# Processamento computacional de anáfora e correferência

## Computational processing of anaphora and coreference

Renata Vieira
Patrícia Nunes Gonçalves
José Guilherme C. de Souza
**Universidade do Vale do Rio dos Sinos**

### Abstract

Referring expressions and their textual design are fundamental factors in discourse interpretation. Computational approaches for interpreting information that is contained in textual bases find in anaphoric and correference processes a great challenge. Recent work shows the use of sophisticated techniques for the discovery of anaphoric relations in texts. This kind of research requires textual data bases where anaphoric relations are identified in an accessible way for computer systems. Building such data bases, also called annotated corpora, is very important for the development of research in this area, considering the Portuguese language. Studies in the area show that semantic and pragmatic knowledge are used not as frequently as other types of anaphoric relations. These are, however, the most difficult cases to solve computationally and recent research in the area show that currently available resources, although very sophisticated, are still not sufficient to deal with the problem in a satisfactory way. On the other hand the importance of anaphora resolution is recognized as essential for other natural language tasks, in this paper we see in particular the case of summarization.

### Keywords

Anaphoric relations; Natural language processing; Summarization.

## Resumo

A estrutura referencial é fundamental para a interpretação do discurso. Abordagens computacionais de interpretação da informação contida em bases textuais encontram nos processos anafóricos e correferenciais um grande desafio. Em trabalhos recentes da área, encontramos o emprego de técnicas sofisticadas para a descoberta de relações anafóricas. Tal pesquisa requer bases textuais em que relações correferenciais estejam identificadas de maneira acessível aos sistemas. A construção dessas bases se faz necessária para viabilizar a realização de pesquisas com a língua portuguesa. Estudos do problema mostram que a freqüência de retomada de um antecedente, baseada em conhecimento (semântico, lexical ou de mundo), é geralmente reduzida em relação ao total de tipos de usos. Apesar de menos freqüente, porém, esse tipo mais elaborado de retomada impõe sérias barreiras ao tratamento computacional do fenômeno. Os recursos disponíveis hoje, apesar de altamente sofisticados, são ainda insuficientes para tratar esse problema de forma satisfatória. No entanto, cada vez mais é reconhecida a importância do tratamento da correferência para outras tarefas de processamento de linguagem natural. Neste artigo analisamos especificamente o exemplo da sumarização automática.

## Palavras-chave

Relações anafóricas; Processamento da linguagem natural; Sumarização automática.

## 1. Introdução

O processamento de anáfora e correferência é uma tarefa relevante e um grande desafio para a área de lingüística computacional. Diversas aplicações, tais como extração de informação, tradução e sumarização, podem se beneficiar do desenvolvimento da área. Tanto questões de informatividade como de legibilidade são intimamente ligadas com o projeto referencial embutido no texto pelo autor. No entanto, esse é um problema de alta complexidade cognitiva e, conseqüentemente, computacional. Sistemas atuais utilizam recursos muito sofisticados, como técnicas de aprendizado de máquina com base em *corpus* anotado. Apesar de sofisticadas, as técnicas desenvolvidas até hoje ainda são em grande parte insuficientes. Neste artigo, apresentamos diferentes casos de anáforas que ilustram a dificuldade do problema e mostramos como os sistemas computacionais procuram resolvê-lo. Por fim, relacionamos o problema da resolução anafórica e correferencial com questões de coesão textual e sua importância na construção de sistemas práticos, tais como sumarizadores automáticos.

## 2. Anáfora e correferência

A anáfora pode ser definida como a retomada de uma expressão apresentada anteriormente em um texto. Quando uma entidade é mencionada pela primeira vez textualmente, temos a *evocação* da entidade. Durante a leitura da seqüência do texto, quando essa entidade é novamente mencionada, temos a realização do *acesso* a essa entidade. A expressão que faz o acesso é dita como *anafórica* e a expressão anterior é dita como seu *antecedente*. A relação entre essas duas expressões (anáfora e antecedente) é dita como *relação de correferência* (JURAFSKY; MARTIN, 2000). De uma forma geral, são os sintagmas nominais as estruturas textuais (expressões) utilizadas para evocação e acesso de entidades mencionadas em um texto. De acordo com Perini (1995), o sintagma nominal pode se tornar uma estrutura bem complexa, pois pode apresentar grandes

diferenças estruturais, por exemplo, apresentar determinantes ou modificadores. Esses elementos podem ser observados nos exemplos seguintes:

- Núcleo nome próprio: "*William Eberhard* descobriu que as larvas provocam mudanças no comportamento da hospedeira".
- Núcleo substantivo comum: "*Pesquisas* em camundongos foram realizadas".
- Determinantes: O uso de determinantes é muito comum em sintagmas nominais, podem ser artigos definidos, indefinidos, adjetivos entre outros. "*Os pingüins* são acostumados a mar aberto".
- Modificadores
  - Pré-modificadores: aparecem antecedendo o núcleo. "*O pequeno astro* vai passar a uma certa distância do Sol".
  - Pós-modificadores: aparecem após o núcleo. "*Amostras celulares de animais ameaçados de extinção* foram coletadas".

