Um Modelo Estatístico Gerativo para o Aprendizado Não Supervisionado das Estruturas Argumentais dos Verbos[1]

Thiago Alexandre Salgueiro Pardo[*], Daniel Marcu[+] e Maria das Graças Volpe Nunes[*]

[*]Núcleo Interinstitucional de Lingüística Computacional (NILC)
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br

[+]Information Sciences Institute (ISI)
4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292
http://www.isi.edu

{thiago@nilc.icmc.usp.br; marcu@isi.edu; gracan@icmc.usp.br}

***Abstract**. This paper presents a statistical generative model for unsupervised learning of verb argument structures. The model is based on the noisy-channel model and is trained with the Expectation-Maximization algorithm. The model was used to induce the argument structures for the 1.500 most frequent verbs in English. The evaluation of a sample of this verb set showed that about 80% of the structures were considered plausible by humans. The structures also show correct patterns of verb usage not present in PropBank, a manually developed semantic resource for verbs.*

Resumo. Apresenta-se, neste artigo, um modelo estatístico gerativo para o aprendizado não supervisionado das estruturas argumentais dos verbos. O modelo proposto baseia-se no modelo *noisy-channel* e é treinado por meio do algoritmo *Expectation-Maximization*. O modelo foi usado para induzir as estruturas argumentais dos 1.500 verbos mais freqüentes da língua inglesa. A avaliação de uma amostra deste conjunto de verbos indicou que 80% destas estruturas foram consideradas plausíveis por humanos. Estas estruturas também apresentaram padrões de uso verbal não previstos no *PropBank*, um repositório de informação semântica para verbos construído manualmente.

1. Introdução

Inspirando-se no impacto causado pela construção e disponibilidade do *Penn Treebank* (Marcus et al., 1993; Marcus, 1994), um conjunto de sentenças em inglês sintaticamente anotadas, muitos esforços têm sido feitos para a criação de repositórios semanticamente anotados. A anotação dos argumentos dos verbos, seus papéis temáticos e comportamentos lingüísticos preferenciais (Levin, 1993) representa uma parcela significativa destes esforços.

O *Penn Treebank* permitiu um avanço considerável no estado da arte em Processamento de Línguas Naturais (PLN) ao possibilitar o desenvolvimento e/ou aprimoramento de ferramentas como *taggers* (etiquetadores morfossintáticos) e *parsers* (analisadores sintáticos) e, conseqüentemente, das aplicações baseadas nestas ferramentas. Espera-se que um repositório semântico proporcione avanços similares, permitindo o desenvolvimento de aplicações mais informadas e sofisticadas. As anotações semânticas focadas neste artigo são as estruturas argumentais dos verbos. Essas estruturas codificam informações conceituais subjacentes ao uso dos verbos, indicando quantos e quais são os possíveis argumentos que os verbos requerem quando os eventos indicados por estes são realizados superficialmente na forma de sentenças.

[1] Este trabalho foi apoiado pelas agências de fomento à pesquisa FAPESP, CAPES, CNPq e Comissão Fulbright.

Várias questões são consideradas problemáticas para a identificação das estruturas argumentais dos verbos. Considere as sentenças (1)-(3) abaixo, as quais expressam padrões de uso diferentes para o verbo em inglês *buy*:
(1) *John bought gifts for them.*
(2) *He bought it 40 years ago.*
(3) *About 8 million home water heaters are bought each year.*

Pode-se perceber que:
- há uma grande variedade de argumentos possíveis para um verbo (por exemplo, em (1), o agente da ação é *John* ; em (2), é *He*);
- os argumentos podem estar realizados em posições e de formas diferentes nas sentenças (em (1), o objeto comprado – *gifts* – é realizado depois do verbo; em (3), o objeto – *About 8 million home water heaters* – é realizado antes do verbo);
- não é simples diferenciar argumentos, que são obrigatórios nas estruturas argumentais, de adjuntos, isto é, termos opcionais (em (2), a expressão de tempo – *40 years ago* – não parece ser obrigatória, pois não aparece na sentença (1));

Além disso, tem-se que:
- é difícil saber com exatidão o número de estruturas argumentais possíveis para um verbo;
- é desejável que se determine os possíveis rótulos ontológicos dos argumentos em uma estrutura, para maior generalidade desta; entretanto, não é simples determinar os graus de abstração mais apropriados para os argumentos de forma que estes sejam representativos e reutilizáveis para todos os verbos.

