Avaliação de *Centering* em Resolução Pronominal da Língua Portuguesa

Ana Margarida Aires[1], Jorge Cesar B. Coelho[2], Sandra Collovini[2], Paulo Quaresma[1] e Renata Vieira[2]

[1] Universidade de Évora, Departamento de Informática, Évora, Portugal
{aaires,pq}@di.uevora.pt}
[2] Universidade do Vale dos Sinos, Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, São Leopoldo, Brasil
{cesar,sandrac,renata}@exatas.unisinos.br

Abstract. This paper presents a corpus study and an evaluation of centering for pronoun resolution in Portuguese texts. Centering is a system of rules and restrictions that govern the relations between referring expressions. The corpus study is to verify specific features related to pronominal anaphora. These features serve as background knowledge for the development of a system for pronouns anaphoric resolution. The final goal of this study is to integrate the pronoun resolution into information retrieval systems.

Key words: pronouns, portuguese anaphora resolution, information retrieval.

Resumo. Este artigo apresenta um estudo de corpus e uma avaliação do método de *centering* para resolução de anáforas pronominais em textos da língua portuguesa. *Centering* é um sistema de regras e restrições que governam as relações entre expressões referenciais. O estudo de corpus tem por objetivo verificar características específicas relacionadas à anaforicidade pronominal no domínio jurídico. Estas características servem como conhecimento inicial para o desenvolvimento de uma aplicação para resolução de anáforas pronominais. O objetivo final desse estudo é investigar se podemos ter benefícios com a integração da resolução automática de anáforas pronominais aos sistemas de recuperação de informação.

Palavras-chave: pronomes, resolução de anáforas para o português, recuperação de informação.

1 Introdução

Este artigo apresenta um estudo de corpus e uma avaliação do método de *centering* [1], para resolução de anáforas pronominais em textos da língua portuguesa. O objetivo final deste estudo é investigar se podemos ter benefícios com a integração da resolução automática de anáforas pronominais aos sistemas de recuperação de informação

do Projeto PGR - Acesso Seletivo aos Pareceres da Procuradoria Geral da República de Portugal[1].

O artigo está organizado como segue: A seção 2 oferece uma visão geral sobre a resolução pronominal e demonstra porque a resolução de anáforas pronominais tem potencial para tornar a pesquisa de informação mais eficaz. A seção 3 apresenta a anotação sintática automática e anotação manual de co-referência pronominal do corpus utilizado na avaliação de *centering*. A seção 4 descreve o algoritmo avaliado. Os resultados obtidos são apresentados na seção 5. A seção 6 reservamos para as considerações finais.

2 Resolução Pronominal

O ato de referência consiste em utilizar formas lingüísticas para evocar entidades que pertencem a universos reais ou fictícios. O processo de referenciação se constrói de maneira progressiva. Usamos constantemente expressões que retomam outras expressões referidas no próprio texto na apresentação do fluxo de idéias.

Dentre as operações de retomada existentes, a que nos interessará aqui particularmente é a anáfora pronominal. Qualifica-se de anafórico um segmento do texto (na maioria das vezes um pronome ou um sintagma definido[2] ou demonstrativo[3]) cuja interpretação remete a outro segmento anteriormente presente no discurso.

Assim definida, a anáfora representa um fenômeno de dependência interpretativa entre duas unidades, em que a segunda (*elas* ou *esse artigo*, no exemplo (1) abaixo) não pode receber um sentido referencial específico, completo, sem ter sido posta em conexão com a primeira (*todas as pessoas* ou *o artigo 27º*, no exemplo (1) abaixo). Quando a segunda unidade é um pronome, denomina-se que ocorre uma anáfora pronominal.

