

Universidade de São Paulo
Escola de Artes, Ciências e Humanidades - EACH

**Resolução e Interpretação de Expressões de Referência
na Geração Automática de Textos**

Ivandr  Paraboni

Projeto CNPq nr. 484015/2007 9
Edital MCT/CNPq 15/2007 - Universal

Resumo

A geração de expressões de referência (GER) é um dos componentes fundamentais de aplicações de geração de linguagem natural (GLN) a partir de dados de entrada não-lingüísticos. Trabalhos existentes nesta área tendem a se concentrar em situações de referência simplificadas, na qual questões como o grau de facilidade de resolução (ou busca) da referência não é relevante. Em trabalho prévio investigamos a geração de expressões de referência em contextos *físicos* estruturalmente complexos, demonstrando que, ao contrário de abordagens tradicionais para o problema, o uso de redundância lógica para facilitar a tarefa de resolução pode ser essencial para o sucesso da referência gerada. Embora tal estudo tenha aplicação imediata em áreas como a geração de diálogos homem-máquina e em ambientes interativos em geral, um grande número (ou talvez a maioria) das aplicações de GLN concentra-se no domínio puramente textual, ou seja, na geração de expressões de referências cujos antecedentes não são objetos de um contexto físico imediato, mas sim objetos do *discurso* no qual estão inseridas. Neste trabalho propomos uma extensão do estudo realizado buscando generalizar seus resultados para situações de referência no discurso, e assim avançar em direção a uma teoria de maior cobertura deste fenômeno.

Palavras-chave: processamento de linguagem natural, geração de linguagem, expressões de referência.

Sumário

1	Introdução	3
2	Geração de Expressões de Referência: Abordagem Tradicional	3
3	O Trabalho Realizado: Facilidade de Resolução em Domínios Físicos.....	5
4	Proposta de Pesquisa.....	9
4.1.	Objetivos.....	9
4.2.	Metodologia.....	9
4.3.	Cronograma de Execução	11
4.4.	Recursos	11
5	Considerações Finais	11
	Bibliografia	12

1 Introdução

Participantes de um discurso escrito ou falado que desejam comunicar uma mensagem normalmente procuram fazer com que sua declaração seja de fácil compreensão, o que de modo geral é confirmado por um grande número de investigações empíricas [11,18,4,1,14]. Nesta proposta discutimos conseqüências dessa observação para a *geração de expressões de referência* (GER), que é um dos componentes tradicionais de aplicações de geração de linguagem natural (GLN) a partir de dados de entrada não-lingüísticos. Para este propósito, diferenciamos dois aspectos da ‘compreensão’ de uma expressão de referência, as quais chamamos interpretação e resolução. Por *interpretação* entendemos o processo pelo qual um leitor ou ouvinte determina o significado ou forma lógica de uma expressão (lingüística) de referência; por *resolução* entendemos a identificação do referente em questão uma vez que o significado da expressão lingüística tenha sido determinado.

Dificuldade de interpretação e resolução são em geral forças opostas. Por exemplo, considere duas alternativas de referência não-ambígua a uma mesma pessoa (i.e., só há um homem vestido de preto no contexto em questão) como a seguir:

- (a) o homem de preto
- (b) o homem de preto, ao lado do piano

A expressão (a) é mais breve (e possivelmente mais fácil de interpretar) do que (b), porém a informação adicional em (b) facilita a resolução (i.e., a busca e identificação da pessoa referenciada) uma vez que a interpretação tenha sido concluída.

A informação adicional presente em (b) (i.e., a referência ao piano) é dita *logicamente redundante* no sentido de que não colabora para a desambiguação do referente. Este conceito é explorado no trabalho descrito em [16], que discute sob quais circunstâncias um algoritmo de GER pode (ou deve) incluir material deste tipo em uma expressão de referência para facilitar sua resolução¹.

A discussão em [16] é limitada à dificuldade de interpretação em domínios *físicos*, nos quais a decisão sobre a quantidade de informação a ser incluída na expressão de referência é influenciado principalmente pelas relações *espaciais* entre os objetos do domínio. No entanto, um grande número de aplicações de GLN concentra-se no domínio puramente *textual* (e.g., aplicações de geração de texto em geral) no qual a decisão sobre a quantidade de informação a ser incluída em uma expressão de referência passa a ser influenciadas pelas relações *discursivas* entre referentes. Possíveis similaridades entre estas duas formas de organização são no mínimo vagas, e atualmente não é possível estimar até que ponto os resultados apresentados para geração de referências a objetos em domínios físicos possam ser generalizados para referência a objetos no discurso. A presente proposta busca assim estender o conhecimento existente sobre este fenômeno, ampliando o escopo do trabalho em [16] para um grande número (ou possivelmente a maioria) das aplicações atuais de GLN, e contribuindo para a elaboração de uma teoria de maior cobertura sobre o necessário equilíbrio entre facilidade de resolução e interpretação na geração destas expressões.

