Uma ferramenta para resolução automática de correferência

Caroline Gasperin , Rodrigo Goulart , Renata Vieira

[1]Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
Unisinos
São Leopoldo, RS, Brasil

{caroline,rodrigo,renata}@exatas.unisinos.br

Resumo. *Este trabalho apresenta a arquitetura de uma ferramenta para resolução automática de correferência em textos. A ferramenta trata corpora analisados sintaticamente e codificados em XML com base nos formatos utilizados pela ferramenta de anotação MMAX. Tais formatos são detalhados e as etapas do processo de resolução das anáforas são descritas.*

Abstract. *This work presents the architecture of a tool for automatic coreference resolution in texts. The tool deals with syntatic analysed corpora encoded in XML based on the formats given by the anotation tool MMAX. The anaphor resolution process and the encoding formats are described.*

1. Introdução

Estamos desenvolvendo uma ferramenta multilíngüe para resolução de expressões anafóricas como as descrições definidas (...*os garotos*...), demonstrativas (...*estes garotos*...) e pronomes (...*eles*...). O principal objetivo é identificar automaticamente os antecedentes para estas expressões anafóricas.

Uma expressão é dita anafórica quando esta se refere a uma entidade que foi previamente referenciada no texto por outra expressão. Ambas as expressões são consideradas correferentes, e a primeira é vista como o antecedente da segunda. A informação sobre quais expressões são correferentes em um texto pode ser de grande utilidade em sistemas de recuperação e extração de informação, classificação de textos, sumarização de textos, entre outros.

O processo de resolução das anáforas é baseado em heurísticas geradas através de um estudo prévio do comportamento das expressões correferentes em textos escritos [Poesio and Vieira, 1998]. A ferramenta proposta tem como formato para entrada, manipulação e saída dos dados a linguagem de marcação XML.

Na seção seguinte, é detalhado o fenômeno da correferência e também os esquemas de anotação de corpus utilizados em nossos estudos. Na seção 3, são apresentados os principais aspectos da ferramenta proposta: o formato dos dados, as heurísticas que serão utilizadas e a arquitetura do sistema. Ao final, são apresentadas as conclusões deste trabalho.

2. Correferência

O fenômeno de correferência que ocorre na linguagem natural consiste em duas ou mais expressões de um texto se referirem a uma mesma entidade do discurso. Quando uma entidade é referenciada pela primeira vez em um texto, a expressão que a descreve é dita nova no discurso. Quando tal entidade é retomada no texto, a expressão que a descreve é dita anafórica, sendo considerado seu antecedente a expressão anterior correferente. As expressões anafóricas podem ser de diferentes tipos, como:

- ◇ pronominais: o termo anafórico é um pronome. Ex: *O menino leu o livro, mas **ele** não gostou.*
- ◇ definidas: compostas por um substantivo antecedido por um artigo definido. Ex: *Comprei uma casa. **A casa** fica longe daqui.*
- ◇ demonstrativas: compostas por um substantivo antecedido por um pronome demonstrativo. Ex: *Comprei uma casa. **Esta casa** será sempre minha.*

2.1. Esquemas XML para anotação de correferência

Em [Vieira et al., 2002b, Vieira et al., 2002a] são detalhados estudos feitos sobre textos escritos em Português e Francês, em que as expressões anafóricas foram marcadas manualmente no corpus, bem como seus antecedentes no texto. Este processo de marcação foi feito com a utilização da ferramenta MMAX (*Multi-Modal Annotation in XML*) [Müller and Strube, 2001b]. O MMAX requer que os dados estejam representados em XML, seguindo uma estrutura determinada. O corpus deve estar de acordo com o formato apresentado na Figura 1.

```
...
  <word id="word_49">milhares</word>
  <word id="word_50">de</word>
  <word id="word_51">refugiados</word>
...
```

Figura 1: Formato do corpus.

De acordo com as tendências para padronização da anotação de corpora apresentadas em [Ide and Romary, 2002], adotou-se o arquivo de palavras, como proposto em [Müller and Strube, 2001a], como o repositório básico do corpus, para o qual todas as demais anotações lingüísticas apontam. Chama-se marcação *stand-off* aquela feita separadamente do corpus, isto é, em que as marcações são armazenadas em outro repositório diferente daquele onde está o corpus. Tais marcações contêm apontadores para os elementos do corpus a que se referem. Com isso, a integridade do texto não é comprometida pela anotação e pode-se ter mais facilmente diversos níveis de anotação sobre um mesmo corpus.