De acordo com o tipo de sintagma, a anáfora pode ser pronominal (ele, ela, seu, sua, isso), ou definida (uma criança... o menino). No caso das anáforas definidas, a relação com o antecedente pode envolver relações semânticas mais simples como identidade (uma criança, o menino), ou relações mais complexas (uma criança, a mãe). Ainda em relação à posição do antecedente, ela pode ser intra-sentencial ou intersentencial. O estudo de correferência textual geralmente inclui todos os tipos de anáfora, pois, muitas vezes, uma mesma cadeia referencial (o conjunto total de expressões utilizadas no texto para evocar e acessar uma entidade) contém diversos tipos de expressões anafóricas.

De acordo com a sua condição em relação aos atos de evocação e acesso, as expressões referenciais podem ter *status* diferenciados. Em Vieira (1998) e Collovini e Vieira (2006a), encontramos uma classificação das expressões referenciais quanto ao seu *status*: novas no discurso, anáforas diretas, anáforas indiretas e associativas.

Quando um sintagma nominal introduz um novo referente (evocação), sem apresentar parte de seu sentido ancorado em uma expressão anterior, é considerado novo no discurso. As expressões dadas como novas no discurso não são anafóricas, já que são mencionadas pela primeira vez. No decorrer do

discurso, outras expressões serão utilizadas fazendo uma referência a uma entidade mencionada anteriormente. Portanto, as expressões novas no discurso podem servir de antecedente para as anáforas.

A anáfora direta é aquela que possui antecedente e estabelece com ele uma relação de identidade; além disso, sua expressão lingüística (sintagma nominal) apresenta o mesmo nome-núcleo do antecedente. Por exemplo:

> "Um grupo que reúne 13 sociedades científicas nacionais enviou *uma carta* ao Senado Federal para pedir mudanças no projeto da nova Lei de Biossegurança. Na carta os cientistas falam sobre células-tronco."

O sintagma nominal "a carta" é considerada uma anafóra direta, pois possui o mesmo nome-núcleo de seu antecedente.

A anáfora indireta é também caracterizada pela relação de identidade com o antecedente, mas o acesso é feito a partir de um sintagma que não possui o mesmo nome núcleo do seu antecedente. Vejamos o exemplo:

> "Os EUA foi (sic) um dos últimos países a assinar *a Declaração de Helsinque. O texto* traça diretrizes para ética em pesquisas..."

Nesse exemplo, o termo "O texto" está se referindo a "a Declaração de Helsinque"; como podemos notar, as expressões não possuem o mesmo nome núcleo, mas os dois termos referem-se à mesma entidade. As anáforas indiretas são geralmente baseadas em processos cognitivos mais complexos, como ocorre nos processos inferenciais nos quais o leitor ativa a representação da informação armazenada em sua memória por meios variados, envolvendo conhecimento semântico e pragmático. A classe anafórica indireta possui, portanto, vários tipos. A seguir temos alguns exemplos:

Relação entre nome próprio e nome comum:

> "Não temos certeza de que aquela carga era ilegal, mas sabemos que 80% da atividade madeireira no Brasil é irregular", disse Rebeca Lerer, ativista brasileira d*o Greenpeace*. Para a *ONG*, há evidências de que as companhias que mais exportam madeira para os EUA estejam envolvidas com o comércio ilegal do produto.

Relação de sinonímia:

> "Isso quer dizer que os *camundongos* transgênicos reduziram a gordura de seu corpo. Os *ratos* estudados..."

Nominalização de verbos:

> "O presidente da Comissão Nacional de Ética em Pesquisa *propôs* na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência.... *A proposta* foi discutida pelos cientistas..."

Hiponímia/hiperonímia:

> "As mudanças nas populações de *pingüins* também serviram como indicativo do problema climático. *Os animais* usavam geleiras para se abrigar e procriar."

A anáfora associativa introduz um novo referente no discurso, entretanto seu significado está fortemente ancorado em uma expressão anterior. A anáfora associativa pode ser de vários tipos. Vejamos alguns exemplos:

Relação conjunto/subconjunto:

> "Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual *as cidades em regiões amazônicas ocupadas de forma predatória* duram por volta de 23 anos. Ele citou como exemplo *as cidades de Paragominas (PA), Açailândia (MA) e Humaitá (AM)*."

Relação grupo/membros:

> "Um tratamento para a obesidade que faz você perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo realizado por *um grupo de cientistas britânicos* será publicado hoje na revista Nature. *Um dos cientistas, John Clapham*, diz que esse é um alvo viável para remédios contra a obesidade."

Relação objeto/substância:

> "*Uma estrela* é composta de *gás hidrogênio* condensado pela gravidade."

Relação entidade/atributo:

> "O mecanismo que faz *as pessoas* sentirem falta de ar em regiões montanhosas...Cientistas descobriram que esses gases atuam na regulação respiratória, fazendo com que *os vasos sanguíneos e vias respiratórias* dilatem."

Relação parte/todo:

> "As larvas ao parasitar *a aranha* provocam mudanças no comportamento da hospedeira. A relação espúria começa no *abdome*."