2 Geração de Expressões de Referência: Abordagem Tradicional

A *geração de expressões de referência* (GER) é uma das tarefas mais tradicionais em aplicações de geração de linguagem natural (GLN). Algoritmos para este fim tratam da construção das expressões lingüísticas que fazem referência aos objetos do discurso que está sendo gerado, e podem ser divididos em dois tipos básicos: de um lado, observamos algoritmos que determinam *o tipo de expressão lingüística* necessária em um determinado ponto do discurso (e.g., a escolha entre um pronome pessoal como “ele”, um nome próprio como “João” ou uma descrição definida como “o vendedor de automóveis”) [12,17]. De outro lado, há algoritmos que se encarregam especificamente da construção da *descrição definida* adequada (e.g., a escolha entre “o vendedor de automóveis”, “o irmão de José”, “o homem citado na frase

¹ Alguns aspectos da dificuldade de *interpretação* de referências são discutidos em [14,15].

acima” etc.) uma vez que o uso deste tipo de expressão já tenha sido decidido [5,6,8,10,13,14,15]. Neste artigo discutimos o segundo tipo de problema.

Um dos mais bem-conhecidos e influentes algoritmo de GER para produção de descrições definidas é o *algoritmo Incremental* de R. Dale e E. Reiter [6]. Este algoritmo recebe como entrada o referente r que se deseja descrever, um contexto formado pelos objetos que são passíveis de referência em um dado ponto do discurso, e suas propriedades semânticas referenciáveis (i.e., aquelas que podem figurar em uma descrição definida). O objetivo do algoritmo é computar uma lista L de propriedades semânticas de r tal que L permita a identificação de r sem ambigüidade, i.e., tal que L possa ser realizada na forma de uma expressão lingüística que descreva apenas o referente (e nenhum outro objeto do contexto)². Por exemplo, considere o contexto da Fig. 1, contendo cinco objetos geométricos que podem ser referenciados de forma única utilizando-se propriedades como cor, tamanho e posição. Neste caso o algoritmo Incremental poderia produzir descrições como “o cilindro preto” e “o cubo branco da direita”, em referência ao terceiro e quarto objetos, respectivamente.

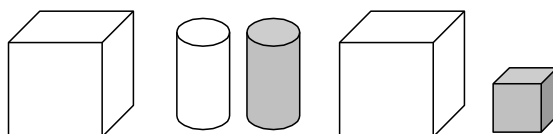


Fig. 1. Um domínio-exemplo contendo objetos geométricos referenciáveis via propriedades visuais como forma (cubo ou cilindro), cor (preta ou branca) e posição (esquerda ou direita).

No algoritmo Incremental as propriedades do referente vão sendo incluídas em L de forma seqüencial segundo uma ordem de preferência determinada pelo domínio, até que uma descrição única seja obtida. Por razões de complexidade computacional (a obtenção de um conjunto mínimo de propriedades identificadoras é um problema NP-difícil cf. [6]), uma propriedade, uma vez selecionada, não pode mais ser removida da lista L , mesmo que o acréscimo posterior de outra propriedade tenha a tornado redundante (daí o nome do algoritmo).

Tradicionalmente, o algoritmo Incremental [6] e abordagens relacionadas [5,10] procuram evitar a inclusão de informações que não sejam estritamente necessárias para a desambiguação do referente, ou seja, consideram apenas a inclusão de propriedades *restritivas*. Assim, as expressões resultantes são ditas *mínimas*³. Esta estratégia se explica pela necessidade de adesão à máxima de *brevidade* proposta por H. P. Grice [7], com o propósito de evitar implicações lógicas que possam frustrar o significado comunicado. Por exemplo, considere as duas expressões a seguir para referência a um mesmo objeto (i.e., uma porta), em um contexto livre de ambigüidade (i.e., só existe uma porta na sala em que se encontram os participantes do discurso).

- (c) Por favor, abra *a porta* para mim.
- (d) Por favor, abra *a porta vermelha* para mim.

Neste caso a informação adicional em (d) pode constituir um obstáculo à resolução, uma vez que a referência (supérflua, do ponto de vista lógico) à cor do referente pode deixar o ouvinte em dúvida sobre o porquê desta observação e, conseqüentemente, sobre o sentido real que se pretende comunicar. Para evitar problemas deste tipo, e também para facilitar a interpretação (lingüística) das expressões geradas, é geralmente aceito que um algoritmo de GER para fins de identificação deva evitar o uso espúrio de propriedades logicamente redundantes. Desta forma, a inclusão de redundância em uma expressão de referência ficaria limitada a usos outros que não identificação, tais como para a expressão de *intenções* do comunicador [9]. Por exemplo, a expressão (d) acima poderia estar sendo usada para fazer alusão a um fato especial relacionado à cor vermelha (e.g., um determinado time de futebol), ou virtualmente a qualquer outro sentido pretendido.

² Neste trabalho enfocamos apenas a determinação do conteúdo semântico destas expressões, deixando de lado questões relacionadas à sua realização textual.

³ O termo ‘mínimo’ aqui é tomado no sentido amplo de ‘número reduzido de propriedades’, mas não necessariamente no sentido de ‘menor número possível’.