Idealmente, todas as possibilidades de estruturas argumentais deveriam ser incluídas na especificação semântica dos verbos. Visando a auxiliar esta tarefa, apresenta-se, neste artigo, uma abordagem não supervisionada, completamente automática, para o aprendizado das estruturas argumentais. Propõe-se um modelo estatístico gerativo baseado no modelo *noisy-channel* (Shannon, 1948) e treinado por meio do algoritmo *Expectation-Maximization* (Dempster et al., 1977). Este modelo foi treinado com textos reais para a indução das estruturas argumentais dos 1.500 verbos mais freqüentes do inglês. Como resultado, foi produzido um repositório nomeado *ArgBank*, que deverá servir de base para diversas aplicações de PLN.

Na seção seguinte, uma breve revisão da literatura é apresentada. O modelo estatístico proposto e o algoritmo de treinamento são descritos na Seção 3. Introduzem-se, na Seção 4, os dados utilizados para o treinamento. Na Seção 5, apresenta-se a avaliação (humana) das estruturas argumentais aprendidas automaticamente. Por fim, na Seção 6, discutem-se as potencialidades e limitações do modelo e os trabalhos futuros.

2. Trabalhos correlatos

Alguns grandes projetos visam a desenvolver repositórios de informação semântica para os verbos. Destacam-se a *FrameNet* (Baker et al., 1998), a *VerbNet* (Kipper et al., 2000) e o *PropBank* (Kingsbury and Palmer, 2002). Estes repositórios são para a língua inglesa e foram desenvolvidos manualmente. Nas Figuras 1, 2 e 3, mostram-se recortes das informações associadas ao verbo *buy* nestes três repositórios. Na *FrameNet*, mostra-se o padrão no qual o verbo ocorre com exemplos representativos; além disso, os verbos são organizados em uma hierarquia que codifica implicitamente como os argumentos dos verbos podem ser herdados de ancestrais. Na *VerbNet*, mostram-se os papéis temáticos que o verbo exige, suas características semânticas e possíveis esquemas de subcategorização; também são exibidos exemplos para cada esquema de subcategorização possível. No *PropBank*, mostram-se os papéis dos argumentos dos verbos, os possíveis esquemas de subcategorização e exemplos para cada um; distinguem-se, também, os argumentos dos adjuntos (na Figura 3, o adjunto é marcado pela extensão MNR,

indicando modo – *manner*, em inglês). É interessante notar que: o *PropBank* é o único repositório que diferencia argumentos de adjuntos; como discutido anteriormente, há baixa concordância sobre o status ontológico dos rótulos dos argumentos (no *PropBank*, o primeiro argumento – ARG1 – é *thing bought*; na *FrameNet*, é *goods*; na *VerbNet*, é *theme*).

Typical pattern:
BUYER buys GOODS from SELLER for MONEY

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

Figura 1 – Anotação do verbo *buy* na *FrameNet*

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 2 – Anotação do verbo *buy* na *VerbNet*

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 3 – Anotação do verbo *buy* no *PropBank*

Algumas pesquisas têm investigado o problema da determinação automática (Brent, 1991; Resnik, 1992; Grishman and Sterling, 1992; Manning, 1993; Framis, 1994; Briscoe and Carroll, 1997; Rooth et al., 1999; McCarthy, 2000; Sarkar and Zeman, 2000; Merlo and Stevenson,