Como podemos observar, evocações e acessos ocorrem de formas muito variadas. Para o processamento de língua natural (PLN), esse fenômeno é de grande relevância no tratamento da informação veiculada, porém impõe grande dificuldade. O fenômeno da anáfora e da correferência têm um papel importante na construção do sentido, tanto na veiculação da informação quanto na estruturação global do texto. Além disso, têm relação com aspectos muito estudados do discurso, como a coesão e coerência, que serão discutidos a seguir.

## 3. Coerência e coesão referencial

Como observado em Koch (2003), as anáforas possuem um papel importante na construção da coerência de um texto. Não apenas na coerência, mas também na compreensão global e sentido do texto. Durante a leitura, o leitor realiza o processamento textual, e, por meio de representações de entidades no texto, faz uso do encadeamento referencial para resolver qual das entidades descritas deve ser selecionada para interpretação do sentido do texto. Enquanto a noção de coerência se relaciona com a linearidade e o sentido do texto, a noção de coesão diz respeito à superfície textual, isto é, ao uso de mecanismos coesivos para realizar a conexão entre termos e frases. A coesão subdivide-se em dois grandes grupos (KOCH; TRAVAGLIA, 1996): coesão referencial e coesão seqüencial. A coesão referencial faz uso do mecanismo de reiteração, utilizando, por exemplo, o emprego de sinônimos, meronímia, hiperonímia e nomes genéricos, ilustrados a seguir.

Sinônimos:

> "Um *garoto* estava correndo. *O menino* estava apavorado".

Meronímia:

> "*O carro* roubado foi encontrado. *Os pneus* não estavam no veículo."

Hiperonímia:

> "Dentre *os mamíferos* estudados para essa pesquisa, *a vaca* foi escolhida."

Nomes genéricos:

> "Todos ouviram o barulho *da moto*. Olharam para o fim da rua e viram *a coisa* chegando rápido."

A coesão seqüencial diz respeito à progressão textual, em que existem elementos que se unem para dar a idéia de seqüencialidade e continuidade da idéia central do texto. Num texto coeso, as partes são interdependentes e importantes para a compreensão geral, fenômeno chamado de progressão textual. Mais especificamente, a coesão seqüencial por progressão é utilizada para possibilitar manutenção temática e encadeamentos. A manutenção temática faz uso de termos com a mesma contigüidade semântica, por exemplo:

> "Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram a descoberta de *uma nova espécie de dinossauro* no Brasil. *O animal* que na cadeia evolutiva d*os dinossauros* ocuparia uma posição n*o grupo Tyrannoraptora*, o mesmo do *Tyrannossaurus Rex*, habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo. *O fóssil, batizado de Santanaraptor placidus***,** é o único a ser encontrado no país com tecidos preservados. Isso pode permitir que os cientistas saibam mais sobre o modo de vida e a evolução d*os répteis*."

Por meio dos termos que estão destacados nesse exemplo, é possível que o leitor ative seu esquema cognitivo, desfazendo ambigüidades e avançando na

perspectiva do texto. Esse encadeamento permite estabelecer relações semânticas entre orações, enunciados ou seqüências textuais. Com base nos exemplos apresentados, podemos ver que a coesão lexical/referencial é um dos meios predominantes de conexão entre sentenças. Os mecanismos mais utilizados na correferência são a repetição (anáfora direta) e a substituição (anáfora indireta). A diversidade observada nos exemplos apresentados está relacionada com a questão da complexidade do tratamento do fenômeno, relativa não apenas ao processamento computacional da língua, mas também ao processamento cognitivo. Em Hickman (1980), são reportados indícios de que a coesão referencial se desenvolve até os 10 anos de idade; em Roth, Spekman e Fye (1995), são apresentados estudos indicando que estudantes com dificuldades de aprendizado têm mais problemas com narrativas do que com sintaxe. Como esses estudos sugerem, a habilidade discursiva não é apenas um problema difícil para o PLN, mas também um problema que apresenta uma alta complexidade para os falantes.

## 4. Resolução de correferência: estudos baseados em *corpus*

Vimos a importância da coesão lexical na manutenção da coerência de um texto. Essa coerência garante o acesso correto ou mais fiel possível à informação intencionada e projetada pelo autor ou interlocutor. Capturar essas relações para reproduzir a informatividade de modo fiel é um dos objetivos dos estudos realizados na área de lingüística computacional e voltados para a resolução automática de correferência. Sendo um problema muito difícil, que envolve uma combinação de vários fatores lingüísticos e extralingüísticos na sua construção, os sistemas computacionais voltados para esse tipo de problema são bastante complexos. Recursos de inteligência artificial, como o aprendizado de máquina, são utilizados nos tratamentos computacionais dados ao problema. Nessas abordagens, um conjunto de características lingüísticas, identificadas num *corpus* previamente anotado com correferência, é analisado automaticamente para que se identifiquem relações entre elas e a presença de ligação entre as expressões.