3 O Trabalho Realizado: Facilidade de Resolução em Domínios Físicos

Abordagens de GER tendem a enfatizar domínios relativamente simples, cujos objetos são descritos em termos de suas propriedades atômicas (e.g., tipo, cor, tamanho etc.). Nos poucos casos em que propriedades relacionais são consideradas (e.g., [5,10]), os exemplos discutidos são tão pequenos que é razoável assumir que os participantes do discurso conhecem antecipadamente todos os fatos sobre o domínio. Por esta razão, a dificuldade de resolução (i.e., a busca pelo referente) não se torna uma questão relevante. Esta simplificação permita o estudo de vários aspectos do fenômeno de referência, mas outros aspectos só podem ser investigados se considerarmos domínios mais extensos e/ou estruturalmente complexos.

Algumas questões interessantes aparecem, por exemplo, quando consideramos um domínio estruturado de forma *hierárquica*. Para o presente propósito, um domínio é dito ‘hierárquico’ se for organizado na forma de uma árvore cujos nós representam objetos e cujas arestas representam relações referenciáveis (e.g., relações espaciais) entre estes, tal que tudo que pertence a um nó n pertença a no mínimo um dos filhos de n , enquanto que tudo que pertence a um dos filhos de n pertence também a n . Hierarquias estão entre nossas maneiras prediletas de organizar o mundo. Exemplos de domínios deste tipo incluem países divididos em estados, províncias e cidades; anos divididos em meses, dias e minutos; prédios divididos em alas, andares e salas, livros divididos em capítulos, seções, parágrafos e sentenças etc. Muitos dos exemplos em [16] e aqui reproduzidos utilizam um domínio hierárquico simples representando o campus de uma universidade e dois de seus prédios (Watts e Cockcroft) com salas numeradas e bibliotecas como a seguir:

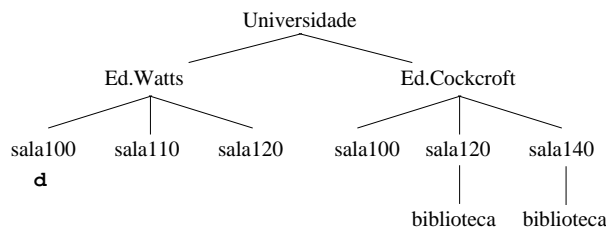


Fig. 2. Um exemplo de domínio espacial estruturalmente complexo. Os participantes do discurso encontram-se na posição d (sala 100 do edifício Watts).

A vantagem em investigar o fenômeno de referência em um domínio deste tipo se torna evidente quando consideramos o grau de conhecimento acessível ao ouvinte. Domínios hierárquicos são estruturalmente complexos - o que os torna interessantes do ponto de vista da resolução de referências - mas mantêm um número finito de relações de semântica precisa, evitando assim a complexidade inerente a um domínio espacial irrestrito. Mais especificamente, relações hierárquicas tendem a estabelecer uma ordem natural de acessibilidade para referência aos seus componentes, nos moldes propostos em [10]. Assim, é possível supor, por exemplo, que um ouvinte localizado na sala 100 do edifício Watts não conheça a estrutura do restante do campus, o que tem consequências importantes para a resolução de expressões de referência. Em um domínio extenso ou estruturalmente complexo, expressões de referência ditas mínimas (i.e., que evitam a inclusão de propriedades que não sejam estritamente necessárias para a desambiguação do referente) podem não apresentar informação suficiente para a identificação do referente. Considere novamente o exemplo da Fig. 2 acima. Suponha que os participantes do discurso se encontrem na sala 100 do edifício Watts onde foi feita a seguinte referência (mínima):

(e) A reunião será realizada na *sala 140*.

Esta é uma forma de referência única (i.e., livre de ambigüidade) pois há apenas uma sala de número 140 no campus inteiro, e é uma referência que seria de fato gerado pela maioria das abordagens de GER existentes (incluindo, por exemplo, o algoritmo Incremental). No entanto, a expressão “sala 140” não é muito útil do ponto de vista da resolução, uma vez que estando o ouvinte nas proximidades de um grupo de objetos do tipo (sala) esperado, cria uma falsa expectativa de que a sala em questão deveria se encontrar nesta região do domínio: há uma tendência (reforçada pela numeração seqüencial da numeração de salas) de procurar o referente de “sala 140” dentre as salas do edifício Watts, onde obviamente ela não pode ser encontrada.

Situações deste tipo têm sido chamadas de ‘falta de orientação’ (*Lack of Orientation*, ou LO em [14,15,16]). O ouvinte da expressão “sala 140” não tem condições de resolver a referência com segurança e, crucialmente, nem mesmo atingindo a região correta do domínio (i.e., o edifício Cockcroft) e encontrando uma sala de número 140, pode ter certeza de ter localizado o referente correto (e de não ter, por exemplo, ignorado alguma outra área onde a ‘verdadeira’ sala 140 se encontra).