(1) *O artigo 27º reconhece a todas as pessoas o direito de acesso às informações sobre elas registradas. Esse artigo também prescreve sobre o direito à informação e acesso.*

Os pronomes representam a classe mais genérica dos nomes. O próprio conceito de pronome carrega em si a idéia de conexão conceitual. Os pronomes apresentam várias subdivisões (substantivos, adjetivos, pessoais, retos, oblíquos, possessivos, demonstrativos, relativos etc.). Nosso estudo concentrou-se nos pronomes pessoais de terceira pessoa: *ele*(s), *ela*(s), *o*(s)[4], *a*(s) e *lhe*(s).

---

[1] Disponível em: http://www.pgr.pt.

[2] Grupo de palavras que inicia por artigo definido e possui núcleo nome (e.g. *o parecer*, *a resolução final*).

[3] Grupo de palavras que inicia por pronome demonstrativo e possui núcleo nome (e.g. *essa resolução*).

[4] Incluímos também as formas especiais dos pronomes *o*(s) e *a*(s): *lo*(s) e *la*(s) quando ligados a terminações verbais *-r, -s* ou *-z*; e *no* e *na* quando ligados a terminações verbais *-am, -em, -ão* ou *-õe*.

Como a expressão anafórica retoma um termo anterior, a identificação automática do antecedente efetua-se com certa complexidade, pois freqüentemente mais de um candidato (grupo nominal) está disponível como possível antecedente. Quando se encontra um só grupo nominal anterior de mesmo número e gênero, a identificação do antecedente de um pronome está assegurada. No entanto, quando vários grupos nominais apresentam mesmo número e gênero, outros requisitos têm de ser incorporados.

De acordo com a teoria de *centering* (detalhada na seção 4), as possibilidades podem ser resolvidas se os grupos nominais candidatos a antecedentes assumem posições diferenciadas na sentença. Em cada sentença, existe um grupo nominal dominante (em foco), que pode ser o grupo mais próximo ou o sujeito da sentença. No exemplo (2), apesar da identidade de gênero (feminino) e número (singular), o sujeito da primeira sentença (*a advogada*) se impõe como antecedente do pronome *ela*, do mesmo modo que o complemento verbal *a ré* é antecedente da primeira ocorrência do pronome *a*. Portanto, a dominância de um grupo nominal permite eleger o antecedente privilegiado de um pronome no caso de vários candidatos.

(2) *O promotor iniciou a seção com perguntas agressivas. Primeiramente, a advogada acalmou a ré. Depois ela a recomendou responder a cada uma evasivamente.*

A próxima seção apresenta o estudo de corpus realizado.

3 Descrição do Corpus

O estudo apresentado aqui foi realizado em um corpus constituído por 16 Pareceres da Procuradoria Geral da República de Portugal. O corpus foi anotado automaticamente com informações morfossintáticas. Para isso, foi utilizado um analisador sintático, PALAVRAS [2]. Essa ferramenta efetua, mesmo em sentenças incompletas ou incorretas, a análise morfossintática. Uma segunda ferramenta foi empregada, Palavras Xtractor [3]. Esta ferramenta converte a saída do analisador sintático em três arquivos XML[5] (*EXtensible Markup Language*): i) *words* (com uma lista de palavras do texto e seus respectivos identificadores), ii) *pos* (com as informações morfossintáticas das palavras do texto) e iii) *chunks* (com a estrutura do texto).

Os pronomes pessoais de terceira pessoa foram anotados manualmente com o auxílio de uma ferramenta de anotação de discurso, MMAX (*Multi-Modal Annotation in XML*) [4]. Essa ferramenta utiliza o arquivo *words* gerado pela ferramenta Palavras Xtractor. O resultado do processo de anotação no MMAX é um arquivo XML.

