrictions from a parsed corpus. In the *Proceedings of the International Conference on Computational Linguistics*, Kyoto, Japan.

Gildea, D. (2002). Probabilistic Models of Verb-Argument Structure. In the *Proceedings of the 17th International Conference on Computational Linguistics*.

Gomez, F. (2004). Building Verb Predicates: A Computational View. In the *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pp. 359-366, Barcelona, Spain.

Green, R.; Dorr, B.J.; Resnik, P. (2004). Inducing Frame Semantic Verb Classes from WordNet and LDOCE. In the *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pp. 375-382, Barcelona, Spain.

Grishman, R. and Sterling, J. (1992). Acquisition of selectional patterns. In the *Proceedings of the International Conference on Computational Linguistics*, pp. 658-664, Nantes, France.

Grishman, R. and Sterling, J. (1994). Generalizing Automatically Generated Selectional Patterns. In the *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, N. 3.

Hovy, E. (1988). *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

Jordan, M. P. (1992). An Integrated Three-Pronged Análisis of a Fund-Raising Letter. In W. C. Mann and S. A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen and M. Stokhof (eds.), *Formal Methods in the Study of Language*, pp. 277-322. Mathematisch Centrum Tracts, Amsterdam.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In the *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas.

Kipper, K.; Dang, H.T.; Palmer, M. (2000). Class-based Construction of a Verb Lexicon. In the *Proceedings of AAAI 17th National Conference on Artificial Intelligence*. Austin, Texas.

McCarthy, D. (2000). Using semantic preferences to identify verbal participation in role switching alternations. In the *Proceedings of the 1st NAACL*, pp. 256-263, Seattle, Washington.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Manning, C.D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In the *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 235-242, Columbus, Ohio.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Marcu, D. (1997a). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1997b). From local to global coherence. A bottom-up approach to text planning. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 629-635. Providence, Rhode Island.

Marcu, D. (1998a). *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1998b). *Instructions for Manually Annotating the Discourse Structures of Texts*. Information Sciences Institute, University of Southern California.

Marcu, D. (1999a). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 365-372. Maryland.

Marcu, D. (1999b). A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In *Proceedings of the Workshop on Levels of Representation in Discourse*, pp. 101-108. Edinburgh, Scotland.

Marcu, D. (2000a). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA. The MIT Press.

Marcu, D. (2000b). Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures à la Grosz and Sidner. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Saarbrueken.

Marcu, D.; Carlson, L.; Watanabe, M. (2000). The Automatic Translation of Discourse Structures. In the *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL'2000), Seattle, Washington.

Maybury, M.T. (1992). Communicative Acts for Explanation Generation. *Int. Journal of Man-Machine Studies 37*, pp. 135-172.

Moore, J.D. and Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol. 19, N. 4, pp. 651-694.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, Vol. 22, N. 3, pp. 409-419.

O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Germany.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Pardo, T.A.S. e Nunes, M.G.V. (2003). *Análise de Discurso: Teorias Discursivas e Aplicações em Processamento de Línguas Naturais*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 196.

Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In the *Proceedings of the XVII Brazilian Symposium on Artificial Intelligence* - SBIA2004. São Luís, Maranhão, Brazil.

Pardo, T.A.S. e Nunes, M.G.V. (2004). *Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 231.

Pinheiro, G.M. e Aluísio, S.M. (2003). *Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web*. Série de Relatórios Técnicos do Instituto de Ciencias Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.

Ratnaparkhi, A. (1996). A Maximum Entropy Part-of-Speech Tagger. In the *Proceedings of the 1st Empirical Methods in Natural Language Processing Conference*. Philadelphia.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge, University Press.

Resnik, P. (1992). Wordnet and distributional analysis: a class-based approach to lexical discovery. In the *Proceedings of AAAI Workshop on Statistical Methods in NLP*.

Rino, L.H.M. (1996a). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-Usp. São Carlos - SP.

Rino, L.H.M. (1996b). A sumarização automática de textos em português. In *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado*, pp. 109-119. Curitiba - PR.

Rooth, M.; Stefan, R.; Prescher, D.; Carroll, G.; Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In the *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104-111, College Park, Maryland.

Schauer, H. (2000). Referential Structure and Coherence Structure. In the *Proceedings of the TALN*.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, pp. 235-255.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, Vol. 27, N. 3, pp. 379-423.