Expressões mínimas podem dificultar a resolução de forma ainda mais acentuada. Suponha que desejamos fazer referência à biblioteca localizada na sala 120. Esta referência não pode ser realizada simplesmente como “a biblioteca” pois neste caso seria ambígua (já que há outra biblioteca na sala 140 do mesmo prédio). Assim, faz-se necessário o uso de uma propriedade relacional (i.e., uma referência à sala onde se encontra a biblioteca em questão) para desambiguá-la:

(f) O livro que você quer está na *biblioteca da sala 120*.

Esta expressão é uma forma de referência única à biblioteca em questão, e seria o tipo de referência produzida por algoritmos capazes de lidar com propriedades relacionais [5,10]. No entanto, estando os participantes do discurso no edifício Watts (onde existe outra sala de número 120), há o risco desta expressão ser interpretada no contexto local, levando o ouvinte a procurar por uma biblioteca na sala 120 mais próxima, onde obviamente não pode ser encontrada. Neste ponto, o ouvinte não tem condições de estimar se houve erro na produção da referência, na sua interpretação, no processo de busca ou se houve alguma alteração desconhecida no próprio domínio (e.g., uma porta trancada). Qualquer que seja a hipótese formulada, entretanto, não é possível prever se a resolução terá continuidade, ou seu grau de sucesso. Problemas desta natureza têm sido chamadas ‘becos sem saída’ (*Dead Ends*, ou DE em [14,15,16]).

Instâncias de LO e DE são problemas recorrentes na resolução de referências mínimas, e não estão limitados a domínios hierárquicos ou espaciais. Por exemplo, (g) abaixo produz uma situação de DE (beco sem saída) se o ouvinte, desconhecendo a planta do prédio em questão, estiver posicionado próximo a outro elevador que não o referenciado. Em (h), há uma situação de LO (falta de orientação) caso o diálogo (ocorrido em junho) seja interpretado como referente ao mês atual, e não o seguinte.

(g) Aguarde na *saída próxima ao elevador*.

(h) Eu volto no *dia 31* (de julho).

Formalmente, DE e LO são definidos como situações em que, na busca por um determinado referente, encontramos um obstáculo que não é o objeto referenciado na descrição, mas que possui exatamente as mesmas propriedades semânticas (DE), ou um subconjunto delas (LO). Nesta definição está implícita a noção de *escopo* das propriedades semânticas realizadas na expressão de referência. Para uma elaboração destes conceitos referenciamos [16].

Na resolução de uma expressão de referência, instâncias de DE e LO são essencialmente situações de incerteza quanto à identidade do referente. Em ambos os casos, não é possível determinar com precisão qual o procedimento de busca a ser adotado, pois embora a expressão de referência não seja ambígua, não fornece informações suficientes para a correta localização do referente. Expressões deste tipo exigem a inclusão de propriedades logicamente redundantes para facilitar a busca. Assim, o risco de incorrer em uma situação do tipo DE ou LO como nos exemplos (e) e (f) acima poderia ser eliminado com uso de uma referência adicional (i.e., logicamente redundante) ao nível hierárquico superior (i.e., o edifício Cockcroft) no qual o referente se encontra.

(i) A reunião será realizada na sala 140 *do edifício Cockcroft*.

(j) O livro que você quer está na biblioteca da sala 120, *no edifício Cockcroft*.

Nestes exemplos, a informação adicional facilita a resolução da expressão, embora aumentando o esforço necessário para a sua interpretação. Dificuldade de resolução e sua relação com dificuldade de interpretação são questões centrais para a geração de referências em domínios extensos ou estruturalmente complexos, embora ainda não sejam exploradas na literatura em GER [2,5,6,10]. O uso de redundância lógica para facilitar a resolução de referências contrasta com o uso de propriedades não-restritivas para expressar relações semânticas [3] ou intenções [9], e difere também do uso de propriedades relacionais com propósito único de desambiguação [5,10]. Apesar destas distinções,

observamos que o uso de propriedades deste tipo não satisfaz nenhum objetivo de comunicação além da mera identificação do referente.

Uma vez que facilitar a interpretação implica maior dificuldade de interpretação, coloca-se a questão de qual seria a quantidade ‘certa’ de informação a ser incluída em uma expressão de referência. De um lado, não é desejável que a expressão seja demasiadamente breve (como as expressões mínimas geradas pelos algoritmos descritos na literatura em GER), pois há o risco de DE ou LO. Por outro lado, expressões excessivamente longas são em geral indesejáveis, seja pelo risco de falsas implicações lógicas ou pelo simples tempo exigido para sua interpretação. Em [16] são propostos dois algoritmos que apresentam um diferente balanço entre dificuldade de resolução e interpretação, chamados ‘Inclusão Total’ (*Full Inclusion* ou simplesmente FI) e ‘Limitado ao Escopo’ (*Scope-Limited*, ou SL). As expressões logicamente redundantes em (i) e (j) acima são exemplos de expressões geradas pelo algoritmo FI. O algoritmo FI elimina o risco de situações do tipo DE ou LO pela inclusão de referências a *todos* os níveis hierárquicos superiores ao referente considerados informativos. O algoritmo produz descrições sucessivas do referente r e de seus ancestrais, até o ponto em que há uma tentativa de referência a um nó ancestral do local d onde a expressão foi produzida (situação que provocaria uma auto-referência do tipo “neste prédio” etc., que pode ser inferida pelo contexto e portanto é desprezada). No exemplo da Fig.2, uma referência à sala 140 seria construída com uma descrição do referente propriamente dito (“sala 140”), então acrescida de uma referência ao nó pai (“sala 140 no edifício Cockcroft”). Como o próximo nível é um ancestral de d (i.e., sua inclusão produziria uma auto-referência do tipo “sala 140 no edifício Cockcroft, *nesta universidade*”) o processo se encerra.