Nesse sentido, foram analisadas em detalhe características (sintáticas, semânticas e discursivas) relacionadas às ocorrências pronominais do corpus. No corpus foram identificadas 302 ocorrências pronominais e seus antecedentes mais próximos foram apontados. As características estudadas foram: i) tipo de sintagma do antecedente pronominal; ii) tamanho das cadeias de co-referência dos pronomes; iii) carga significativa do antecedente pronominal em relação à cadeia de co-referência e iv) janela de retomada pronominal. Observamos que o tipo de antecedente predominante foi descrição definida (69% dos casos) e o tamanho médio das cadeias foi de 17 (maior 44 e

[5] Disponível em: http://www.w3.org/XML.

menor 8)[6]. Ao analisar a carga significativa dos antecedentes pronominais, verificamos dois aspectos: se o antecedente era plenamente definido (e.g. *o regime de separação de bens*) ou uma retomada incompleta (e.g. *esse regime*). Nessa análise averiguou-se que, mesmo com predomínio do tipo descrição definida, houve equilíbrio entre os dois aspectos (43% dos casos plenamente definidos e 57% genericamente marcados). Quanto à janela pronominal (i.e. distância entre o pronome e seu antecedente), detectamos que, predominantemente (81% dos casos), na mesma sentença do pronome encontrava-se o seu antecedente. É necessário ponderar que, no domínio jurídico, as sentenças apresentam uma extensão considerável (i.e. compostas geralmente por mais de três orações). Tanto a anotação manual quanto o estudo de características têm em vista a avaliação do algoritmo (apresentada na seção 5).

4 Descrição do Algoritmo Avaliado

O algoritmo desenvolvido e avaliado nesse trabalho é baseado em [1], uma extensão de [5]. Ambos os métodos, durante o processamento do texto, determinam o foco central do discurso a fim de localizar as entidades a que cada pronome faz correspondência. O objetivo é atingindo através das relações entre as entidades de sentenças distintas. Existem diferentes tipos de transações entre as sentenças. No primeiro estudo consideravam-se três, *continuing*, *retaining* e *shifting*, sendo a estas, acrescentado mais um estado *shifting-1*, na metodologia em que a aplicação foi baseada, para classificar situações mais ambíguas.

Mais especificamente, *centering* é um sistema de regras e restrições que governam as relações entre o assunto que um texto trata e as escolhas lingüísticas feitas pelo(s) seu(s) autor(es) para exprimir o fluxo de idéias.

Considerando que Un representa uma sentença (*Utterance*), vejamos as estruturas necessárias ao modelo *centering*. Um segmento de discurso consiste numa seqüência de sentenças, U1...Un. A cada sentença Un está associado um Cf(Un) (*foward looking center*), constituído pelas entidades referidas em Un (pronomes ou grupos nominais), possíveis de serem foco na próxima sentença. Convenciona-se que o primeiro elemento desta lista é o Cp(Un) (*preferred center*), i.e., o elemento que tem maior possibilidade de ser um elemento central na próxima sentença. A última estrutura, denomina-se Cb (*backward looking center*). O Cb corresponde ao Cf da sentença anterior, especificamente ao primeiro elemento desta lista. Refere-se, portanto, a uma entidade que já tinha sido introduzida no texto e que continua presente nele. Veja o exemplo (3):

(3) *O réu conduzia um Alfa Romeu. Ele trafegava acima da velocidade permitida. O radar registrou que ele estava a 183 km/h.*

---

[6] O tamanho da cadeia de co-referência corresponde ao número de vezes que uma entidade é evocada no discurso, e.g., "*As associações públicas poderão requerer o direito de posse* (...) *essas associações, contudo, não poderão tomar posse da área em período processual, a elas será concedida posse, 22 dias a contar a data da resolução final*". Nesse exemplo, o tamanho da cadeia é 3.

