UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-002/2017

Anotação semiautomática de expressões de referência

Danillo da Silva Rocha, Alex Gwo Jen Lan, Ivandré Paraboni[1]

Junho - 2017

Série de Relatórios Técnicos

PPgSI-EACH-USP
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo
03828-000 – São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

Anotação semiautomática de expressões de referência

Danillo da Silva Rocha, Alex Gwo Jen Lan, Ivandré Paraboni

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brazil

rochadan60@gmail.com, alex.lan@usp.br, ivandre@usp.br

Resumo. *Estudos em Geração de Expressões de Referência (GER) frequentemente realizam a coleta de córpus de descrições definidas produzidas por sujeitos humanos em experimentos controlados. Recursos deste tipo, que são essenciais para o estudo de fenômenos de referência e outras questões de pesquisa da área, acarretam um considerável esforço de anotação de grandes massas de dados. Como forma de simplificar esta tarefa, o presente trabalho apresenta um método para anotação semiautomática destas descrições baseado em regras heurísticas. O método é avaliado com base em córpus publicamente disponíveis, e suas vantagens são discutidas.*

1. Introdução

Na pesquisa em Geração de Língua Natural (GLN), é frequente a necessidade de coleta de formas linguísticas a partir de experimentos controlados envolvendo sujeitos humanos. Esta necessidade se explica pelo interesse em conhecer não apenas o texto produzido pelos participantes do experimento, mas também de exercer controle sobre o estímulo (ou contexto) que o motivou.

De especial interesse para o presente trabalho, estudos da subárea de Geração de Expressões de Referência (GER), que trata do problema computacional de determinar o conteúdo semântico de expressões deste tipo [Krahmer e van Deemter 2012], frequentemente fazem uso de experimentos psicolinguísticos para coleta de córpus de descrições definidas. Em experimentos deste tipo, estímulos de interesse (e.g., imagens) são exibidas aos participantes que têm a tarefa de descrever um determinado objeto de interesse de forma única. Exemplos de córpus desenvolvidos para a pesquisa em GER incluem o córpus TUNA [Gatt et al. 2007], GRE3D3/7 [Dale e Viethen 2009, Viethen e Dale 2011], Craft [Mitchell et al. 2010], GenX [FitzGerald et al. 2013], ReferItGame [Kazemzadeh et al. 2014], Wally [Clarke et al. 2013], Stars [Teixeira et al. 2014], Stars2 [Paraboni et al. 2017a], b5-ref [Paraboni et al. 2017c, dos Santos et al. 2017] e outros.

Um exemplo de estímulo utilizado na construção de recursos deste tipo é ilustrado pela figura 1, do córpus TUNA [Gatt et al. 2007], no qual a tarefa do participante humano é descrever (ou referenciar) o objeto em destaque (no exemplo, uma cadeira) utilizando expressões como ‘a cadeira verde’, ‘a cadeira no canto inferior esquerdo’ etc.

Córpus para GER são essenciais para o estudo de fenômenos de referência e outras questões de pesquisa da área, mas possuem um elevado custo de desenvolvimento. Em especial, o estudo de questões de seleção de conteúdo em GER - que é um fenômeno de natureza altamente semântica - depende da anotação de grandes conjuntos de dados coletados, e acarreta assim um esforço considerável a cada nova investigação. Como forma de simplificar esta tarefa, o presente trabalho apresenta um método para anotação semiautomática de descrições definidas coletadas a partir de experimentos envolvendo

Figura 1. Um contexto do córpus TUNA-Furniture, adaptado de Gatt et al. [2007].

sujeitos humanos. O método é baseado em regras heurísticas simples, e foi avaliado com base em quatro córpus de GER publicamente disponíveis para o idioma inglês. O objetivo de longo prazo desta iniciativa é o de incorporar este método a um ambiente online de implementação de experimentos de GER capaz de validar as descrições produzidas pelos participantes, e fornecer uma anotação semântica preliminar (e sujeita à validação manual posterior) dos dados coletados.

2. Conceitos básicos

Córpus de descrições definidas são recursos necessários para o estudo de estratégias humanas de referência, e para o desenvolvimento de algoritmos de GER. De modo resumido, a tarefa computacional de GER consiste em, dado um contexto de comunicação que inclui o objeto-alvo a ser descrito, decidir quais informações (ou propriedades semânticas) devem ser selecionadas para compor uma expressão deste tipo [Krahmer e van Deemter 2012]. As propriedades semânticas manipuladas por um algoritmo de GER são tipicamente representadas na forma de pares atributo-valor, como *gender-male*.