O resultado da anotação de correferência é apresentado como mostra a Figura 2, onde cada expressão referencial analisada é representada por um elemento <markable>, cujo atributo *span* indica as palavras que a formam, e o atributo *pointer* indica o identificador do antecedente. Os demais atributos correspondem à informação que pode ser especificada pelo usuário para a marcação. Em nosso caso, marcação de correferência, é necessário especificar os novos atributos a serem incluídos nos markables para registrar a informação sobre correferência, como por exemplo, sua classificação (direta, indireta, etc.).

Em nossos experimentos, a marcação manual foi feita por mais de um anotador (lingüistas) para verificar a concordância entre suas marcações ao final. No entanto, em nosso primeiro experimento, verificamos que a concordância entre os anotadores foi prejudicada por algumas dificuldades no processo de marcação, pois toda a anotação devia ser feita em uma única tarefa: identificar as expressões anafóricas, identificar seu antecedente (expressão correferente) e finalmente classificar a coreferencia em cada caso. Concluímos, então, que eram muitas decisões a serem tomadas pelo anotador em um único passo.

Desta forma, para facilitar o trabalho dos anotadores e, conseqüentemente, melhorar sua concordância, decidimos dividir o processo de marcação em 4 passos:

1. selecionar as expressões anafóricas: esta tarefa é feita por apenas um anotador, já que não há discordância sobre quais são as expressões anafóricas do texto.
2. identificar o antecedente para cada expressão anafórica, se possível;
3. classificar os correferentes: se a expressão anafórica tem um antecedente e este é correferente, a relação deve ser classificada.
4. classificar os não correferentes: se a expressão anafórica não tem um antecedente ou se seu antecedente não é correferente, a relação deve ser classificada.

O MMAX não estava completamente adaptado a trabalhar com diversos passos para a marcação, ou seja, a ter o processo de marcação dividido, não garantindo a consistência entre a anotação feita em passos diferentes. Para tratar este problema, desenvolvemos uma nova maneira de especificar os esquemas de marcação do usuário no MMAX e a propusemos aos desenvolvedores do MMAX, que implementaram nossas idéias.

Primeiramente, os esquemas de anotação do MMAX não permitiam relacionar um atributo com outro, de forma que seus valores eram independentes. Assim, o usuário não podia restringir a marcação de um atributo dependendo do valor marcado previamente para outro atributo. Por causa disso, todas as combinações de valores de atributos eram disponibilizadas ao anotador, causando confusão.

O aprimoramento do MMAX consiste em descrever o esquema de marcação em XML de acordo com a DTD mostrada na Figura 3.Isso permite mostrar ao anotador apenas as opções coerentes com o que foi marcado nos passos anteriores, evitando erros. O atributo *next* restringe as opções que estarão disponíveis ao anotador de acordo com o valor associado no passo anterior ou no passo corrente. Além disso, os atributos marcados nos passos anteriores podem ser indicados como “read-only” nos próximos passos. A Figura 4 apresenta como exemplo o esquema de anotação para o passo 3. Neste exemplo, o atributo *form* foi marcado no passo 1 e o atributo *coreference* foi marcado no passo 2. Se este último atributo tiver o valor “coreferent”, então o anotador deve marcar o atributo *classification*.

Após dividir nossos experimentos em passos e utilizar os novos esquemas para o

```
...
  <markable id="markable_3"
    span="word_135..word_136"
    pointer="markable_8" np_form="demNP"
    classification="indirect"/>
...
```

Figura 2: Formato da marcação.

MMAX, a concordância entre as marcações dos anotadores aumentou, como esperado. Os resultados da anotação serão utilizados para validar os resultados de nossa ferramenta para resolução automática de correferência, que irá gerar a anotação na forma de elementos <markable>, seguindo os formatos do MMAX.

```
<!ELEMENT schema (level+)>
<!ELEMENT level (value+)>
<!ATTLIST level
  id ID \#REQUIRED
  attribute CDATA \#REQUIRED
  read_only (yes|no) "no"
>
<!ELEMENT value EMPTY>
<!ATTLIST value
  id ID \#REQUIRED
  name CDATA \#REQUIRED
  default (yes | no) \#IMPLIED
  next CDATA \#IMPLIED
>
```

Figura 3: DTD do esquema de marcação.

```
<annotationscheme>
  <level id="level_1" attribute="form">
    <value id="value_1" name="defNP" next="level_2"/>
    <value id="value_2" name="other_NP"/>
    <value id="value_3" name="part_of_sentence"/>
    <value id="value_4" name="sentence"/>
    <value id="value_5" name="other"/>
  </level>
  <level id="level_2" attribute="coreference">
    <value id="value_6" name="coreferent" next="level_3"/>
    <value id="value_7" name="non_coreferent"/>
    <value id="value_8" name="none"/>
  </level>
  <level id="level_3" attribute="classification">
    <value id="value_9" name="direct"/>
    <value id="value_10" name="indirect"/>
    <value id="value_11" name="other"/>
    <value id="value_12" name="none"/>
  </level>
</annotationscheme>
```

Figura 4: Exemplo de esquema.