O réu conduzia um Alfa Romeu.
Cf = {réu, Alfa Romeu}
Cb = { }
Cp = {réu}

Ele trafegava acima da velocidade permitida.
Cf = {Ele}
Cb = {réu}
Cp = {Ele}

O radar registrou que ele estava a 183 km/h.
Cf = {radar, ele}
Cb = {Ele}
Cp = {ele}

As transações já introduzidas são determinadas dependendo de dois fatores: o foco do texto numa sentença Cb(Un) é ou não idêntico ao da sentença anterior Cb(Un-1) e esta mesma entidade é ou não igual ao Cp(Un) – Figura 1. Estas quatro transações representam a forma como as sentenças se relacionam num texto coerente. A coerência é uma noção importante nesta metodologia. Ela compreende que todas as proposições a referir sobre uma entidade são feitas consecutivamente sem que se introduza uma nova entidade no discurso (*continuing*). A coerência mantém-se mesmo que se faça referência a uma nova entidade, se conservado o foco do discurso da sentença anterior (*retaining*). Sempre que se muda o foco do discurso, se pretende mantê-lo nas próximas sentenças (*shifting-1*) ou sempre que se muda de foco sem que se tenha intenção deste continuar a sê-lo na próxima sentença (*shifting*), a coerência permanece.

	Cb(Un) = Cb(Un-1)	Cb(Un) # Cb(Un-1)
Cb(Un) = Cp(Un)	*Continuing*	*Shifting -1*
Cb(Un) # Cp(Un-1)	*Retaining*	*Shifting*

Fig.1. Fatores que determinam as transações do modelo *centering*.

Além das transações, existem algumas restrições e regras. Estas permitem a correspondência correta entre os pronomes e as entidades que estes referem. As restrições são três: Existe apenas um valor para Cb(Un), i.e., para cada sentença existe apenas um elemento central. Todas as entidades contempladas em Cf(Un) são referidas em Un. E Cb(Un) é o elemento mais importante de Cf(Un-1). As regras são duas. A primeira diz que se algum elemento de Cf(Un-1) é pronome então em Cb(Un) também o será. A segunda, mostra como é feita a escolha do pronome e do antecedente após a sua classificação. A ordem de preferência da resolução é: *continuing* > *retaining* > *shifting -1* > *shifting*. Dada a abordagem, passamos a implementação.

4.1 Implementação

Numa primeira fase, o objetivo é identificar todas as anáforas pronominais e seus possíveis candidatos. Nesta etapa, são gerados dois arquivos XML: um com as anáforas e outro com as suas possíveis soluções. Deste modo, todos os elementos cuja categoria gramatical é igual a pronome pessoal são identificados como anáforas e todos os pronomes pessoais são entendidos como candidatos à solução. De acordo com as definições das estruturas e regras apresentadas, são criadas as estruturas Cb e Cf.

A criação de elementos Cf – constituídos, nesse momento, por um par (anáfora, possível solução) – passa por uma checagem de gênero e número (pois o antecedente deve concordar em gênero e número com o pronome). Convencionou-se a janela de retomada para quatro sentenças anteriores. Ao aplicar-se a última folha de estilo desta fase, é construída uma nova estrutura a que se chamou Item. Esta é constituída por um valor de Cb, que é único para cada sentença, e pelas combinações de cada par anáfora e sua solução para uma mesma sentença. Desta forma, cada Item tem um valor de Cb e um valor Cf com tantas anáforas quanto as presentes na sentença que está a ser analisada.

4.1.1 Análise/Filtragem

Nesta segunda fase, o objetivo é eliminar Item's que à partida não constituem uma solução correta. Essencialmente ocorrem duas situações. Primeira, quando o elemento que constitui o Cb(Un) não é o mais relevante de Cf(Un-1). Durante a implementação, na construção do Cb, tem-se esta informação em conta e constrói-se somente os Cb's que não serão eliminados numa próxima fase. A segunda situação diz que a mesma anáfora não pode apontar para dois grupos nominais diferentes. Logo, ao analisar cada Cf(Un), sempre que há repetição de uma anáfora no mesmo Item, exige-se repetição da possível solução, caso contrário o Item é eliminado.