2001; Sarkar and Tripasai, 2002; Gildea, 2002) e semi-automática (Korhonen, 2002; Green et al., 2004; Gomez, 2004) das estruturas argumentais dos verbos (incluindo a tarefa correlata de determinação dos esquemas de subcategorização dos verbos). Em geral, estas abordagens baseiam-se em informações sintáticas e/ou dicionários de subcategorização para identificar os argumentos de um verbo em uma sentença, e/ou assumem como conhecidos os tipos de estruturas de um verbo no que se refere ao número e à ordem de seus argumentos. O principal objetivo destas pesquisas é identificar os lexemas mais prováveis para serem argumentos dos verbos. Outras pesquisas (Grishman and Sterling, 1994; Framis, 1994; Lapata, 1999; Korhonen, 2002; Gomez, 2004) vão além dos lexemas e realizam operações de generalização sobre as estruturas aprendidas, calculando a similaridade entre argumentos de estruturas semelhantes ou usando repositórios de informação lexical, como a *WordNet* (Fellbaum, 1998) e as classes verbais de Levin (1993). A maioria destas abordagens implementa uma etapa de filtragem das estruturas aprendidas, na qual estruturas inadequadas são descartadas manualmente ou automaticamente com base em medidas de freqüência.

Neste artigo, propõe-se uma abordagem diferente das existentes para o problema de determinação das estruturas argumentais dos verbos. O modelo proposto, descrito na seção seguinte, possui as seguintes características principais:

- não se assume que os tipos de estruturas argumentais dos verbos são conhecidos; ao contrário, as estruturas são aprendidas automaticamente a partir de textos reais;
- em adição ao aprendizado das estruturas mais prováveis, a abordagem proposta também ordena essas estruturas de acordo com suas probabilidades (não apenas se sabe que duas estruturas são prováveis para um mesmo verbo, mas que uma estrutura é mais provável do que a outra);
- usam-se ferramentas simples, a saber, um *tagger* e um reconhecedor de entidades mencionadas (REM), para anotar os dados;
- pelo uso do REM, é possível generalizar as estruturas, aprendendo abstrações (entidades) em vez de lexemas, quando isso se mostra apropriado.

3. Descrição do modelo

O aprendizado de estruturas argumentais foi modelado probabilisticamente com base no modelo *noisy-channel* (Shannon, 1948). Recentemente, o modelo *noisy-channel* passou a ser aplicado para uma grande gama de problemas em PLN, avançando consideravelmente o estado da arte. Vide, por exemplo, os trabalhos recentes de tradução automática (Brown et al., 1990, 1993; Koehn et al., 2003), perguntas e respostas (Soricut and Brill, 2004) e sumarização de textos (Daumé and Marcu, 2004). Diz-se que este modelo é gerativo pelo fato de se basear em uma história gerativa de como os dados são produzidos/transformados. Em tradução automática, há uma história de como uma sentença em inglês se transforma em/gera uma sentença em português, por exemplo; em perguntas e respostas, explica-se como a pergunta gera a resposta; em sumarização, mostra-se como um sumário pode se transformar no texto ao qual se refere. Neste trabalho, explica-se como uma sentença é gerada a partir de sua estrutura argumental, como especificado a seguir. Para mais detalhes sobre o modelo *noisy-channel*, sugere-se a leitura de Marcu e Popescu (2005).

Assume-se que sentenças em língua natural são produzidas pelo seguinte processo gerativo estocástico:

1. (a) o verbo v da sentença é escolhido com probabilidade P(v); (b) determina-se que o verbo v requer M argumentos com probabilidade narg(M|v); (c) cada argumento é gerado com probabilidade arg(argumento|v), podendo ser uma abstração (entidade mencionada, neste caso) ou um lexema;
2. a partir da estrutura argumental gerada no passo anterior, determina-se que N abstrações/palavras extras (isto é, que não são argumentos) devem ser produzidas com probabilidade phi(N|v);

3. cada abstração/palavra extra p é produzida estocasticamente com probabilidade ew(p);
4. cada abstração c produzida durante esse processo é mapeada em uma palavra p com probabilidade t(p|c);
5. todas as palavras (provenientes das estruturas argumentais e do conjunto de palavras extras) são reordenadas para produzir uma sentença gramatical.

Por exemplo, a sentença *John bought gifts for them* é gerada pelo seguinte processo:

1. (a) o verbo *bought* é escolhido com probabilidade P(*bought*); (b) a esse verbo são associados três argumentos com probabilidade narg(3|*bought*), (c) que são *PERSON1*, *gifts* e *PERSON2* com probabilidades arg(*PERSON1*|*bought*), arg(*gifts*|*bought*) e arg(*PERSON2*|*bought*), respectivamente, produzindo a estrutura argumental *bought*(*PERSON1*,*gifts*,*PERSON2*);
2. uma palavra extra é produzida com probabilidade phi(1|*bought*);
3. a palavra extra *for* é escolhida com probabilidade ew(*for*);
4. as abstrações *PERSON1* e *PERSON2* são mapeadas nas palavras *John* e *them*, respectivamente;
5. as palavras são reordenadas para produzir a sentença *John bought gifts for them*.