Essas características geralmente resultam de processamento automático realizado em outros níveis (sintáticos e/ou estatísticos). Exemplificamos adiante algumas das características utilizadas por esses sistemas. Os exemplos a seguir apresentam características utilizadas no desenvolvimento de um sistema para resolução de correferência da língua portuguesa:

1. Comparação de núcleo: caso o núcleo dos dois sintagmas seja igual, o valor desse atributo é verdadeiro. Caso contrário, falso.
2. Distância: os valores possíveis desse atributo são números inteiros maiores que 0. Determinam a distância em frases entre os dois sintagmas. Se os dois sintagmas estão na mesma frase, a distância é 0. Se a anáfora está uma frase adiante, o valor é 1 e assim por diante.
3. Antecedente é pronome: caso o núcleo do sintagma nominal antecedente seja um pronome, recebe verdadeiro. Caso contrário, recebe falso.
4. Anáfora é pronome: caso o núcleo do sintagma nominal da anáfora seja um pronome, recebe verdadeiro. Caso contrário, recebe falso.
5. Concordância de gênero: caso o gênero dos núcleos dos dois sintagmas coincidam, recebe verdadeiro. Caso não coincidam, o valor desse atributo é falso.
6. Concordância de número: se o núcleo dos dois sintagmas concordam em número (ou seja, ambos estão no singular ou ambos no plural), o valor desse atributo é verdadeiro. Caso contrário, é falso.
7. Sujeito: esse atributo é verdadeiro, caso ambos sintagmas sejam sujeitos e falso se ocorrer o contrário.
8. Concordância semântica: caso os dois nomes núcleos sejam diferentes e possuam tipos semânticos idênticos ou similares, o valor desse atributo é verdadeiro. Caso contrário, é falso.

Entre os vários sistemas reportados na literatura, o número de características observadas é bastante variado. Em McCarthy e Lehnert (1995), Fisher *et al.* (1995) e McCarthy (1996), o estudo é realizado com base em oito características para cada par, sendo três delas informações específicas do domínio dos textos que compunham o *corpus* utilizado para os experimentos da ferramenta. Assim, o sistema está atrelado ao domínio dos textos do *corpus* utilizado. Soon, Ng e Lim (2001) apresentam uma abordagem para o problema da resolução de correferência em textos de qualquer domínio e para qualquer tipo de sintagma nominal. Foram consideradas 12 características como indicativas de anaforicidade, contendo informações posicionais, sintáticas, morfológicas e semânticas. Para as informações semânticas, foi utilizada a base de dados lexical WordNet (FELLBAUM, 1998). Assim como Soon, Ng e Lim (2001), outro trabalho que

apresenta uma solução independente de domínio é o desenvolvido em Vieira e Poesio (2000). O sistema proposto processa descrições definidas. Foram desenvolvidos métodos heurísticos para: (a) resolver descrições definidas anafóricas diretas, (b) identificar descrições novas no discurso e (c) identificar uma âncora da descrição associativa e a relação semântica entre a descrição associativa e sua âncora. Cabe ressaltar, no entanto, que os métodos não são baseados em aprendizado de máquina supervisionado e sim em heurísticas desenvolvidas pelos autores.

Soon, Ng e Lim (2001) adotam a abordagem de aprendizado de máquina supervisionado, com um conjunto de características igual a 53. As informações para a composição das características são provenientes de dados lexicais, de dados semânticos e baseados em conhecimento, além de 26 características gramaticais que contêm uma série de restrições lingüísticas e preferências. Apesar do aumento do número de características, o efeito sobre as taxas de acerto não é muito significativo. Esses resultados corroboram a dificuldade constatada do tratamento desse fenômeno. A seguir, apresentamos o *corpus* Summ-it, desenvolvido para embasar estudos sobre correferência textual e o seu tratamento computacional, bem como a relação da sumarização automática e o processo de correferência.

### 4.1. *Corpus* Summ-it

Para estudar em detalhe o problema exposto até aqui, consideramos um *corpus* anotado com informações de relações anafóricas e correferenciais, denominado Summ-it (COLLOVINI *et al*., 2007). O *corpus* constitui-se de 50 textos jornalísticos da *Folha de São Paulo*, retirados do caderno de ciências do jornal, escritos em português do Brasil, e disponibilizado através do Projeto PLN-BR. O *corpus* foi processado pelo analisador sintático PALAVRAS (BICK, 2000) e anotado manualmente com informações de correferência, utilizando-se a ferramenta MMAX (MÜLLER; STRUBE, 2001). O processo de anotação foi baseado em projetos anteriores, tais como MUC,[1] VENEX[2] e MATE.[3] O *corpus* Summi-it possui um total de 5047 sintagmas nominais, compondo 560 cadeias de correferência. A cadeia mais extensa possui 16 elementos. Nas tabelas 1 e 2, a seguir, podemos verificar a distribuição das configurações morfossintáticas encontradas no *corpus*.