A inclusão total das propriedades relacionais informativas do referente elimina o risco de DE e LO (portanto facilitando a resolução da referência) mas traz conseqüências para sua interpretação: expressões geradas por FI podem ser desnecessariamente longas, uma vez que o algoritmo inclui propriedades logicamente redundantes mesmo quando não há risco de DE ou LO. Considere o mesmo exemplo de campus universitário, porém sem a segunda biblioteca (da sala 140).

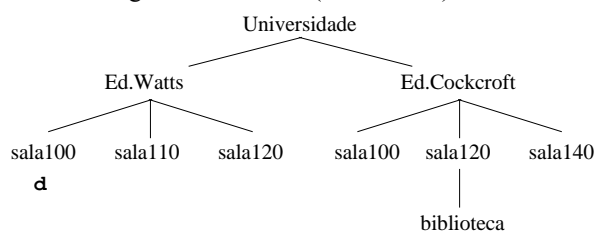


Fig. 3. Exemplo de campus universitário com apenas uma biblioteca.

Neste caso, assumindo-se a mesma localidade d onde a referência é produzida, o algoritmo FI produziria a descrição “a biblioteca na sala 120 do edifício Watts”. Entretanto, como não há risco de confusão com outras salas ou bibliotecas (i.e., não há risco de DE ou LO), a descrição mínima “a biblioteca” teria sido suficiente. Este é o princípio básico do algoritmo SL: utilizando-se da noção de escopo de propriedades do referente, o algoritmo evita a inclusão de redundância lógica que não seja estritamente necessária para a prevenção de DE e LO. Assim, SL proporciona um maior equilíbrio entre dificuldade de resolução e interpretação, favorecendo a geração de referências logicamente redundantes quando há risco de DE ou LO (exatamente como o algoritmo FI), mas favorecendo a geração de descrições mínimas quando não há este risco (exatamente como um algoritmo de GER ‘tradicional’, e.g., [5]). Para detalhes de ambos algoritmos e sua especificação formal, referenciamos [16].

Algoritmos de geração de descrições mínimas envolvendo relações como [5] e os propostos FI e SL são apenas três das muitas maneiras possíveis de produzir uma expressão de referência. Em cada abordagem são produzidas descrições com diferentes graus de dificuldade de resolução e interpretação. Mas qual a quantidade certa de informação a ser incluída? Como um primeiro passo em direção a um melhor entendimento desta questão, é descrito em [14] um experimento no qual 15 sujeitos foram instruídos a declarar suas preferências por descrições mínimas e redundantes em várias situações de referência (incluindo situações de possível DE, LO e situações neutras). Para contornar as dificuldades práticas de experimentação em um domínio físico extenso (como o campus universitário ilustrado neste trabalho), foi utilizado como exemplo de domínio hierárquico um documento estruturado em seções e subseções contendo figuras. Mais especificamente, os participantes foram instruídos a simular uma

situação de *autoria* do documento-exemplo escolhendo as expressões de referência que teriam empregado em cada situação proposta.

Os resultados detalhados em [14] demonstram que os participantes evitaram sistematicamente situações potenciais de DE e LO, escolhendo nestes casos o uso de uma expressão logicamente redundante (FI ou SL). Nos casos em que não havia risco de DE ou LO, entretanto, os resultados foram menos conclusivos: cerca de metade dos participantes do experimento demonstrou preferência por expressões logicamente redundantes, e metade optou por expressões mínimas. Esta divisão parece sugerir que não havendo dificuldade de resolução da referência, o uso de propriedades logicamente redundantes dependeria da preferência pessoal do comunicador.

Embora o experimento em [14] tenha sido bem-sucedido em demonstrar que DE e LO são preocupações necessária na geração de expressões de referência, esta abordagem não foi isenta de falhas: em primeiro lugar, observamos que os participantes do experimento foram instruídos a se *imaginar* como autores de referências pré-definidas, e não a produzir suas próprias expressões livremente. Além disso, o experimento em [14] não nos fornece uma estimativa de quão difícil realmente seria resolver uma expressão em uma situação de DE ou LO. Um experimento complementar relatado em [15,16] procurou aprofundar esta questão.

