Uma forma simples de resolver concordância com gramáticas livres-de-contexto modificadas

Leonardo Brito

www.cin.ufpe.br/~lmpb

maio - agosto 2011

Problema: uso de gramática livre-de-contexto para gerar cadeias de linguagens naturais que sejam sintaticamente corretas.

Numa GLC pura não há o conceito de concordância entre palavras duma frase, pois os terminais da gramática não contém informações adicionais além de seus próprios símbolos. Numa GLC do português, o terminal "gato" pode ser encontrado do lado direito da regra referente à variável SUBSTANTIVO, mas a princípio não se sabe mais nada sobre o terminal. Para que se desenvolva uma GLC com concordância sintática adequada, é necessário armazenar metainformações sobre os terminais, e.g. "gato" tem número singular e gênero masculino.

Uma maneira de garantir a concordância seria aumentar o número de regras de forma que houvesse variáveis distintas para cada conjugação possível. Essa abordagem seria um tanto dispendiosa, ainda mais em línguas como o português, onde abundam conjugações e flexões.

Uma solução mais simples e genérica é atribuir metainformações às variáveis e terminais de modo a restringir as opções de terminais que podem ser escolhidos. Assim abre-se espaço para vários tipos de restrição, não só referentes à sintaxe (e.g. restrições semânticas). Vamos seguir nesta direção.

Proposta: agregar informações adicionais aos terminais duma GLC que representa uma linguagem natural afim de garantir a corretude sintática das cadeias geradas pela gramática.

Definição: uma conjugação é uma n-upla {x1, x2, ..., xn} onde xi ϵ Ci, e C é uma m-upla {c1, c2, ..., cn} onde ci é alguma restrição desejada à linguagem. e.g, uma possível conjugação é {feminino, plural}, onde x1 = gênero e x2 = número.

Definição: uma anáfora é a 4-upla {v, α, β, c}, onde v pertence ao conjunto das variáveis da GLC, α ϵ {FORTE, FRACO, NEUTRO}, β pertence ao conjunto dos identificadores de anáfora e c pertence ao conjunto de conjugações possíveis.

Definição: definimos a GLC estendida como contando com a seguinte modificação às GLCs convencionais:

- Cada variável do lado esquerdo da regra pode receber uma anáfora como parâmetro (α = FRACO) ou retornar uma anáfora definida do lado direito da regra (α = FORTE)
- Cada variável do lado direito da regra pode receber uma anáfora como parâmtro (α = FRACO) ou definir uma anáfora (α = FORTE)
- Variáveis sem metainformação sintática tem α = NEUTRO

No caso das variáveis do lado direito da regra, escrevemos na seguinte notação:

<NOME_VAR x > para uma variável NOME_VAR que define uma anáfora x

<NOME_VAR(x)> para uma variável NOME_VAR que recebe como parâmetro uma anáfora x.

O mesmo se passa para as variáveis do lado esquerdo, com a diferença que uma variável <NOME_VAR x > retornará a anáfora x definida por alguma variável do lado direito da regra.

Como exemplo, vamos supor a seguinte gramática livre de contexto, subconjunto da gramática da língua portuguesa:

<ORACAO> → <SUJEITO> <PREDICADO>
<SUJEITO> → <SUJ_SIMPLES> <COMPLEMENTO_NOMINAL>
<SUJ_SIMPLES> → <ARTIGO> <SUBSTANTIVO>
<COMPLEMENTO_NOMINAL> → “d “ <SUJEITO> | <ADJETIVO > “d “ <SUJEITO> | <ADJETIVO > | ε
<PREDICADO> → <VERBO> “d” <SUJEITO>
<ARTIGO> → “a” | “o” | “as” | “os”
<ADJETIVO> → “rápido” | “fortes” | “cheirosas” | ”grande”
<SUBSTANTIVO> → “menina” | “flores” | “cavalo”
<VERBO> → “gosta” | “cheiram” | “comem” | “comeu”

Fig. 1: gramática G

Reescrevendo a gramática G de acordo com a GLC modificada proposta, temos:

<ORACAO> → <SUJEITO x> <PREDICADO(x)>
<SUJEITO x> → <SUJ_SIMPLES x> <COMPLEMENTO_NOMINAL(x)>
<SUJ_SIMPLES> → <ARTIGO(x)> <SUBSTANTIVO x> RETURN(x)
<COMPLEMENTO_NOMINAL(x)> → “d “ <SUJEITO> | <ADJETIVO(x)> “d “ <SUJEITO> | <ADJETIVO(x)> | ε
<PREDICADO(x)> → <VERBO(x)> <SUJEITO>
<ARTIGO> → “a” | “o” | “as” | “os”
<ADJETIVO> → “rápido” | “fortes” | “cheirosas” | ”grande”
<SUBSTANTIVO> → “menina” | “flores” | “cavalo”
<VERBO> → “gosta” | “cheiram” | “comem” | “comeu”