TABELA 1
Configuração dos sintagmas nominais

| Sintagmas | # (%) |
|---|---|
| Definidos | 2068 (40,95%) |
| Sem determinante | 1134 (22,46%) |
| Nome próprio com determinante definido | 386 (7,64%) |
| Indefinido | 383 (7,58%) |
| Nome próprio sem determinante | 308 (6,10%) |
| Determinante numeral | 155 (3,07%) |
| Determinante quantificador | 110 (2,18%) |
| Coordenados | 98 (1,94%) |
| Demonstrativo | 90 (1,78%) |
| Possessivo | 73 (1,45%) |
| Interrogativo | 2 (0,04%) |
| Total | 4804 (95,18%) |

TABELA 2
Configuração dos sintagmas pronominais

| Pronomes | # (%) |
|---|---|
| Pessoal | 152 (3,01%) |
| Demonstrativo | 35 (0,69%) |
| Numeral | 27 (0,53%) |
| Indefinido | 23 (0,46%) |
| Interrogativo | 6 (0,12%) |
| Possessivo | 0 (0%) |
| Total | 243 (4,82%) |

A tabela 3, a seguir, ilustra os resultados da anotação das descrições definidas, seguindo a classificação nas quatro classes: novas no discurso, direta, indireta e associativas.

TABELA 3
Classificação de descrições definidas

| Classificações | Quantidade |
|---|---|
| Novas no Discurso | 1428 |
| Anáforas Associativas | 183 |
| Anáforas Diretas | 407 |
| Anáforas Indiretas | 291 |
| Total de descrições definidas classificadas | 2309 |

Como podemos observar por essa análise, grande parte dos sintagmas do *corpus* é do tipo descrições definidas (40%); entre essas, uma parcela significativa é anafórica. As anáforas diretas são mais numerosas. Os outros tipos (associativas e indiretas), são aqueles em que o conhecimento refinado semântico e pragmático se faz mais presente no processo de resolução. Os tipos de conhecimento envolvidos no processo de referenciação são variados. Como citado na Seção 3, a retomada de referentes mais concretos pode se dar por repetição (muito usada) ou por substituição. O desempenho obtido pelos sistemas de resolução é relativo, em sua grande maioria, aos casos de retomada nos quais a necessidade de conhecimento semântico/pragmático e de mundo se dá de forma mais básica, como no caso de pronomes, repetição simples, ou entidades nomeadas. Mas, muitas vezes, a retomada é baseada em relações semânticas. A complexidade dessas relações é bastante variada, como pode ser observado nos exemplos a seguir (extraídos do *corpus*):

- uma gripe mortal – a doença (este exemplo, uma relação simples de hipo e hiperonímia se estabelece entre referente e antecedente.)
- pesquisadores – a equipe (este caso possui uma relação semântica menos óbvia, em que está implícita a idéia de equipe de pesquisadores. O interessante nesse par é que ele viola concordância de número, uma restrição geralmente adotada pelos sistemas)
- patenteamento de genes – o assunto (este exemplo, existe uma grande distância semântica entre referente e antecedente, devido ao fato de ser um referente bastante genérico e abstrato.)
- a vespa – o inseto – o parasita – o invasor (Esta seqüência inicia com uma relação semântica simples e continua com relações que

são tipicamente baseadas no discurso textual. Neste texto, a vespa é um parasita da aranha, e viola concordância de gênero.)

- a aranha – a hospedeira – o anfitrião – o aracnídeo – a vítima (como no exemplo anterior, as relações são dependentes de informações do contexto textual e violam restrições básicas, como o genêro.)

Um outro exemplo, retirado de um *corpus* da língua inglesa, mostra uma situação na qual um nome próprio tem duas opções distintas de antecedentes. Um se refere à empresa e outro, à pessoa, ambos com o mesmo nome. Adicionalmente, a interpretação só é possível a partir do contexto textual: "*Snyder Communications Inc. of Bethesda" – "Daniel M. Snyder – "Snyder"*

Recentemente, tratamentos semânticos têm sido propostos ao problema da resolução. Ponzetto e Strube (2006) avaliam o impacto de diversos recursos, em especial para a língua inglesa, em que tanto a atividade como a disponibilidade de recursos é mais abundante. Exemplos desses recursos são: Internet, Wikipedia e WordNet. Apesar da sofisticação dos recursos, os resultados não são muito animadores, a qualidade de respostas em pesquisas baseadas em *corpus* (medida em termos de F-measure, uma medida de balanceamento entre abrangência e precisão) fica em torno de 70%. Outros problemas tratados em PLN, como a análise sintática, reportam resultados acima de 90%. Cabe lembrar que, assim como reportado em Soon, Ng e Lim (2001), grande parte dos acertos obtidos pelos sistemas referem-se a casos de similaridade lexical (repetição). Para um avanço na qualidade desses sistemas, é preciso tratar os casos de substituição (anáforas indiretas e associativas), e, para isso, uma maior compreensão do fenômeno e esforços interdisciplinares são necessários.

### 4.2 Sistemas para a LP

Ainda que de forma mais tímida do que para a língua inglesa, existem iniciativas de tratamento da resolução anafórica para o português. O trabalho desenvolvido por Coelho e Carvalho (2005) implementa o algoritmo de Lappin e Leass (1994) para resolução anafórica pronominal em textos da língua portuguesa. Para esse trabalho, um *corpus* anotado com informações morfológicas e sintáticas foi utilizado. Baseado nessa informação, o algoritmo procura pronomes em um texto e busca reconhecer seu antecedente. Chaves (2007) apresenta uma adaptação do algoritmo de Mitkov (2002) para a língua portuguesa.