3. Ferramenta de resolução automática de correferência

Com base no estudo de corpus referido na seção anterior, estamos desenvolvendo uma ferramenta capaz de identificar automaticamente os antecedentes das expressões anafóricas. A ferramenta apresenta os resultados da resolução de acordo com os formatos do MMAX. Nas etapas intermediárias do processo são utilizados diferentes esquemas de marcação, contendo os elementos <anaphor> e <antecedent>.

Em [Rossi et al., 2001], é apresentada uma ferramenta para resolução automática de correferência baseada nas heurísticas levantadas anteriormente. Esta ferramenta utilizava a linguagem PROLOG para armazenar os dados. Com base nesse trabalho anterior, estamos desenvolvendo uma nova ferramenta para atender as tendências atuais de anotação de corpora.

3.1. Formato dos dados

Para representação dos dados de entrada e saída da ferramenta proposta, adotamos a linguagem XML e respeitamos os formatos usados no MMAX. Além disso, informações adicionais são extraídas do corpus, baseadas na análise sintática feita pelo analisador sintático PALAVRAS para o português [Bick, 2000]. A saída desse analisador é convertida em um conjunto de arquivos XML: o arquivo de palavras (elementos <word>), compatível com o MMAX; um arquivo com as categorias morfo-sintáticas (POS - Part Of Speech) das palavras do corpus, como na Figura 5 (onde o elemento *n* indica um substantivo); e um arquivo com as estruturas sintáticas das sentenças, representadas por "chunks". Um <chunk> representa a estrutura interna da sentença e pode conter sub-chunks, como mostrado na Figura 6, onde o <chunk> pai 1 (atributo id =”chunk_1”) é um sintagma nominal (atributo form=”np”) e o <chunk> filho 2 é um substantivo (atributo form=”n”) e também o núcleo do sintagma nominal (atributo function=”h”).

```
...
  <word id=''word_51''>
    <n gender=''M'' number=''P''>
  </word>
...
```

Figura 5: Categorias morfo-sintáticas.

```
<chunk id="chunk_1" function="subj" form="np" span="word_49..word_51">
  <chunk id="chunk_2" function="h" form="n" span="word_51..word_51"/>
</chunk>
```

Figura 6: Chunks.

A quantidade de informação sintática utilizada pode variar de acordo com as heurísticas de resolução de anáforas utilizadas.

3.2. Heurísticas

A escolha do antecedente de uma anáfora é feita através de um conjunto de heurísticas, cada uma gerando um conjunto de candidatos a antecedente, que podem trazer diferentes resultados. Os candidatos utilizados nas heurísticas são os sintagmas nominais.

Neste trabalho iremos abordar a heurística mais simples de resolução de anáforas (anáforas diretas), para fins de ilustração do processo de resolução. Segundo Poesio em [Poesio and Vieira, 1998], as anáforas diretas são aquelas cujo núcleo do termo anafórico é igual ao de seu antecedente.

Tomemos como exemplo as sentenças "Três acidentes graves marcaram o final de semana. Os acidentes lotaram o hospital local". O núcleo "acidentes" da descrição definida na segunda sentença é igual ao núcleo do sujeito da primeira sentença, o que caracteriza uma relação de anáfora direta.

3.3. Extração dos termos anafóricos e dos candidatos a antecedente

De acordo com as heurísticas utilizadas, um conjunto de chunks será gerado para viabilizar o processo de resolução. Baseado na heurística descrita na seção 3.2, são gerados os chunks das descrições definidas (sintagmas nominais iniciados por artigo definido), ou seja, seleção dos termos anafóricos. Para seleção dos candidatos a antecedente, extrai-se todos os sintagmas nominais do texto.

3.4. Arquitetura

A ferramenta em desenvolvimento corresponde a um conjunto de 3 fases compostas por uma ou mais tarefas codificadas através de folhas de estilo (stylesheets) (Figure 7). Cada tarefa é conectada a outra através de "pipes", e filtram o fluxo de dados[1].