Para que o treinamento deste modelo seja possível, algumas simplificações são feitas. Assume-se que a subseqüência de passos 1.a-c ocorre em uma única etapa: um evento verbo(arg_1,...,arg_n) é gerado em um único processo estocástico com probabilidade event(verbo(arg_1,...,arg_n)). Pelo fato de REMs possuírem precisão acima de 90%, assume-se também que não é necessário mapear as abstrações/entidades nos lexemas durante o processo gerativo: as palavras do corpus de treino são substituídas pelas entidades correspondentes e, durante o treinamento do modelo, estruturas generalizadas, isto é, com entidades, são aprendidas diretamente. Do ponto de vista da história gerativa, isto significa a exclusão do passo 4 acima. Outra simplificação consiste em considerar que a reordenação das palavras (passo 5) possui distribuição uniforme, ou seja, qualquer ordem é igualmente provável. Estas escolhas simplificam significativamente o modelo e seu processo de aprendizado.

De acordo com o modelo esquematizado, a probabilidade de uma sentença S é dada pela seguinte fórmula:

$$P(S) = \sum_{A} P(S,A) = \sum_{A} P(A) \times P(S|A) = \sum_{A} \text{event}(A) \times \text{phi}(N|\text{verbo}) \times \prod_{i=1}^{N} \text{ew}(p_i)$$

em que A é uma estrutura argumental possível, N é o número de abstrações/palavras extras geradas e p_i é a abstração/palavra extra de número i sendo gerada. Sob essa perspectiva, a probabilidade da sentença *John bought gifts for them* é dada por event(*bought*(*PERSON1*,*gifts*,*PERSON2*)) × phi(1|*bought*) × ew(*for*).

O algoritmo *Expectation-Maximization* (EM) (Dempster et al., 1977) é usado para estimar os parâmetros do modelo (que são inicializados uniformemente), isto é, os conjuntos de probabilidade event, phi e ew. O algoritmo EM é usualmente utilizado para o treinamento de modelos *noisy-channel* e busca os melhores valores para os parâmetros considerados no problema pelo cômputo exaustivo de todos os possíveis parâmetros e seus valores para uma determinada instância de treinamento. Para mais detalhes sobre o algoritmo EM, sugere-se a leitura da obra de referência ou Pardo e Nunes (2005).

Para tornar o aprendizado mais efetivo, considera-se que uma estrutura argumental pode ter, no máximo, 3 argumentos e que esses argumentos podem ser somente palavras de classe aberta (verbos, advérbios, adjetivos e substantivos) e pronomes. Para isso, o corpus de treino também é etiquetado por um *tagger* (Ratnaparkhi, 1996). Com isso, o algoritmo EM é aplicado e todas as estruturas argumentais possíveis de uma sentença são consideradas no treinamento.

Como exemplo, na Figura 4, as estruturas argumentais possíveis para a sentença *John bought gifts for them* são mostradas. As setas partem do verbo e apontam para os argumentos. As palavras não apontadas são consideradas palavras extras. Por simplicidade, as entidades mencionadas e as etiquetas morfossintáticas não são exibidas.

É importante dizer que, por causa do uso do algoritmo EM, probabilidades baixas são naturalmente atribuídas às estruturas incomuns ou inadequadas. Portanto, o processo de filtragem dos resultados não é necessário.