Essas abordagens resolvem somente anáforas pronominais e que não utilizam aprendizado de máquina nem conhecimento semântico.

Em Collovini (2005) e Collovini e Vieira (2006b), são apresentados experimentos com o objetivo de classificar de forma automática as descrições definidas em quatro classes: novas no discurso, anáforas diretas, anáforas indiretas e associativas. Para essa tarefa, foram extraídas 16 características morfológicas e sintáticas para o aprendizado de máquina. Um dos problemas com o aprendizado em relação às classes que consideramos mais interessantes (associativas e indiretas) é que o baixo número de exemplos faz com que o classificador, que é inferido automaticamente, tenda a privilegiar as outras classes (mais numerosas). Os resultados obtidos para essas classes são menos favorecidos. Por esse motivo, em Collovini e Vieira (2006a), uma técnica de balanceamento de *corpus* por repetição de exemplos é avaliada com o objetivo de melhorar os resultados, reportando uma melhora na classificação.

Em Coelho *et al*. (2006), é apresentado um primeiro estudo de *corpus* sobre resolução das descrições definidas utilizando a informação semântica fornecida pelo analisador sintático PALAVRAS (BICK, 2000). Esse trabalho teve como objetivo abordar especificamente a resolução de anáforas associativas e indiretas. Ribeiro Jr. *et al*. (2007) propõem uma combinação das duas técnicas apresentadas nos trabalhos de Collovini e Vieira (2006a) e Coelho *et al*. (2006), utilizando tanto as informações semânticas para classificação das expressões nas quatro classes como a técnica de balanceamento de corpus. Foram implementadas as características inicialmente apresentadas em Collovini e Vieira (2006a) mais outras duas baseadas em informações semânticas (fornecidas pelo parser PALAVRAS). Em Souza (2007) é apresentado o primeiro sistema de resolução de correferência para a língua portuguesa, baseado em corpus e utilizando técnicas de aprendizado de máquina. A partir do surgimento desses sistemas e da possibilidade de tratarmos o problema de uma forma automática, passamos a investigar a utilidade desses resultados em outras aplicações de tratamento textual. A seguir, discutimos a aplicação de correferência na sumarização automática.

## 5. Aplicando resolução de correferência em sumarização

Uma das aplicações de PLN que podem se beneficiar da existência de um sistema de resolução anafórica é a sumarização. Uma das técnicas de sumarização mais utilizada é a da sumarização extrativa em que o processamento indica

sentenças mais relevantes através da freqüência de palavras e outras técnicas similares e apresenta um sumário constituído pela seqüência das sentenças com maior pontuação. É claro que os sumários extrativos (por eliminarem partes do texto) podem facilmente corromper a coesão de um texto e, conseqüentemente, sua coerência.

Na principal conferência de avaliação de sistemas de sumarização automática, a *DUC Document Understanding Conference* (http://duc.nist.gov/), a avaliação de qualidade de sumários apresenta como um dos critérios de avaliação a clareza referencial, assim descrita:

– Deve ser fácil identificar a quem ou a que os pronomes e sintagmas nominais do sumário se referem;
– Se uma pessoa ou outra entidade for mencionada, seu papel na história deve ser claro;
– Uma referência não é clara, se uma entidade for referenciada, mas sua identidade ou relação com o resto da estória não estiver clara.

Esses critérios ilustram a importância da questão de resolução anafórica nesse contexto. Recentemente, temos investigado a aplicação da resolução de correferência em sumarização. São duas as questões principais envolvidas nessa relação:

– a correferência pode guiar a seleção de sentenças em sumarização extrativa?
– a correferência pode ser usada para recuperar coesão referencial de sumários?

Nossos estudos estão focados na segunda questão. Estamos atualmente desenvolvendo e avaliando métodos de recuperação de coesão textual dos sumários. Como a construção do *corpus* Summ-it objetivava o estudo de sumarização, além da informação de correferência, cada texto do *corpus* possui um sumário manual feito por sumarizadores humanos (COELHO, 2007). Além disso, são disponibilizados os extratos ideais, formados pelos textos-fonte com a indicação das sentenças mais relevantes e os sumários extraídos automaticamente com o sumarizador Gist-Summ (PARDO, 2005). A partir desses sumários e extratos, podemos realizar diversas análises, por exemplo, observar as sentenças em comum encontradas nos extratos e nos sumários manuais e verificar o processo de reescrita na sumarização. Um exemplo é apresentado a seguir:

Sumário

> Pesquisadores do *Museu Nacional* do Rio de Janeiro anunciaram a descoberta de uma nova espécie de *dinossauro* no Brasil. O animal que na cadeia evolutiva dos dinossauros ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo. *O fóssil, batizado de Santanaraptor placidus, é o único a ser encontrado no país com* tecidos preservados. Isso pode permitir que os cientistas saibam mais sobre o modo de vida e a evolução dos répteis.

Extrato

> *Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com* restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. "É como se o *dinossauro* tivesse sido enterrado ontem", disse Alexander Kellner, geólogo do Setor de Paleovertebrados do *Museu Nacional* e coordenador da expedição que encontrou o fóssil na região da Chapada do Araripe, Ceará (veja mapa). O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem do fóssil só foi concluída nove anos mais tarde.