A saída do algoritmo de GER é assim um conjunto de propriedades semânticas capaz de distinguir o alvo de todos os distraidores no contexto. Por exemplo, em um contexto C envolvendo um grupo de pessoas e suas propriedades, no qual há uma pessoa em especial (o objeto-alvo t) que se deseja distinguir das demais (os distraidores), um algoritmo de GER pode computar uma lista de propriedades de t como $L = \{$gender-male, shirtcolour=black$\}$, que poderia ser posteriormente realizada - como uma tarefa à parte, e que não faz parte do problema de seleção de conteúdo - na forma superficial 'o homem de preto' [Pereira e Paraboni 2007, 2008, de Novais e Paraboni 2012].

Cabe destacar, entretanto, que a tarefa de GER não se limita há produzir descrições 'corretas' ou livres de ambiguidade. Da mesma forma que na área de GLN como um todo, grande ênfase é dada à questão da plausibilidade (do ponto de vista psicológico) da expressão gerada. Assim, um algoritmo de GER deve idealmente produzir não apenas descrições que sejam identificáveis, mas que sejam tão próximas quanto possível das descrições que um locutor humano produziria em circunstâncias idênticas (i.e., dado o mesmo contexto). Isso inclui, por exemplo, decisões sobre como combinar propriedades atômicas e relacionais [Dale e Viethen 2009, Barclay 2010, dos Santos Silva e Paraboni 2015], como controlar o grau de superespecificação da expressão gerada [Arts et al. 2011, Engelhardt et al. 2006, 2011, Paraboni et al. 2006, 2017b], a questão da variação humana

na produção destas expressões [Gupta e Stent 2005, Fabbrizio et al. 2008, Bohnet 2007, 2008, Viethen e Dale 2010, Ferreira e Paraboni 2014b] e muitas outras. Uma visão detalhada do problema computacional de GER e seus principais desafios são discutidos em van Deemter [2016].

Sendo entretanto um problema de natureza altamente semântica, o desenvolvimento de modelos computacionais de GER normalmente exigem a disponibilização de dados linguísticos acompanhados de anotação semântica, constituindo o que costuma-se denominar um córpus para GER. A partir de um córpus deste tipo é possível, por exemplo, produzir modelos baseados em métodos de aprendizagem de máquina [Viethen et al. 2013, Ferreira e Paraboni 2014a, 2017] ou soluções puramente procedurais [Dale 2002, Dale e Reiter 1995, Paraboni 2000, de Lucena et al. 2010].

À primeira vista, um córpus para GER parece guardar forte semelhança com córpus textuais comumente utilizados em diversas linhas de pesquisa da grande área de Processamento de Línguas Naturas (PLN). Córpus para GER são no entanto recursos linguístico-computacionais altamente especializados, e a analogia com córpus textuais de uso geral é sob vários aspectos limitada. Ao contrário de tarefas de interpretação de língua natural que exploram a ampla gama de córpus textuais existentes (ou facilmente obtidos a partir da WEB, por exemplo), poucas tarefas de GLN podem de fato ser implementadas com base apenas no texto resultante do processo de produção da língua. Para muitos estudos de GLN - incluindo o caso da tarefa de GER - faz-se necessário conhecer não apenas os tipos de texto que se pretende gerar, mas *as condições iniciais* ou *contextos* que os motivaram (e.g., o discurso subjacente, o contexto visual etc.).

A construção de córpus para GER é assim normalmente implementada na forma de experimentos controlados envolvendo participantes humanos, estabelecendo ligações entre o texto produzido e os fatores (ou contexto) que o motivaram. Experimentos deste tipo fornecem estímulo textual ou visual aos participantes, e registram suas reações manifestas na forma oral, escrita, ou por ações de interação com o ambiente de experimentação. Após a coleta, processamento e anotação dos dados, o resultado final é assim formado pela representação semântica dos contextos de referência e das descrições neles produzidas.