Neste segundo experimento, mediu-se o esforço de resolução (mas não de interpretação) exigido por diferentes expressões de referência com base no volume de navegação efetuada em um documento on-line. Para este fim, 20 documentos sobre assuntos variados foram apresentados em ordem aleatória a 42 participantes. O conteúdo da posição atual de cada documento (contendo uma expressão de referência solicitando ao leitor “clique” em uma determinada figura do documento) era sempre visível, mas para atingir outras porções do documento (e assim localizar cada figura solicitada) era necessário navegar utilizando *hyperlinks* representando sua divisão em seções e subseções. A seguir reproduzimos um fragmento de um destes documentos.

Document 1 out of 20: Astrology

Document structure

- Section 1
 - [Part A](#)
 - [Part B](#)
 - [Part C](#)
- Section 2
 - [Part A](#)
 - [Part B](#)
 - [Part C](#)
- Section 3
 - [Part A](#)
 - [Part B](#)
 - [Part C](#)

Section 2

Part B

Let's talk about scales.
Please click on picture 4 in part C of Section 3.

no pictures here

Fig. 4. Fragmento de um dos documentos utilizados para medir o esforço de resolução de expressões de referência: neste exemplo, o leitor se encontra na seção 2-B do documento e é instruído a localizar uma figura na seção 3-C utilizando a estrutura de *hyperlinks* da esquerda.

O experimento registrou todos os passos de navegação (i.e., “clicks” do mouse) efetuados pelos participantes em resposta a cada expressão de referência (mínima, FI ou SL) fornecida. O experimento objetivou descobrir se expressões ditas longas (i.e., logicamente redundantes) geradas por FI e SL realmente auxiliam na resolução de referências, em especial nas situações ditas problemáticas (i.e., DE e LO). Os resultados apresentados em [16] demonstram que (a) em situações de DE e LO, descrições mínimas exigem maior esforço de busca (i.e., resolução) do que descrições geradas por FI e SL e (b) o ganho (i.e., economia no esforço de resolução) obtido por uma descrição longa sobre uma descrição mínima é maior nas situações ditas problemáticas. Além disso, considerando-se que o acréscimo de informações talvez *sempre* leve a uma redução do esforço de navegação, foi demonstrado que (c) o ganho obtido por descrições mais completas do que as produzidas por FI (e.g., incluindo auto-referências como no exemplo da seção anterior “sala 140 no edifício Cockcroft, *nesta universidade*”) é relativamente pequeno se comparado ao ganho em (b).

4 Proposta de Pesquisa

Em trabalho prévio [16] investigamos o equilíbrio entre facilidade de interpretação e resolução na geração de expressões de referência em domínios físicos. Tal estudo tem aplicações imediatas na área de geração de diálogos homem-máquina e em ambientes interativos nos quais um sistema computadorizado precisa se comunicar com o usuário em um contexto (e.g., um ambiente virtual) por meio de sentenças em linguagem natural. No entanto, um grande número (ou talvez a maioria) das aplicações de GLN concentra-se no domínio puramente textual, ou seja, na geração de texto estruturado contendo expressões de referências cujos antecedentes não são objetos do contexto físico imediato, mas sim objetos do discurso.

De forma superficial, poderíamos argumentar que ambos os casos – contexto físico e discursivo – são de natureza semelhante: assim como certos tipos de domínios físicos podem ser estruturados hierarquicamente em função de graus de saliência [14], também a estrutura do discurso parece seguir uma organização similar [4,12,17]. Uma análise mais detalhada, no entanto, revela diferenças importantes, e sobre as quais até o momento podemos apenas especular: por exemplo, sabemos que o grau de dificuldade de resolução de uma expressão de referência depende da saliência relativa dos objetos no domínio. No entanto, não é evidente que saliência física (e.g., proximidade física, tamanho, cor etc.) seja de alguma forma equivalente à saliência no discurso. Pela mesma razão, é pouco provável que referência a relações espaciais entre objetos (e.g., “o livro *sobre a mesa*”) tenha o mesmo impacto na resolução da expressão do que, digamos, uma referência anafórica (e.g., “o livro *dele*”).

Atualmente não é possível afirmar com precisão até que ponto problemas de interpretação e resolução de referências característicos de contextos físicos (e os respectivos algoritmos propostos para sua solução) se aplicam à geração de referências em um contexto discursivo. A elaboração desta questão, que é uma lacuna importante no conhecimento atual em GER, representa uma oportunidade de complementação de trabalhos anteriores - e de grande interesse para aplicações de GLN que produzem texto estruturado – em direção a uma teoria de maior cobertura do fenômeno. Uma proposta de pesquisa desta natureza é detalhada nos itens a seguir.

4.1. Objetivos

O objetivo principal da presente proposta de pesquisa é investigar *se e quando* aplicações de geração de textos devam usar redundância lógica em referência a objetos do discurso. Mais especificamente, a pesquisa proposta procura estender os resultados obtidos em trabalho prévio [16] relativos à interpretação e resolução de referências em contextos *físicos* para o domínio do *discurso*, investigando a possível generalização das questões identificadas e a aplicabilidade de algoritmos propostos.