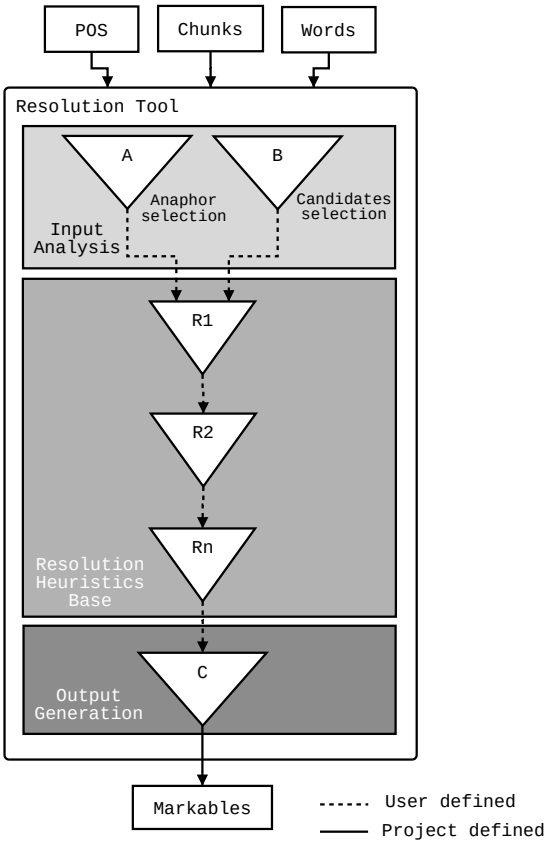

Figura 7: Arquitetura da ferramenta

A primeira fase, chamada *Input Analisys*, realiza duas tarefas (A e B) para extração de termos anafóricos e candidatos (descrito na seção 3.3). Os arquivos de palavras (words), de categorias morfo-sintáticas (POS) e de chunks são utilizados na geração de duas novas árvores de nodos que serão utilizadas no próximo passo. Uma é formada por nodos <anaphor>, cujo o atributo *span* corresponde a uma descrição definida do texto (Figura8(a)). Estes também têm nodos filhos que indicam informação extra (por exemplo, o *span* do núcleo da expressão anafórica). A outra árvore (Figura 8(b)) é formada por nodos <candidate> que indicam através de seu *span* os sintagmas nominais do texto (nos experimentos preliminares utilizamos todos os sintagmas nominais como candidatos).

A próxima fase, chamada *Resolution Heuristics Base*, é composta por um conjunto de regras (R1 à Rn), que implementam as heurísticas e filtram os nodos <anaphor> em busca dos seus antecedentes. Cada uma implementa uma heurística diferente[2], e a ordem em que são aplicadas pode ser modificada, ou seja, o usuário configura a ordem de aplicação das heurísticas visando alcançar melhores resultados.

[1]Estratégia baseada em Pipes & Filters Design Pattern, proposta por Gamma [Gamma et al., 1995].

[2]Com a finalidade de enfatizarmos a arquitetura global da ferramenta, o conjunto completo de heurísticas não é apresentado neste trabalho.

```
<anaphor span="word_5..word_11">
  <header>início</header>
</anaphor>
<anaphor span="word_8..word_11">
  <header>final</header>
</anaphor>
<anaphor span="word_13..word_14">
  <header>acidentes</header>
</anaphor>
<anaphor span="word_16..word_18">
  <header>hospital</header>
</anaphor>
```

(a)

```
<candidate span="word_1..word_3">
  <header>acidentes</header>
</candidate>
<candidate span="word_5..word_11">
  <header>início</header>
</candidate>
<candidate span="word_8..word_11">
  <header>final</header>
</candidate>
<candidate span="word_13..word_14">
  <header>acidentes</header>
</candidate>
<candidate span="word_16..word_18">
  <header>hospital</header>
</candidate>
```

(b)

```
<markables>
  <markable id="markable_1" pointer="" span="word_5..word_11" classification="discourse_new"/>
  <markable id="markable_2" pointer="" span="word_8..word_11" classification="discourse_new"/>
  <markable id="markable_3" pointer="markable_5" span="word_13..word_14" classification="direct"/>
  <markable id="markable_4" pointer="" span="word_16..word_18" classification="discourse_new"/>
  <markable id="markable_5" pointer="" span="word_1..word_3"/>
</markables>
```

(c)