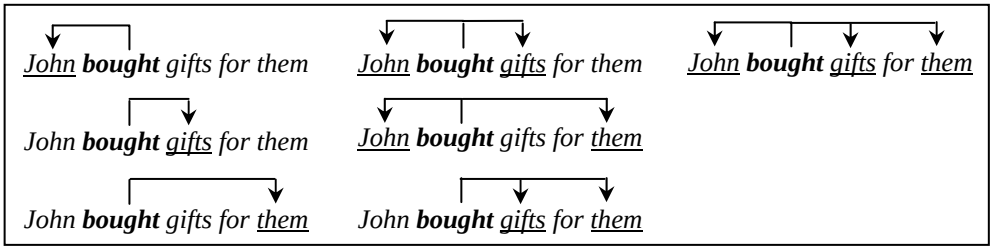

Figura 4 – Estruturas argumentais possíveis para a sentença *John bought gifts for them*

4. Preparação dos dados

Foram extraídas do conjunto de dados da TREC'2002 (*Text REtrieval Conference*) todas as sentenças com os 1.500 verbos mais freqüentes do inglês. Devido às limitações do modelo estatístico proposto para lidar com sentenças longas (aquelas cujos verbos contêm complementos sentenciais, principalmente), foram selecionadas somente as sentenças com 10 palavras, no máximo. Optou-se por utilizar os dados da TREC por estes já estarem anotados por um REM, o *BBN Identifinder* (Bikel et al., 1999). Depois de selecionadas, as sentenças foram etiquetadas morfossintaticamente pelo *tagger*. Outras modificações feitas nos dados são:

- todos os números são substituídos pela entidade genérica *number*;
- com exceção de *it*, *they* e *them*, todos os pronomes pessoais são substituídos pela entidade *person*; *it*, *they* e *them* são considerados tanto *person* quanto a entidade genérica *thing* (isto é, qualquer coisa que não seja *person*), pois estes pronomes podem se referir a qualquer coisa.

O uso de um REM não é necessário para o modelo. Se as entidades não são utilizadas, as estruturas aprendidas são completamente lexicalizadas; caso contrário, aprendem-se estruturas lexicalizadas e generalizadas. Como esperado, o uso das entidades atenua o problema dos dados serem esparsos. É importante notar que o nível mais apropriado de abstração para os argumentos é aprendido automaticamente pelo algoritmo EM.

A Figura 5 mostra uma amostra dos dados anotados, com as entidades mencionadas em negrito e as etiquetas morfossintáticas após a barra. É fácil de se notar que algumas sentenças são completamente lexicalizadas, sem entidades (por exemplo, a última sentença da figura), enquanto outras possuem várias entidades. Na primeira sentença, pode-se notar que um erro foi introduzido pelo REM: o termo *8 million* foi classificado como *money*. Tais erros são naturalmente descartados pelo algoritmo EM por serem pouco freqüentes no corpus.

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*
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 5 – Amostra dos dados de aprendizado

5. Avaliação

Para avaliar se as estruturas argumentais aprendidas são plausíveis ou não, dois experimentos foram realizados.

Para três verbos escolhidos aleatoriamente, foram avaliadas as 20 estruturas argumentais mais prováveis aprendidas pelo modelo proposto. Estas estruturas foram apresentadas a três humanos lingüistas computacionais para, independentemente, julgá-las em termos de sua corretude/plausibilidade. Cada estrutura podia ser classificada como "correta", "incorreta" ou "não é possível dizer". A Tabela 1 exibe a porcentagem de estruturas corretas consideradas por cada juiz para cada verbo. Como se pode notar, por volta de 90% delas foram julgadas corretas. A concordância entre os juízes também foi alta: a estatística kappa (Carletta, 1996) foi de 0.77. Um valor entre 0.60 e 0.80 indica alta concordância.