Fig. 2: gramática G’

Algoritmo de resolução: numa GLC comum, expande-se as variáveis mais à esquerda ou mais à direita até que só haja terminais. Na GLC modificada isso não é mais possível por conta das restrições e dependências geradas pelas anáforas. O seguinte algoritmo basta para resolver todas as regras da GLC modificada:

1. Da esquerda para a direita, visite todas as variáveis do lado direito da regra:
    a. Se a variável v tiver anáfora com α = FORTE faça:
        i. Resolva a variável v e atualize sua anáfora.
        ii. Se a variável do lado esquerdo da regra tem anáfora com α = FORTE e o mesmo identificador (β) da variável v, atualize a anáfora da variável do lado esquerdo
    b. Se a variável v tiver anáfora com α = NEUTRO, resolva-a.
2. Da esquerda para a direita, visite novamente todas as variáveis do lado direito da regra:
    a. Se a variável v tiver anáfora com α = FRACO faça:
        i. Se a variável do lado esquerdo da regra tem anáfora com α = FORTE e o mesmo identificador (β) da variável v, passe a anáfora como parâmetro e resolva a variável v
        ii. Senão, busque nas variáveis do lado direito da regra uma variável u com anáfora com α = FORTE e o mesmo identificador (β) da variável v. Passe esta anáfora como parâmetro e resolva a variável v

Exemplo de funcionamento: afim de explorar o funcionamento de G', vamos derivar o seguinte exemplo: "*o cavalo rápido da menina comeu as flores*", começando sempre pelas variáveis mais à esquerda.

Começamos pela variável inicial:

<ORACAO> → <SUJEITO x> <PREDICADO(x)>

O trecho <SUJEITO x> significa: anafora(sujeito) = x , e o trecho <PREDICADO(s)> significa: anafora(predicado) = x , ou seja, a resolução da variável PREDICADO em terminais dependerá da sintaxe fornecida pela variável SUJEITO.

Destrinchando do lado esquerdo, começando por <SUJEITO x>, temos a seguinte regra:

<SUJEITO x> → <SUJ_SIMPLES x> <COMPLEMENTO_NOMINAL(x)>

A variável do lado esquerdo, SUJEITO, tem uma anáfora forte com identificador "x". Isso significa que esta regra irá retornar a anáfora com identificador "x" definida por alguma variável com anáfora forte do lado direito da regra, neste caso SUJ_SIMPLES.

<SUJ_SIMPLES x> → <ARTIGO(x)> <SUBSTANTIVO x>

Aqui acontece o mesmo. SUJ_SIMPLES retornará a SUJEITO a anáfora definida em SUBSTANTIVO.

Numa GLC pura iríamos primeiro tentar resolver a variável ARTIGO, mas nesta gramática a variável ARTIGO tem uma dependência sintática em relação à variável SUBSTANTIVO. Portanto, a primeira variável mais à esquerda que podemos resolver é a variável SUBSTANTIVO, que fornecerá a conjugação x tanto a ARTIGO quanto a SUJ_SIMPLES (que, apesar de ter uma anáfora forte, é variável do lado esquerdo e portanto necessita ter sua anáfora definida fora da regra).

<SUBSTANTIVO> → "menina" | "flores" | "cavalo"

Escolheremos a palavra "cavalo" de modo que sigamos o exemplo proposto.

No caso, a palavra “cavalo” tem a seguinte conjugação: sintaxe(“cavalo”) = { singular, masculino }. A anáfora x da variável SUBSTANTIVO na regra anterior será, portanto:

{ SUBSTANTIVO, FORTE, x, {singular, masculino} }.

<SUJ_SIMPLES> → <ARTIGO(x)> “cavalo”

Vamos resolver ARTIGO:

<ARTIGO> → “a” | “o” | “as” | “os”

A variável terá a restrição de seguir a conjugação fornecida por x, i.e. terá de ser da forma {singular, masculino}. Somente o terminal “o” respeita essa restrição, portanto ele é o escolhido.

A variável SUJ_SIMPLES está completamente resolvida:

<SUJ_SIMPLES> → “o cavalo”

Temos agora:

<SUJEITO> → “o cavalo” <COMPLEMENTO_NOMINAL(x)>

COMPLEMENTO_NOMINAL se resolve da maneira esperada:

<COMPLEMENTO_NOMINAL(x)> → “d “ <SUJEITO x> | <ADJETIVO(x)> “d “ <SUJEITO> | <ADJETIVO(x)> | ε

Para fins de exemplo, escolhemos a regra → ADJETIVO “d” SUJEITO.