4.2. Metodologia

A presente investigação consiste da elaboração e execução de experimentos para medir o esforço de resolução⁴ de expressões de referência, cuja metodologia acompanha a série de experimentos descritos em [16]. Através do monitoramento da estratégia de resolução de referências sob condições controladas, espera-se coletar informações conclusivas sobre a forma como estas expressões devam ser geradas (e.g., de modo a serem psicologicamente plausíveis etc.).

Em linhas gerais, uma coleção de textos com múltiplas situações de referência é apresentada a um grupo de leitores através de uma interface computadorizada, solicitando que estes identifiquem o termo antecedente de certas expressões de interesse (e.g., pronomes, descrições definidas com graus variados de redundância etc.). Os leitores teriam então de realizar uma busca no texto apresentado e marcar o termo a que a expressão se refere, ao passo que uma ferramenta de software subjacente se encarregaria de coletar dados sobre o tempo gasto na resolução, o número de acertos, os tipos de erros cometidos, a forma de navegação pelo texto etc. A seguir descrevemos brevemente cada uma das atividades previstas, seu cronograma de execução e os recursos necessários.

1-Revisão Bibliográfica

Estudo dos avanços recentes em GER e do fenômeno de referência, bem como teorias de representação do discurso, saliência etc.

⁴ Como em [16], pressupomos que o esforço de interpretação é simplesmente uma função do número de propriedades semânticas da expressão.

2-Seleção de Casos de estudo

Seleção dos tipos de expressão de referência a serem investigados, possivelmente envolvendo descrições definidas com grau variável de redundância lógica e outras.

3-Elaboração de Experimentos

Enunciado de hipóteses compatíveis com as questões do estudo original em [16] e definição das condições de cada experimento. Serão necessárias questões de pesquisa para (a) decidir o sucesso e fracasso de expressões mínimas *versus* expressões redundantes e (b) decidir o sucesso e fracasso de expressões acima do limite superior de redundância útil considerada [15,16].

4-Preparação de Materiais

Uma das razões de sucesso dos experimentos em [16] foi o fato de o contexto de referência ser do tipo físico, com objetos em tipo e número bem definidos, e um número também conhecido de propriedades semânticas facilmente identificáveis. Uma investigação no domínio discursivo, entretanto, coloca a difícil questão de como obter estas mesmas informações a partir de um contexto linguístico suficientemente realista. Por exemplo, para a expressão “este simples exemplo” desta sentença seria necessário definir qual seu termo antecedente (neste caso não há), quais suas propriedades semânticas (e.g., *tipo=exemplo*, *qualidade=simples*) e, o que é mais difícil, quais os outros objetos deste discurso que potencialmente competem por esta mesma descrição.

Como forma de contornar algumas destas dificuldades, uma alternativa a ser considerada é o uso do *cópus semântico TUNA*, já utilizado no *First NLG Shared Task and Evaluation Challenge on Attribute Selection for Referring Expressions Generation*⁵. O *cópus TUNA* consiste de uma coleção de situações de referências livre de realização textual, contendo apenas especificações das entidades envolvidas e suas propriedades, além de uma realização semântica garantidamente correta, obtida através de experimentos com participantes humanos em situações reais de referência. Neste caso, a preparação dos textos a serem empregados no experimento envolveria a criação de textos (possivelmente fictícios) nos quais seriam inseridas situações de referência reais obtidas do *cópus TUNA*, acrescidas das variações necessárias para controle do experimento.

5-Implementação

Desenvolvimento da ferramenta de software para realização do experimento (incluindo interface e mecanismos de monitoramento dos atos de navegação e resolução de referências desempenhados pelos participantes).

6-Execução

Seleção de grupos de participantes e realização dos experimentos definidos.

7-Análise de Dados

Coleta de dados estatísticos sobre os experimentos realizados e sua interpretação.

8-Avaliação Comparativa

Comparação dos resultados obtidos com o do estudo original (para contextos físicos) e enunciado de conclusões sobre a possível generalização do mesmo para o domínio discursivo.

9-Divulgação de resultados

Os resultados de cada etapa do trabalho serão divulgados em conferências e outros eventos da área de PLN à medida que se tornarem disponíveis.

⁵ <http://www.csd.abdn.ac.uk/research/evaluation/>

4.3. Cronograma de Execução

Atividade	Ano 1				Ano 2			
1-Revisão bibliográfica								
2-Seleção de casos de estudo								
3-Elaboração dos experimentos								
4-Preparação de Materiais								
5-Implementação								
6-Execução								
7-Análise de Dados								
8-Avaliação								
9-Divulgação de resultados								

Fig. 5. Cronograma de atividades propostas.

4.4. Recursos

A pesquisa proposta tem em vista o seu desenvolvimento junto à EACH - Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (USP Leste), instituição que se encontra em fase de implantação de seus cursos e linhas de pesquisa, incluindo o Processamento de Linguagem Natural⁶. As atividades acima serão executadas (especialmente os itens 4,5 e 6) em colaboração com alunos de Iniciação Científica da instituição e com pesquisadores do NILC/USP São Carlos, grupo com a qual este autor compartilha uma série de interesses na área⁷. Além disso, contemplamos ainda a possível colaboração com a University of Aberdeen (Aberdeen, Escócia), instituição que colaborou no desenvolvimento do trabalho que deu origem à presente proposta [16].