Um exemplo proeminente de córpus para GER é o resultado do projeto TUNA [Gatt et al. 2007], desenvolvido especificamente para o estudo de fenômenos de referência e algoritmos deste tipo. O córpus TUNA contempla situações de referência em dois domínios distintos: peças de Mobília (*Furniture*), e fotos de Pessoas (*People*). A figura 1 anterior apresenta um exemplo de contexto empregado para coleta de descrições definidas no domínios *Furniture*. Um exemplo do domínio *People* é ilustrado pela figura 2. Em ambos os casos, a tarefa dos participantes do experimento era descrever de forma única (i.e., livre de ambiguidade) o objeto-alvo em destaque. Nestes exemplos, isso poderia levar à produção de descrições definidas como, e.g., 'a cadeira verde' e 'o homem de barba branca, olhando para o lado', respectivamente.

As descrições TUNA foram geradas a partir de experimentos controlados realizados com 60 participantes que usavam o Inglês como idioma nativo, ou possuíam fluência no mesmo. no total, o córpus TUNA contém 2280 expressões (780 singulares e 1500 plurais) e seus respectivos contextos. Assim como em outros recursos deste tipo, tanto as descrições como seus contextos são acompanhados de anotação semântica. Sendo

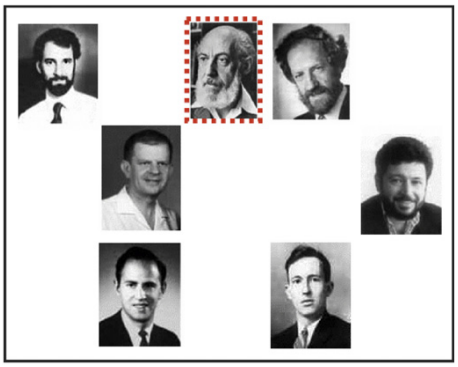

Figura 2. Um contexto do córpus TUNA-People, adaptado de Gatt et al. [2007].

o primeiro recurso de maior escala deste tipo a ser disponibilizado publicamente para pesquisa, o córpus TUNA foi utilizado para treinamento e teste de algoritmos de GER em uma grande variedade de projetos, incluindo três competições (ou *shared tasks*) da área [Gatt e Belz 2007, Gatt et al. 2008, 2009].

Outro recurso amplamente utilizado na área é o córpus GRE3D3 [Dale e Viethen 2009] e sua extensão GRE3D7 [Viethen e Dale 2011], em ambos os casos tratando da questão do uso de relações espaciais em contextos visuais tridimensionais simplificados. Nestes dois experimentos, participantes foram instruídos a descrever objetos geométricos do tipo esfera e cubo, produzindo descrições como 'a bola ao lado do cubo vermelho' etc. Um exemplo deste tipo de domínio é ilustrado na figura 3. Novamente, a tarefa do participante recrutado para a construção do córpus era descrever de forma única o objeto-alvo em destaque (neste caso apontado por uma seta). No exemplo, isso poderia levar à produção de descrições como 'a bolinha verde', 'a esfera verde, mais para a esquerda' e muitas outras.

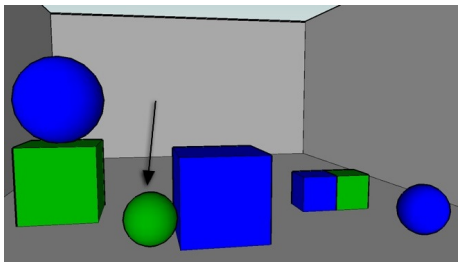

Figura 3. Um contexto do córpus *GRE3D7*, adaptado de Viethen e Dale [2011].

Juntos, os córpus GRE3D3 e GRE3D7 somam cerca de 5110 descrições produzidas por 350 participantes, e são possivelmente o maior conjunto de dados do gênero disponível para fins de pesquisa, e também um dos poucos a contemplar o fenômeno de referência espacial.

2.1. Anotação de córpus para GER

A anotação de um córpus de GER implica, em um primeiro momento, a tarefa de rotular os objetos de cada contexto (e.g., imagem) com propriedades semânticas na forma atributo-valor. Para este fim, é utilizado um esquema pré-definido de acordo com os objetivos do estudo que motivou a coleta de dados. Segue-se então a anotação das descrições

obtidas segundo o mesmo esquema empregado na anotação das imagens. Um exemplo de esquema de anotação utilizado nos córpus GRE3D37 é ilustrado na tabela 1.