<ADJETIVO> → “rápido” | “fortes” | “cheirosas” | ”grande”

Os terminais “rápido” e “grande” se adequam à restrição ({singular, masculino}). Para fins de exemplo escolhemos “rápido”.

<COMPLEMENTO_NOMINAL(x)> → “rápido d” <SUJEITO>

Destrinchando a variável SUJEITO:

<SUJEITO> → <SUJ_SIMPLES x> <COMPLEMENTO_NOMINAL(x)>

Lembrando que queremos obter a frase “o cavalo rápido da menina comeu as flores”, devemos fazer as seguintes escolhas:

<SUJ_SIMPLES> → <ARTIGO(x)> <SUBSTANTIVO x>

<SUBSTANTIVO> → “menina” | “flores” | “cavalo”

<SUJ_SIMPLES> → <ARTIGO(x)> “menina”

<ARTIGO> → “a” | “o” | “as” | “os”

<SUJ_SIMPLES> → “a menina”

Lembrando que tínhamos:

<COMPLEMENTO_NOMINAL(x)> → “rápido d” <SUJEITO x>

Temos então:

<COMPLEMENTO_NOMINAL(x)> → “rápido da menina”

Temos a variável SUJEITO completa:

<SUJEITO> → “o cavalo rápido da menina” RETURN(x)

Na oração, temos:

<ORACAO> → “o cavalo rápido da menina” <PREDICADO(x)>

Nota-se que o predicado terá de concordar com o sujeito, mais especificamente com o núcleo do sujeito, “cavalo”, i.e. receberá como parâmetro a anáfora { SUBSTANTIVO, FORTE, x, {singular, masculino} }.

<PREDICADO(x) → <VERBO(x)> <SUJEITO x>

<SUJEITO> → <SUJ_SIMPLES x> <COMPLEMENTO_NOMINAL(x)>

<SUJ_SIMPLES> → <ARTIGO(s10)> <SUBSTANTIVO s10>

<SUBSTANTIVO> → “menina” | “flores” | “cavalo”

<SUJ_SIMPLES> → <ARTIGO(x)> “flores”

<ARTIGO> → “a” | “o” | “as” | “os”

<SUJ_SIMPLES> → “as flores”

<SUJEITO> → “as flores” <COMPLEMENTO_NOMINAL(x)>

<COMPLEMENTO_NOMINAL(x)> → “d “ <SUJEITO> | <ADJETIVO(x)> “d “ <SUJEITO> | <ADJETIVO(x)> | ε

Obviamente, escolhemos a regra da cadeia vazia. Temos o objeto do predicado completo:

<SUJEITO> → “as flores”

Temos o seguinte predicado:

<PREDICADO(x) → <VERBO(x)> “as flores”

Resolvendo o verbo:

<VERBO> → “gosta” | “cheiram” | “comem” | “comeu”

Escolhemos “comeu”.

<PREDICADO(in) → “comeu as flores”

Enfim, a frase está completamente derivada:

<ORACAO> → “o cavalo rápido da menina comeu as flores”

Exemplos: alguns exemplos de frases geradas utilizando-se a gramática G' proposta (já implementada em Java, exemplo disponível em cin.ufpe.br/~lmpb/newgen.php):

A menina comeu as flores cheirosas do cavalo.

Os cavalos duns cavalos cheiram uns cavalos fortes.

O cavalo rápido da menina comeu uma menina duns cavalos.

Os cavalos duma menina comem um cavalo rápido duma menina.

Os cavalos fortes das flores cheiram um cavalo grande da menina.

Umas flores fortes cheiram uns cavalos.

Um cavalo grande duns cavalos comeu a menina grande dos cavalos.

O cavalo duma menina comeu umas flores fortes da menina.

As flores fortes cheiram umas flores cheirosas duns cavalos.

Uma menina grande do cavalo comeu uns cavalos fortes duma menina.

A menina grande dum cavalo gosta o cavalo rápido.

Uns cavalos fortes da menina comem os cavalos fortes dumas flores.

Conclusão: pode-se utilizar a metainformação introduzida na GLC modificada de várias outras formas além da concordância sintática proposta: pode-se exprimir alguma forma de semântica (e.g. em vez de "conjugações" armazenadas na anáfora, teríamos grupos ontológicos), ponto de vista (privilegiar uso de palavras consideradas positivas ou negativas de acordo com o tom que se quer obter do texto gerado) ou mesmo uma concordância mais refinada (especificando pessoa, tempo e modo verbal).