Tabela 1 – Corretude/plausibilidade das 20 estruturas argumentais mais prováveis

Verbo	**Juiz 1**	**Juiz 2**	**Juiz 3**
ask	89	89	95
buy	95	100	89
cause	87	87	94

Para um melhor entendimento das potencialidades e limitações do modelo proposto, um segundo experimento foi realizado. Para um conjunto de 20 verbos escolhidos aleatoriamente (incluindo os três anteriores), as estruturas argumentais foram comparadas com as estruturas previstas pelo *PropBank*, um dos grandes repositórios construídos manualmente para a especificação semântica dos verbos. O primeiro juiz do experimento anterior (o mais severo dos juízes) avaliou a corretude/plausibilidade de todas as estruturas argumentais com probabilidade maior do que um *threshold* de 10^{-6} e que tiveram suporte de pelo menos 3 sentenças do corpus de treinamento. A Tabela 2 sintetiza os resultados para os 20 verbos selecionados em termos de precisão e cobertura. Precisão indica quantas das estruturas aprendidas são julgadas corretas pelo humano (da mesma forma que no experimento anterior), enquanto cobertura indica quantas das estruturas previstas pelo *PropBank* são aprendidas pelo modelo proposto. Durante o cálculo da cobertura, para cada estrutura do *PropBank*, o juiz procurou pela estrutura aprendida correspondente, a qual podia ser lexicalizada (quando havia correspondência entre as palavras das estruturas) ou não (quando havia correspondência entre as entidades das estruturas aprendidas e os tipos dos argumentos das estruturas do *PropBank*). É importante dizer que a precisão não foi calculada em relação ao *PropBank* pelo fato deste repositório não ser completo e pelo modelo proposto aprender estruturas não previstas no repositório, como é discutido a seguir. Nessas condições, o modelo atingiu precisão média de 76% e cobertura média de 86%. A segunda coluna da Tabela 2 mostra o número de sentenças de treino utilizadas para cada verbo. A terceira coluna lista o número total de estruturas avaliadas para cada verbo. Em média, foram consideradas 142 estruturas por verbo. Essa mesma avaliação foi realizada pelo mesmo juiz para as 10 (*Top-10*) e 20 (*Top-20*) estruturas mais prováveis dos 20 verbos selecionados para se verificar o número de estruturas corretas em relação ao decremento de suas probabilidades. A Tabela 3 mostra a média dos resultados para cada verbo. Como esperado, pode-se notar que, ao se considerar mais estruturas, a precisão diminui e a cobertura aumenta.

Foram detectadas duas principais razões para o aprendizado de estruturas inadequadas pelo modelo proposto: o tratamento impróprio dos advérbios e dos *phrasal verbs* (isto é, a combinação de um verbo e uma partícula que, juntos, formam uma unidade semântica). Em princípio, advérbios deveriam ser adjuntos e, portanto, não deveriam ser incluídos nas estruturas argumentais. Entretanto, em alguns casos, os advérbios são muito freqüentes e parecem essenciais para o significado das sentenças, como em *He asked rhetorically* e *He asked incredulously*. Corroborando esse fato, o *PropBank* inclui advérbios em algumas estruturas argumentais. Os *phrasal verbs*, por sua vez, não são detectados pelo modelo proposto, dadas as

limitações deste. Por exemplo, para a sentença *He gave up*, aprende-se que *up* é um possível argumento de *gave* ou que *gave* requer um argumento (*He*, neste caso). Ambos os casos são considerados inadequados.

Como exemplo de estruturas aprendidas, a Figura 6 mostra as 10 estruturas mais prováveis para o verbo *buy* e a probabilidade de cada uma. É interessante notar: as estruturas 5 e 6 são muito similares (na primeira, uma pessoa compra uma organização; na segunda, uma organização é comprada por uma pessoa); a estrutura 7 tem um item lexicalizado (*house*); na estrutura 8, há um erro causado pela inclusão do advérbio (*anyway*) na estrutura; na estrutura 9, há um erro causado pelo *phrasal verb buy down*.

Tabela 2 – Desempenho do modelo para uma amostra de 20 verbos

Verbo	**Num. de sentenças para treino**	**Num. de estruturas argumentais julgadas**	**Precisão & Cobertura (%)**
ask	3.179	212	P=83, C=100
begin	4.042	166	P=81, C=50
believe	910	45	P=87, C=100
buy	1.106	75	P=79, C=80
cause	957	23	P=74, C=100
change	2.648	152	P=88, C=100
die	4.154	257	P=72, C=100
earn	952	72	P=83, C=100
expect	6.039	422	P=73, C=75
find	4.679	210	P=82, C=100
give	3.581	249	P=67, C=100
help	1.663	53	P=75, C=40
kill	3.253	240	P=54, C=100
like	968	74	P=69, C=100
move	2.118	162	P=71, C=60
offer	1.599	59	P=85, C=67
pay	2.076	142	P=76, C=88
raise	1.268	79	P=71, C=50
sell	1.538	98	P=81, C=100
spend	1.322	62	P=61, C=100
Média	**2.402**	**142**	**P=76, C=86**