Tabela 1. esquema de anotação do córpus GRE3D7 [Viethen e Dale 2011]

atributo	valores possíveis
type	{cube, ball}
colour	{green, red, blue, yellow}
size	{large, small}
near(x), above(x), right(x), left(x), below(x)	identificador do objeto x

Ambos os conjuntos de anotações - de imagens e descrições - costuma seguir a notação XML originalmente adotado em Gatt et al. [2007], e que é amplamente difundido na área. Por exemplo, supondo-se que os objetos próximos à seta na cena de exemplo do córpus GRE3D7 da figura 3 anterior são identificados como $b3$ (bola) e $c4$ (cubo), a anotação semântica de uma descrição como 'green ball next to blue cube' poderia ser representada em formato XML como ilustrado pela figura 4.

```
<CONTEXT ID="13" SEQ="6">
        <ATTRIBUTE-SET TARGET="b3" LANDMARK="c4"
                STRING="green ball next to blue cube">
            <ATTRIBUTE NAME="type" VALUE="ball" />
            <ATTRIBUTE NAME="colour" VALUE="green" />
            <ATTRIBUTE NAME="near" VALUE="c4" />
            <ATTRIBUTE NAME="landmark-type" VALUE="cube" />
            <ATTRIBUTE NAME="landmark-colour" VALUE="blue" />
        </ATTRIBUTE-SET>
</CONTEXT>
```

Figura 4. Descrição GRE3D7 em notação XML.

3. Trabalho realizado

A tarefa de anotação de grandes massas de descrições definidas é um dos pontos críticos da construção de um córpus de GER, e uma das suas etapas de maior demanda de recursos humanos especializados. Uma vez que os resultados dos futuros estudos baseados nestes dados depende fundamentalmente da acurácia da anotação, não seria realista supor que este processo possa (ou deva) ser totalmente automatizado. No entanto, assim como em inúmeras tarefas de anotação de córpus típicas da pesquisa em PLN, a anotação semântica de descrições definidas pode em princípio beneficiar-se de métodos semiautomáticos para facilitar sua execução. Dentre diversas maneiras possíveis de auxiliar nesta tarefa, destacamos neste trabalho a anotação semiautomática baseada em um pequeno conjunto de exemplos anotados, e que podem prover uma anotação inicial (e possivelmente incompleta) sujeita à revisão manual posterior. Além disso, observamos que as descrições definidas do tipo normalmente coletado em experimentos psicolinguísticos como TUNA [Gatt et al. 2007] e GRE3D3/7 [Dale e Viethen 2009, Viethen e Dale 2011] possuem estrutura sintática relativamente simples, o que sugere a viabilidade de métodos mais superficiais aplicados a esta tarefa.

O método de anotação semiautomática a ser considerado neste trabalho faz uso de regras heurísticas simples para estabelecer associações entre as palavras do *string* a ser anotado e as propriedades semânticas que as representam. Para este fim, o método faz uso de uma base de conhecimento na forma de propriedades mapeadas para as palavras que as representam, e seus sinônimos mais comuns. Esta base, a ser construída de forma manual para cada domínio de interesse (e.g., a partir de um conjunto de exemplos de descrições definidas), representa o aspecto 'semiautomático' do método. Além desta base de conhecimento dependente da aplicação, o único outro recurso externo utilizado é o dicionário DELAF-Pt [Muniz 2004] para identificação de classes gramaticais.

O método proposto funciona da seguinte forma. Seja um domínio D composto de todos objetos (alvos e distraidores) do córpus a ser anotado, e suas propriedades possíveis na forma de pares atributo-valor. A definição do domínio D é tipicamente parte do projeto de construção do córpus de GER, e contempla os objetos e propriedades de interesse para o estudo ao qual o córpus se destina. Dado um *string* S representando uma descrição definida como uma lista de palavras, e um conjunto de mapeamentos M de propriedades-palavras no domínio D, objetiva-se produzir um conjunto L de propriedades anotadas que corresponda ao subconjunto de palavras de S que puderam ser identificadas nos mapeamentos em M. O algoritmo a seguir ilustra este procedimento.

```
1  Heuristic(S, M, D, lang)
2      if lang == English then
3          Reverse(S)
4      end
5      L ← ∅
6      Z ← Split(S, D)
7      for z_i ∈ Z do
8          for w_j ∈ z_i do
9              np ← NearestNoun(w_j, z_i)
10             p ← M[w_j + np]
11             if p ≠ null then
12                 L ← L ∪ p
13             end
14             else
15                 p ← M[w_j]
16                 if p ≠ null then
17                     L ← L ∪ p
18                 end
19             end
20         end
21         return L
22     end
```