No exemplo apresentado, destacamos os termos comuns e podemos observar que há uma sentença com maior grau de similaridade (sentença sublinhada). Nessa sentença, a referência encontrada no sumário foi reescrita, e nessa reescrita houve deslocamento do aposto e agrupamento de síntese (tecidos). Pode-se observar que essa reescrita envolve reestruturação da expressão referencial. Esta é uma questão interessante a ser observada na sumarização. Outra análise possível é a comparação entre cadeias presentes no sumário e no extrato, conforme destacado no trecho seguinte:

Sumário

> Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram a descoberta de *uma nova espécie de dinossauro* no Brasil. *O animal* que na cadeia evolutiva dos dinossauros ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex,

> habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo. *O fóssil, batizado de Santanaraptor placidus*, é o único a ser encontrado no país com tecidos preservados. Isso pode permitir que os cientistas saibam mais sobre o modo de vida e a evolução dos répteis.

Extrato

> *Batizado de Santanaraptor placidus, o fóssil* é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. "É como se *o dinossauro* tivesse sido enterrado ontem", disse Alexander Kellner, geólogo do Setor de Paleovertebrados do Museu Nacional e coordenador da expedição que encontrou o fóssil na região da Chapada do Araripe, Ceará. *O exemplar de Santanaraptor* encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem *do fóssil* só foi concluída nove anos mais tarde.

No exemplo anterior, podemos observar que as cadeias resultantes no sumário e no extrato compartilham duas de três expressões (fóssil e dinossauro). No sumário, é preservada uma construção mais típica de introdução do referente a partir do sintagma nominal indefinido. Isso poderia também indicar uma preferência na composição do sumário. Uma vez indicado o elemento fóssil como relevante no texto, as sentenças poderiam ser escolhidas mediante a observação dos elementos de progressão textual.

Por fim, temos a análise do impacto da substituição de expressões para a coesão dos sumários. Em alguns casos, a sentença selecionada para um sumário extrativo contém uma anáfora cujo antecedente não está incluído no sumário. O seguinte exemplo demonstra isso:

> Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não conectadas.

Obviamente, ao analisar esse exemplo, percebemos que não é possível atender aos critérios de coesão apontados na avaliação do DUC. Não é possível interpretar o pronome "ele" no início da frase. A cadeia, recuperada do texto-fonte, é formada pela seguinte seqüência de expressões:

– Barry Ellman, do Centro para Estudos Urbanos e Comunitários de a Universidade de Toronto, Canadá
– o pesquisador
– ele
– Ellman
– ele
– o pesquisador

Se realizarmos a substituição do pronome no sumário pelo item mais expressivo da cadeia, teremos um sumário mais coeso e coerente, como segue:

> Segundo Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não conectadas.

Um estudo detalhado dessa última questão, a substituição de cadeias de correferência em sumários extrativos, é apresentado em Gonçalves (2008). Com a disponibilidade de sistemas que realizam de forma automática a composição das cadeias em um texto, poderíamos contribuir para a melhoria da tarefa de sumarização automática, um recurso que é relevante e muito desejado em diversos domínios da atividade humana. Atualmente, a maior dificuldade encontrada por esses sistemas de resolução reside no reconhecimento dos ligamentos semânticos e pragmáticos. Entender melhor a habilidade humana em produzir e compreender o discurso, e, em particular, a habilidade de gerar e recuperar discursos coesos, são desafios que só podem ser encarados de forma colaborativa e interdisciplinar, unindo o processamento de linguagem natural e os estudos de linguagem e cognição.

## 6. Conclusão

Este trabalho aponta para a importância e complexidade do desenvolvimento de sistemas para a resolução de correferência. A pesquisa nessa área busca meios de melhorar os resultados até então obtidos pelos sistemas. Um estudo mais detalhado dos casos nos mostra a complexidade semântica e pragmática desse fenômeno, e, aparentemente, temos ainda pouca compreensão sobre esses processos do ponto de vista da cognição humana. O entendimento e o tratamento

computacional desse fenômeno são relevantes para o desenvolvimento da tecnologia de informação. Esses recursos serão cada vez mais necessários para o tratamento da informação em grande quantidade, cenário que já se configura nas mais diversas atividades humanas. Discutimos e exemplificamos o caso específico da sua aplicação na tarefa de sumarização automática. Apontamos para a complexidade do problema e a necessidade da pesquisa interdisciplinar.

## Notas

[1] http://www-nlpir.nist.gov/related_projects/muc/

[2] http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf

[3] http://mate.nis.sdu.dk/

## Referências Bibliográficas

BICK, E. *The Parsing System "PALAVRAS" - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* 2000. Tese (Doutorado) – Department of Linguistics, University of Århus, DK., 2000.

CHAVES, A. R. *A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov*. 2007. Dissertação (Mestrado) – Universidade Federal de São Carlos, 2007.

COELHO, J. C. B. *Uso de Informação de Correferência e Anáfora para Verificação da Coesão e Coerência Textual na Sumarização Automática.* Junho 2007. Trabalho de Conclusão de Curso de Letras. Unisinos - São Leopoldo.