Tabela 3 – Desempenho do modelo para as 10 e 20 estruturas mais prováveis

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

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

Figura 6 – Estruturas argumentais para o verbo *buy*

Para muitos verbos, o modelo aprendeu sentidos e comportamentos não previstos pelo *PropBank*. Por exemplo, para o verbo *raise*, foram aprendidas estruturas para seu sentido de "crescer" (como em *Peter was raised in a big city*). Para o verbo *die*, os vários comportamentos seguintes foram aprendidos:

(a) *In date, person died.*

(b) *Person died in date.*

(c) *Person died in date in location.*
(d) *Person died in location in date.*

6. Conclusões

Os experimentos relatados neste artigo fazem explícitas as vantagens e limitações do modelo proposto. Pelo lado positivo, o modelo é capaz de aprender estruturas argumentais com grande precisão, sem esforço de anotação, usando ferramentas relativamente simples. Aprendem-se tanto estruturas lexicalizadas quanto generalizadas. Além de aprender a maioria das estruturas contidas no *PropBank*, aprendem-se também estruturas não previstas nesse repositório. O lado negativo é que o modelo ainda não é suficientemente robusto para lidar apropriadamente com *phrasal verbs*, advérbios e complementos verbais sentenciais complexos. Como a *FrameNet* e a *VerbNet*, também não se diferencia argumentos de adjuntos nas estruturas aprendidas. Cada uma das limitações acima será objeto de pesquisa futura. Além disso, uma avaliação mais robusta, com mais verbos, deverá ser realizada.

O repositório de estruturas argumentais aprendidas para os 1.500 verbos mais freqüentes do inglês, chamado *ArgBank*, já se encontra disponível para uso pela comunidade de pesquisa (em http://www.nilc.icmc.usp.br/~thiago/ArgBank/index.html). Como próximo passo deste trabalho, um repositório semelhante deve ser produzido para o português brasileiro, utilizando-se o Corpus NILC (Pinheiro e Aluísio, 2003).

Referências

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.

Brown, P.; Cocke, J.; Della Pietra, S.; Della Pietra, V.; Jelinek, F.; Lafferty, J.; Mercer, R.; Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, Vol. 16, N. 2, pp. 79-85.

Brown, P.; Della Pietra, S.; Della Pietra, V.; Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, N. 2, pp. 263-311.

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

Daumé, H. and Marcu, D. (2004). A Phrase-Based HMM Approach to Document/Abstract Alignment. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Fellbaum, C. (1998). *Wordnet: an electronic lexical database*. The MIT Press. Cambridge.

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.

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.

Koehn, P.; Och, F.J.; Marcu, D. (2003). Statistical Phrase-Based Translation. In the *Proceedings of the Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics*.

Korhonen, A. (2002). Semantically Motivated Subcategorization Acquisition. In the *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon*, pp. 51-58.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In the *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 394-404.

Levin, B. (1993). *Towards a lexical organization of English verbs*. Chicago University Press, Chicago.

Manning, C. (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.

Marcu, D. and Popescu, A.M. (2005). Towards Developing Probabilistic Generative Models for Reasoning with Natural Language Representations. In the *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 88-99.

Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, Vol. 19, N. 2, pp. 313-330.

Marcus, M. (1994). The Penn Treebank: A revised corpus design for extracting predicate-argument structure. In the *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ.

Merlo, P. and Stevenson, S. (2001). Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, Vol. 27, N. 3.

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.

Pardo, T.A.S. e Nunes, M.G.V. (2005). *Investigação e Desenvolvimento de Modelos Estatísticos para Análise Discursiva Automática*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 251.

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.

Ratnaparki, A. (1996). A Maximum Entropy Part-Of-Speech Tagger. In the Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania.

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*.

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.

Sarkar, A. and Zeman, D. (2000). Automatic extraction of subcategorization frames for Czech. In the *Proceedings of the 18th International Conference on Computational Linguistics*.

Sarkar, A. and Tripasai, W. (2002). Learning Verb Argument Structures from Minimally Annotated Corpora. In the *Proceedings of the 19th International Conference on Computational Linguistics*.

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

Soricut, R. and Brill, E. (2004). Automatic Question Answering: Beyond the Factoid. In the *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*.