Algoritmo 1: Anotação Heurística

O método inicia verificando se o idioma *lang* da descrição é o Inglês (linha 2). Em caso afirmativo, o *string* de entrada S é invertido (linha 3) de modo que suas palavras sejam posteriormente examinadas considerando-se primeiramente o núcleo (que é tipicamente um substantivo) como seria o caso em Português. Assim, descrições como 'dark guy' são transformadas em 'guy dark' e tratadas da mesma forma que 'rapaz moreno'. A seguir, é criado um conjunto vazio de propriedades L a ser retornado como resultado (linha 5), e a função auxiliar *Split* (não detalhada) é invocada para dividir o *string* de entrada

S em k subcomponentes $z_{1..k}$ separados por propriedades relacionais (6). O objetivo deste procedimento é o de tratar individualmente cada objeto referenciado na expressão. Por exemplo, em 'green ball *near* blue cube' a propriedade relacional 'near' separa a porção esquerda do *string*, que se refere ao objeto-alvo (bola) da porção da direita, que se refere ao objeto *landmark* (cubo).

A tarefa de decidir se uma determinada propriedade é do tipo relacional (como *near*) ou atômica (i.e., intrínseca ao objeto, como *colour*) a partir da definição do domínio D é implementada de forma trivial pela função *Split* em virtude de que os valores destas propriedades denotam outros objetos do contexto, e não características do objeto em si. Por exemplo, a propriedade *near-c4* é reconhecida como sendo relacional porque o valor $c4$ representa o identificador de um objeto em D, e não uma característica intrínseca do objeto-alvo, como seria o caso de uma propriedade atômica como *colour-green*.

Uma vez delimitado, cada *substring* z_i é tratado individualmente (linha 7) e suas palavras são consideradas em associação ao substantivo mais próximo (9-13) e, se necessário, de forma isolada (14-20). Em um primeiro momento, o substantivo mais próximo np (que é o provável núcleo do sintagma ao qual w_j é subordinada) é localizado pela função auxiliar *NearestNoun* (não detalhada no pseudocódigo) na direção do idioma inglês ou português conforme definido no início do algoritmo (9). A seguir, a combinação da palavra atual w_j e núcleo np é consultada na base de mapeamentos M para verificar se esta corresponde a alguma propriedade p (10). Em caso afirmativo (11), a propriedade p correspondente é acrescentada ao conjunto de propriedades anotadas L (12).

Apesar da simplicidade, este procedimento permite a identificação correta da maioria dos casos de dependência sem necessidade de análise sintática, e pode assim ser considerado suficiente para a aplicação em discussão. Por exemplo, é possível determinar o sentido correto de 'dark' em expressões como 'dark man' e 'man with dark beard' como sendo *haircolour-dark* ou *beardcolour-dark*, respectivamente. Cabe destacar entretanto que, nos córpus de teste considerados na avaliação deste trabalho (cf. próxima seção), este tipo de dependência é relativamente raro, e limitado quase que exclusivamente ao domínio TUNA-People. Para a grande maioria dos casos encontrados, a associação direta entre palavras individuais e propriedades é a mais comum.

Caso não haja um mapeamento da combinação $w_j + np$ para nenhuma propriedade do domínio, a palavra w_j é então consultada individualmente, ou seja, sem considerar possíveis associações a nenhum núcleo específico (15). Caso haja um mapeamento entre a palavra e uma propriedade p (16), p é acrescentada ao conjunto de propriedades anotadas L (17). O processo é repetido até que todas as palavras de todos os *substrings* tenham sido consideradas, e então o conjunto de anotações L identificadas é retornado (21). Ao longo deste processo, palavras não identificadas são desconsideradas, e portanto a anotação L resultante pode permanecer incompleta ou mesmo vazia.

4. Avaliação

Esta seção descreve a avaliação do método proposto com base em quatro córpus de descrições definidas em inglês já disponibilizados com anotação semântica. O objetivo da avaliação foi o de medir o grau de proximidade entre a anotação existente em cada córpus e aquela que seria obtida de forma semiautomática pelo uso do método Heurístico.

Métodos avaliados Para fins de avaliação do método Heurístico proposto, consideramos como sistema de *baseline* uma alternativa de anotação semiautomática implementada com uso da ferramenta *nlpnet* [Fonseca e Rosa 2013]. Esta ferramenta, que originalmente oferece resultados que representam o estado da arte em *POS-tagging* para o Português brasileiro, foi neste caso adaptada a uma aplicação menos típica, ou seja, à anotação de propriedades semânticas de cada palavra de uma descrição definida fornecida.