COELHO, J. C. B. *et al*. Resolving portuguese nominal anaphora. In: VIEIRA, R. *et al*. (Ed.). *7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006).* Itatiaia, RJ: Springer, 2006.

COELHO, T. T. *Resolução de anáfora pronominal em português utilizando o algoritmo de Lappin e Leass.* 2005. Dissertação (Mestrado) – Departamento de Computação, Universidade Estadual de Campinas - Unicamp, 2005.

COLLOVINI, S. *Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa*. 2005. Dissertação (Mestrado) – Departamento de Computação, Universidade do Vale do Rio dos Sinos – Unisinos, 2005.

COLLOVINI, S. *et al*. Summit: Um *corpus* anotado com informações discursivas visando à sumarização automática. In: *5o Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007).* Rio de Janeiro, RJ: Proceedings of the SBC, 2007.

COLLOVINI, S.; VIEIRA, R. Anáforas nominais definidas: balanceamento de corpus e classificação. In: *IV Workshop de Tecnologia da Informação e Linguagem Humana TIL*. Ribeirão Preto, SP: Proceeding of the Brazilian Symposium on Artificial Intelligence, 2006a.

COLLOVINI, S.; VIEIRA, R. Análise de expressões referenciais em corpus anotado da língua portuguesa. In: *V Best MSc dissertation/PhD thesis contest (CTDIA'2006).* Ribeirão Preto, SP: Proceedings of the SBIA-IBERAMIA, 2006b.

FELLBAUM, C. *WordNet: An Electronical Lexical Database*. Cambridge, MA: The MIT Press, 1998.

FISHER, D. *et al*. Description of the umass system as used for muc-6. In: *MUC6 '95: Proceedings of the 6th conference on Message understanding*. Morristown, NJ, USA: Association for Computational Linguistics, 1995. p. 127-140.

GONÇALVES, P. N. Aplicando *Cadeias de Correferência na revisão de Sumários Extrativos.* Dissertação (Mestrado) – Departamento de Computação, Universidade do Vale do Rio dos Sinos – Unisinos, 2008 (em preparação).

HICKMAN, M. Creating referents in discourse: a developmental analysis of linguistic cohesion. In: OJEDA, J.; KREIMAN, A. E. (Ed.). *Papers form the parasession on pronouns and anaphora*. Chicago: Linguistic Society, 1980. p. 192-203.

JURAFSKY, D.; MARTIN, J. Speech and language processing. In: . [S.l.]: Alan Apt, 2000. cap. *Discourse*, p. 670-718.

KOCH, I. G. V. *Desvendando os Segredos do texto*. [S.l.]: São Paulo: Cortez, 2003.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coesão textual*. [S.l.]: São Paulo: Contexto, 1996.

LAPPIN, S.; LEASS, H. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, v. 20(4), p. 535-561, 1994.

MCCARTHY, J. F. *A trainable approach to coreference resolution for information extraction*. 1996. Tese (Doutorado) – Director-Wendy G. Lehnert.

MCCARTHY, J. F.; LEHNERT, W. G. Using decision trees for coreference resolution. In: *Proceedings of the 14th IJCAI*. Montreal, Canada: [s.n.], 1995. p. 1050-1055.

MITKOV, R. *Anaphora Resolution*. [S.l.]: Longman, 2002.

MÜLLER, C.; STRUBE, M. Mmax: A tool for the annotation of multimodal corpora. In: *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle, Washington: [s.n.], 2001. p. 45-50.

PARDO, T. *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades* [S.l.], 2005.

PERINI, M. A. *Gramática descritiva do português*. São Paulo: Ática, 1995. 308 p.

PONZETTO, S. P.; STRUBE, M. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA: Association for Computational Linguistics, 2006. p. 192–199. Disponível em: <http://www.aclweb.org/anthology/N/N06/N06-1025>.

RIBEIRO Jr, L. C. *et al*. Uso de informações semânticas na identificação de anáforas indiretas e associativas. In: *5o Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*. Rio de Janeiro, RJ: Proceedings of the SBC, 2007.

ROTH, F. P.; SPEKMAN, N. J.; FYE, E. C. Reference cohesion in the oral narratives of students with learning disabilities and normally achieving students. In: *Learning Disability Quarterly*. [S.l.: s.n.], 1995. v. 18, n. 1, p. 25-40.

SOON, W. M.; NG, H. T.; LIM, D. C. Y. *A machine learning approach to coreference resolution of noun phrases*. v. 27, n. 4, p. 521–544, 2001. Disponível em: <http://www.aclweb.org/anthology/J01-4004.pdf>.

SOUZA, J. G. C. de. *Resolução automática de correferência aplicada à língua portuguesa.* Novembro 2007. Trabalho de conclusão.

VIEIRA, R. *Definite description processing in unrestricted text*. 1998. Tese (Doutorado) – University of Edinburgh, Edinburgh, 1998.

VIEIRA, R.; POESIO, M. An empirically-based system for processing definite descriptions. *Computational Linguistics*, v. 26, n. 4, p. 539-594, 2000.