Figura 8: Nodos <anaphor>(a), <candidate>(b), e <markable>(c)

Considerando, por exemplo, que a regra R2 implementa a resolução de anáforas diretas, para cada nodo <anaphor> ainda sem antecedente (com o atributo *pointer* vazio) busca-se um <candidate> com o mesmo núcleo, e *span* anterior ao nodo <anaphor>. Duas situações podem ocorrer com cada nodo <anaphor>:

1. Um candidato casa: o atributo *pointer* é preenchido com o *span* do candidato e o atributo *classification* com o valor correspondente ao da heurística em questão (por exemplo *direct*).
2. Nenhum candidato casa: este nodo <anaphor> será então analisado pela próxima regra.

Este processo se repete a cada regra e se encerra quando a última regra for executada.

A última fase, chamada *Output Generation*, gera os resultados no formato utilizado pelo MMAX para permitir a comparação dos resultados entre a marcação automática e a manual. Cada nodo <anaphor> gera um <markable> com *ID* único (por exemplo, ID="markable_1"). O atributo *pointer* recebe o *ID* do markable correspondente ao *span* do atributo *pointer* do nodo <anaphor>. Se este markable ainda não foi criado, um novo nodo <markable> é gerado utilizando este *span*. Nodos <anaphor> com *pointer* vazio geram nodos <markable> classificados como "discourse_new".

4. Conclusão

Apresentamos a arquitetura de uma ferramenta para resolução automática de correferência. Essa ferramenta processa automaticamente corpus analisado sintaticamente e representado na linguagem XML, seguindo recomendações de padrões em desenvolvimento para anotação de corpora (XCES [Ide and Romary, 2002], ISO TC37 SC4). A vantagem de se ter os dados codificados em XML é a possibilidade de utilizar-se as ferramentas já existentes para manipulação dos dados nessa linguagem.

Para avaliar-se a precisão dos resultados da ferramenta, os dados de saída serão comparados com os dados obtidos na marcação manual realizada em estudos prévios do corpus [Salmon-Alt and Vieira, 2002, Vieira et al., 2002b, Vieira et al., 2002a].

Como esse trabalho se insere no contexto do projeto de cooperação internacional Brasil-França COMMOn-REFs, a ferramenta foi concebida para tratar corpora multilíngüe, ou seja, que possa ser utilizada para a resolução de correferência em outras línguas além do português, visto que as heurísticas podem ser acopladas a ferramenta incrementalmente através de folhas de estilo independentes. Em estudos de corpus prévio sobre o francês [Salmon-Alt and Vieira, 2002, Vieira et al., 2002b, Vieira et al., 2002a], também foi realizada a anotação manual de correferência, o que poderá ser utilizado para avaliação da ferramenta em relação a esta língua.

Agradecimentos

Agracemos ao CNPq e INRIA, pelo financiamento do projeto COMMOn-REFs; ao Eckhard Bick, por sua ajuda no uso do analisador PALAVRAS; ao Paulo Quaresma que desenvolveu o parser XML para o analisador PALAVRAS; ao Christoph Müller e Michael Strube, por proverem a base de nossos esquemas de anotação e acolherem nossas sugestões em relação ao MMAX; e a Susanne Salmon-Alt, pela cooperação no desenvolvimento dos estudos de corpora.

Referências

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Århus University, Århus.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional Computing Series. Addison-Wesley Publishing Company, New York.

Ide, N. and Romary, L. (2002). Standards for language resources. In *Proceedings of the LREC 2002*, pages 839–844, Las Palmas de Gran Canaria.

Müller, C. and Strube, M. (2001a). Annotating anaphoric and bridging expressions with MMAX. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, pages 90–95, Aalborg, Denmark.

Müller, C. and Strube, M. (2001b). MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the IJCAI 2001*, pages 45–50, Seattle.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Rossi, D., Pinheiro, C., Feier, N., and Vieira, R. (2001). Resolução de correferência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica*, 1(2).

Salmon-Alt, S. and Vieira, R. (2002). Nominal expressions in multilingual corpora: Definites and demonstratives. In *Proceedings of the LREC 2002*, Las Palmas de Gran Canaria.

Vieira, R., Salmon-Alt, S., Gasperin, C., Schang, E., and Othero, G. (2002a). Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the DAARC 2002*, Estoril.

Vieira, R., Salmon-Alt, S., and Schang, E. (2002b). Multilingual corpora annotation for processing definite descriptions. In *Proceedings of the PorTAL 2002*, Faro.