O uso da ferramenta *nlpnet* com este propósito será aqui denominado o método de *baseline* POS. A ferramenta é inicialmente treinada a partir de um conjunto de exemplos rotulados manualmente (e que representa o aspecto 'semiautomático' do método) e, em uma segunda etapa, o modelo resultante é aplicado a um conjunto de teste a ser anotado.

Um exemplo de preparação de dados de treinamento da descrição 'green ball near blue cube' é ilustrado a seguir. Nesta representação, palavras que não correspondem a nenhuma propriedade do domínio D (incluindo o uso de artigos, preposições etc.) são anotadas como OOV (*out of vocabulary*).

green_TARGET_COLOUR
ball_TARGET_TYPE
near_NEAR
blue_LANDMARK_COLOUR
cube_LANDMARK_TYPE

Do ponto de vista lógico, ambos os métodos são equivalentes, já que utilizam o mesmo conhecimento de entrada, e diferenciam-se apenas pela forma de representação deste conhecimento, e pela forma como cada método é utilizado. O método Heurístico tende entretanto a ser mais conveniente para uso em tempo de execução, como no caso de desejar-se a anotação (e eventual tratamento de inconsistências etc.) da descrição já durante o experimento de coleta de dados. O método POS, por outro lado, é mais adequado ao processamento em lote de conjuntos de descrições previamente coletadas.

Conjuntos de dados Para avaliação dos métodos de anotação Heurístico e POS, consideramos a anotação semântica disponibilizada nos quatro conjuntos de descrições definidas de objetos em cenas visuais já discutidos, e amplamente difundidos na pesquisa em GLN: os córpus TUNA-Furniture e TUNA-People [Gatt et al. 2007], e os córpus GRE3D3 [Dale e Viethen 2009] e GRE3D7 [Viethen e Dale 2011]. Para cada um dos quatro córpus, uma pequena porção de dados (de 14 a 18%) foi utilizada para treinamento de cada modelo, e o restante foi reservado para teste. A tabela 2 apresenta o número de instâncias (i.e., descrições) de treinamento e teste, e exemplos linguísticos observados em cada domínio.

Tabela 2. Córpus de teste.

Domínio	Treinamento	Teste	Exemplo
TUNA-Furniture	63	288	the large red couch
TUNA-People	54	303	the man with gray beard and glasses
GRE3D3	90	540	the small green cube
GRE3D7	624	3856	the red ball next to a large cube

Procedimento No caso do método Heurístico, o conjunto de treinamento foi utilizado para extração de mapeamentos entre palavras e suas propriedades, conforme discutido na

seção anterior. No caso do método POS, o conjunto de treinamento foi rotulado manualmente com etiquetas representando as propriedades de interesse e utilizado no treinamento de um etiquetador conforme discutido. A avaliação propriamente dita consistiu em aplicar os dois modelos previamente treinados ao conjunto de teste de cada domínio, comparando-se seus resultados com a anotação de referência disponibilizada em cada córpus por meio de coeficiente Dice [Dice 1945] e Acurácia. Coeficientes Dice variam de 0 até 1, sendo que o valor 1 representa coincidência total entre a anotação semiautomática e a do córpus. A Acurácia representa a proporção de casos em que as duas anotações eram idênticas.

Resultados A tabela 3 apresenta os valores médio de coeficientes Dice e Acurácia obtidos pelos dois métodos avaliados para cada um dos córpus de teste. As diferenças estatísticas significativas entre os dois métodos são destacadas.

Tabela 3. Resultados

	Heurístico		POS	
Córpus de teste	Dice	Acc.	Dice	Acc.
TUNA-Furniture	**0,83**	**0,47**	0,63	0,09
TUNA-People	0,38	0,01	**0,50**	**0,12**
GRE3D3	**0,99**	**0,96**	0,76	0,74
GRE3D7	**0,97**	0,86	0,95	0,92

O método Heurístico apresenta resultados de modo geral superiores ao *baseline* baseado em POS-tagging com exceção do domínio TUNA-People. Observa-se resultados altamente positivos - e próximos de 100% nos domínios GRE3D3 e GRE3D7, e resultados mais modestos no caso dos domínios TUNA-Furniture e, de modo especial, TUNA -People. Os domínios TUNA são assim mais indicativos da complexidade da tarefa, diferença esta que pode ser em grande parte explicada pela maior variedade lexical (especialmente no caso de TUNA-People), a qual não é suficientemente representada pelo pequeno conjunto de treinamento utilizado na avaliação.