Os recursos necessários para a execução do projeto são equipamentos básicos de computação para desenvolvimento dos experimentos e verba para participação em eventos científicos onde será feita a divulgação do trabalho:

- 3 computadores e software de apoio, a serem utilizados por alunos de Iniciação Científica no desenvolvimento dos experimentos, redação de artigos etc.
- 1 impressora;
- bens permanentes de apoio (e.g., mobiliário etc.).
- 10 diárias nacionais;
- 2 passagens nacionais;
- 5 diárias internacionais;
- 1 passagem internacional.

5 Considerações Finais

O plano de pesquisa descrito consiste de uma extensão de trabalhos prévios deste autor na área de geração de expressões de referência. A extensão proposta representa um avanço substancial no entendimento das técnicas computacionais para geração destas expressões, e é de interesse imediato para o estudo e desenvolvimento de aplicações de geração de linguagem natural em geral. Este trabalho vem a complementar um projeto de pesquisa relacionado já em andamento na instituição, e também de responsabilidade deste autor, que trata da resolução de expressões de correferência no contexto de sistemas de Tradução Automática⁸, com a perspectiva de longo prazo de consolidação da área de pesquisa em Processamento de Linguagem Natural nesta instituição.

⁶ Atualmente três alunos de Iniciação Científica da USP / EACH desenvolvem atividades relacionadas à presente proposta, contando com apoio PIBIC/CNPq e da Pró-reitoria de Pesquisa da Universidade de São Paulo.

⁷ Incluindo a co-orientação de um aluno de Iniciação Científica em projeto na área de Tradução Automática (Processo Fapesp nro. 2006/04803-7).

⁸ Processo Fapesp nro. 2006/03941-7 (pesquisa individual).

Bibliografia

1. Arts, Anja. Overspecification in instructive texts. Tese de doutorado. Tilburg University, Holanda. Wolf Publishers, Nijmegen (2004).
2. Appelt, D. e A. Kronfeld A computational model of referring. IJCAI (1987).
3. Cheng, H. e C. Mellish. An Empirical Analysis of Constructing Non-restrictive NP modifiers to Express Semantic Relations. 1st International Conference on Natural Language Generation (INLG-2000). Mitzpe Ramon, Israel (2000).
4. Cremers, Anita. Reference to Objects; an empirically based study of task-oriented dialogues. Tese de doutorado. University of Eindhoven (1996).
5. Dale, R. e N. Haddock. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4) (1991).
6. Dale, R. e E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* (19) (1995).
7. Grice, H. P. Logic and Conversation. In P. Cole and J. L. Morgan (eds.) *Syntax and Semantics*, Vol. iii: Speech Acts. New York, Academic Press, (1975) 41-58.
8. Horacek, Helmut. Generating referential descriptions under conditions of uncertainty. 10th European workshop on Natural Language Generation (ENLG-2005). Aberdeen, 58-67.
9. Jordan, P. W. Can Nominal Expressions Achieve Multiple Goals?: An Empirical Study. ACL-2000, Hong Kong, outubro (2000).
10. Krahmer, E. e M. Theune. Efficient Context-Sensitive Generation of Referring Expressions. In *Information Sharing Reference and Presupposition in Language Generation and Interpretation*. Kees van Deemter and Rodger Kibble (eds.) CSLI Publications, Stanford, California (2002) 223-264.
11. Levelt, W.J.M. *Speaking: From Intention to Articulation*. MIT Press, Cambridge (1989).
12. McCoy, K. F. e M. Strube. Generating Anaphoric Expressions: Pronoun or Definite Description? ACL-1999 Workshop "The Relation of Discourse/Dialogue Structure and Reference". University of Maryland (1999).
13. Paraboni, I. e K. van Deemter. Referring via document parts. 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006). Springer Lecture Notes in Computer Science. Cidade do México, fevereiro (2006) 299-310.
14. Paraboni, I. e K. van Deemter. Generating Ease References: the case of Document Deixis. 2nd Intl. Conference on Natural Language Generation (INLG-2002), New York, 113-119.
15. Paraboni, I., J. Masthoff e K. van Deemter. Overspecified reference in hierarchical Domains: measuring the benefits for readers. 4th Intl. Natural Language Generation Conference (INLG-2006). Sydney, Australia (2006).
16. Paraboni, I., K. van Deemter e J. Masthoff. Generating Referring Expressions: Making Referents Easy to Identify. *Computational Linguistics* 33(2) junho (2007) 229-254.
17. Passonneau, R. Interaction of Discourse Structure with Explicitness of Discourse Anaphoric Noun Phrases. M. Walker, A. Joshi, and E. Prince (eds.) *Centering Theory in Discourse*. Clarendon Press, Oxford, UK (1998).
18. Sonnenschein, S. The effect of redundant communication on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research* 13, (1984) 147-166.