A comparação entre coeficientes Dice médios dos dois métodos foi realizada com o uso do teste de Wilcoxon. O método Heurístico apresentou acurácia média significativamente superior à acurácia média do método POS nos domínios TUNA-Furniture (W=23027, Z=1,85, p <0.001), GRE3D3 (W=10072, Z=10.15, p <0.001) e GRE3D7 (W=17180, Z=3.03, p =0.0024). No domínio TUNA-People, um efeito contrário foi observado (W=-14630, Z=-6.13, p <0.001).

Finalmente, a comparação entre valores de acurácia média dos dois métodos foi realizada com o uso do teste de Qui-quadrado. O método Heurístico apresentou acurácia média significativamente superior à acurácia média do método POS nos domínios TUNA-Furniture (χ^2 = 35.81, df =1, p<0.01), e GRE3D3 (χ^2 = 18.98, df =1, p=0.00013). No caso do domínio TUNA-People, um efeito contrário foi observado (χ^2 = 9.95, df =1, p=0.001604), e no caso do domínio GRE3D7, a acurácia média dos dois métodos não apresenta diferença significativa.

5. Considerações finais

Este trabalho apresentou um método de anotação semiautomática de descrições definidas baseado em heurísticas de associação de palavras às propriedades semânticas que elas representam. O método objetiva simplificar a tarefa de anotação de grandes massas de dados que são a saída típica de experimentos de coleta de dados para a pesquisa em GER, e pode ser especialmente adequado para processamento de descrições individuais durante a execução do experimento de coleta de dados (por exemplo, para fins de validação em tempo real), já que este método não depende de treinamento prévio.

O método apresentado evidentemente não faz uso de recursos ou técnicas avançadas de PLN, mas é uma ferramenta potencialmente útil para a pesquisa em GER/GLN, e pode em princípio ser adaptado para outras tarefas de anotação semântica desta natureza. Como trabalho, futuro, espera-se aprimorar o método proposto e incorporá-lo a um ambiente online para realização de experimentos de GER no qual participantes do experimento de coleta de dados receberiam *feedback* a respeito das respostas fornecidas. Isto permitiria corrigir erros comuns (e.g., como os causados pela falta de atenção por parte do participante etc.) e, consequentemente, coletar dados de maior qualidade e já acompanhados de uma anotação inicial a ser revisada posteriormente.

Agradecimentos

Este trabalho conta com apoio FAPESP nro. 2016/14223-0.

Referências

Arts, A., Maes, A., Noordman, L. G. M., e Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43:361–374.

Barclay, M. (2010). *Reference Object Choice in Spatial Language: Machine and Human Models*. Phd thesis, University of Exeter.

Bohnet, B. (2007). IS-FBN, IS-FBS, IS-IAC: The adaptation of two classic algorithms for the generation of referring expressions in order to produce expressions like humans do. Em *UCNLG+MT: Language Generation and Machine Translation*, páginas 84–86, Copenhagen, Denmark.

Bohnet, B. (2008). The fingerprint of human referring expressions and their surface realization with graph transducers. Em *5th International Natural Language Generation Conference*, páginas 207–210, Salt Fork, Ohio, USA.

Clarke, A. D. F., Elsner, M., e Rohde, H. (2013). Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4(329).

Dale, R. (2002). Cooking up referring expressions. Em *Proceedings of ACL-2002*, páginas 68–75, Philadelphia, PA, USA.

Dale, R. e Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19.

Dale, R. e Viethen, J. (2009). Referring expression generation through attribute-based heuristics. Em *12th European Workshop on Natural Language Generation*, ENLG '09, páginas 58–65, Athens, Greece.

de Lucena, D. J., Paraboni, I., e Pereira, D. B. (2010). From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

de Novais, E. M. e Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

dos Santos, V. G., Paraboni, I., e Silva, B. B. C. (2017). Big five personality recognition from multiple text genres. Em *Text, Speech and Dialogue (TSD-2017) (to appear)*, Prague, Czech Republic. Springer-Verlag.

dos Santos Silva, D. e Paraboni, I. (2015). Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition & Computation*, 15(03):186–225.

Engelhardt, P. E., Baileyand, K., e Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573.

Engelhardt, P. E., Demiral, S. B., e Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77(2):304–314.

Fabbrizio, G. D., Stent, A., e Bangalore, S. (2008). Trainable speaker-based referring expression generation. Em *12th Conference on Computational Natural Language Learning*, páginas 151–158, Manchester, UK.

Ferreira, T. C. e Paraboni, I. (2014a). Classification-based referring expression generation. Em *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science 8403*, páginas 481–491, Kathmandu, Nepal. Springer.

Ferreira, T. C. e Paraboni, I. (2014b). Referring expression generation: taking speakers' preferences into account. Em *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 8655*, páginas 539–546, Brno, Czech republic. Springer.

Ferreira, T. C. e Paraboni, I. (2017). Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering*, páginas 1–22.

FitzGerald, N., Artzi, Y., e Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. Em *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, páginas 1914–1925.

Fonseca, E. R. e Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. Em *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, páginas 98–107.

Gatt, A. e Belz, A. (2007). The attribute selection for GRE challenge: Overview and evaluation results. Em *UCNLG+MT: Language Generation and Machine Translation*, Copenhagen, Denmark.

Gatt, A., Belz, A., e Kow, E. (2008). The TUNA challenge 2008: Overview and evaluation results. Em *Fifth International Natural Language Generation Conference (INLG-2008)*, páginas 198–206, Salt Fork, Ohio, USA.

Gatt, A., Belz, A., e Kow, E. (2009). The TUNA challenge 2009: Overview and evaluation results. Em *12nd European Workshop on Natural Language Generation*, páginas 174–182, Athens, Greece.

Gatt, A., van der Sluis, I., e van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. Em *Proceedings of ENLG-07*, Schloss Dagstuhl, Germany.

Gupta, S. e Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. Em *1st workshop on using corpora in NLG*, Birmingham, UK.

Kazemzadeh, S., Ordonez, V., Matten, M., e Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. Em *Proceedings of EMNLP-2014*, páginas 787–798, Doha, Qatar.

Krahmer, E. e van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Mitchell, M., van Deemter, K., e Reiter, E. (2010). Natural reference to objects in a visual domain. Em *Proceedings of INLG-2010*, Dublin, Ireland.

Muniz, M. C. M. (2004). A construção de recursos linguístico-computacionais para o Português do brasil: o projeto de Unitex-PB. Master's thesis, ICMC / USP São Carlos.

Paraboni, I. (2000). An algorithm for generating document-deictic references. Em *Procs. of workshop Coherence in Generated Multimedia, associated with First Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon*, páginas 27–31.

Paraboni, I., Galindo, M., e Iacovelli, D. (2017a). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 51(2):439–462.

Paraboni, I., Lan, A. G. J., de Sant'Ana, M. M., e Coutinho, F. L. (2017b). Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43(2).

Paraboni, I., Masthoff, J., e van Deemter, K. (2006). Overspecified reference in hierarchical domains: measuring the benefits for readers. Em *Proceedings of the fourth international natural language generation conference (INLG-2006)*, páginas 55–62, Sydney, Australia.

Paraboni, I., Monteiro, D. S., e Lan, A. G. J. (2017c). Personality-dependent referring expression generation. Em *Text, Speech and Dialogue (TSD-2017) (to appear)*, Prague, Czech Republic. Springer-Verlag.

Pereira, D. B. e Paraboni, I. (2007). A language modelling tool for statistical NLP. Em *5th Workshop on Information and Human Language Technology (TIL-2007)*, páginas 1679–1688, Rio de Janeiro.

Pereira, D. B. e Paraboni, I. (2008). Statistical surface realisation of Portuguese referring expressions. Em *Gotal-2008, Lecture Notes in Artificial Intelligence 5221*, páginas 383–392, Gothenburg, Sweden. Springer-Verlag.

Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., e Yamasaki, A. K. (2014). Generating relational descriptions involving mutual disambiguation. Em *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science 8403*, páginas 492–502, Kathmandu, Nepal. Springer.

van Deemter, K. (2016). *Computational Models of Referring. A Study in Cognitive Science*. MIT Press, Cambridge, Massachusetts, USA.

Viethen, J. e Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. Em *Proceedings of the Australasian Language Technology Association Workshop 2010*, páginas 81–89, Melbourne, Australia.

Viethen, J. e Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. Em *UCNLG+Eval: Language Generation and Evaluation Workshop*, páginas 12–22, Edinburgh, UK.

Viethen, J., Mitchell, M., e Krahmer, E. (2013). Graphs and spatial relations in the generation of referring expressions. Em *14th European Workshop on Natural Language Generation*, páginas 72–81, Sofia, Bulgaria.