4.1.2 Classificação e Seleção

Nesta última fase, cada Item recebe uma classificação. As estruturas são analisadas e comparadas entre si, e, de acordo com as definições da Figura 1, atribuem-se as classificações. Antes que a seleção seja feita, o XML intermédio, nessa fase do algoritmo, sofre ainda ação de dois filtros, a fim de evitar mais comparações na aplicação da última folha de estilo. Neste momento são eliminados os Item's vazios assim como os Item's em que a mesma anáfora é contemplada mais do que uma vez. Isto porque estes últimos conjuntos resultam das combinações feitas anteriormente, porém representam repetições desnecessárias.

Por fim é aplicada a última folha de estilo que verifica a classificação de cada Item e escolhe segundo esta ordem decrescente de preferência: *continuing* > *retaining* > *shifting*-1 > *shifting*.

Na próxima seção são apresentados os resultados.

5 Resultados

Os resultados apresentados são o produto de testes sobre o corpus (seção 3). Para que se tenha uma visão geral, apresentam-se as tabelas 1 e 2 que dizem respeito a uma análise conjunta de todos os arquivos (micro-média) e a uma média das percentagens obtidas em cada arquivo (macro-média), respectivamente.

Tabela 1. Análise conjunta de todos os arquivos.

Total de Anáforas	302
Anáforas mal resolvidas	128
Anáforas mal identificadas	20
Anáforas bem resolvidadas	154
Percentagem de sucesso	51.00%

Tabela 2. Percentagem de sucesso obtida em cada parecer jurídico.

2.txt	3.txt	4.txt	7.txt	10.txt	11.txt	12.txt	13.txt	14.txt	30.txt	Média
41%	75%	45%	74%	40%	69%	43%	40%	42%	39%	51.00%

Os resultados aqui apresentados mostram um percentual de acerto abaixo dos resultados reportados em aplicações dessa mesma teoria em corpus da língua inglesa. Em [6], um sistema baseado em *centering*, avaliado em um corpus de artigos de jornal e textos de ficção, apresenta resultados acima de 75%. *Centering* também foi testado para diálogos em espanhol, atingindo resultados acima de 75% [7]. Em [8], um estudo para a resolução pronominal do português é apresentado. Nesse estudo, a teoria de foco de Sidner [9] é integrada ao DRT [10]. Os testes apresentam bons resultados (acima de 77%) mas são limitados para casos específicos como elipses e anáforas em cláusulas relativas.

6 Considerações finais

A resolução de anáfora pode modificar o peso de importância de termos em documentos para pesquisa de informação. Esse projeto tem por objetivo verificar se isso vem a ser um passo relevante no processo de indexação de documentos no projeto PGR. O estudo do corpus nos permitiu observar que os pronomes pessoais de terceira pessoa retomam informações com bastante freqüência, a média de cadeias de co-referência em que os pronomes estão envolvidos (17 retomadas) indica que o seu tratamento, se considerado nas medidas de relevância, possa vir a modificar os pesos identificados para os termos. Para isso precisamos desenvolver algoritmos eficazes na resolução pronominal. Esse trabalho apresenta uma primeira avaliação da teoria de *centering* aplicada à resolução pronominal em um corpus jurídico. Como trabalhos futuros, uma análise detalhada dos resultados será realizada. Outros algoritmos de resolução pro-

nominal – tais como Lappin and Leass [11] e Hobbs [12] – serão avaliados e posteriormente experimentos com aprendizado de máquina serão realizados.

Agradecimentos

Este trabalho foi parcialmente financiado pelas seguintes agências: CAPES, CNPq, FAPERGS e GRICES.

Referências

1. Brennan, Susan E., Friedman, Marilyn W., Pollard, Carl J.: A centering approach to pronouns. Proceedings of the 25th conference on Association for Computational Linguistics, Stanford, California (1987) 155-162
2. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese. Constraint Grammar Framework. PH. D. thesis, Aarhus University Press (2000)
3. Gasperin, C., Vieira, R., Goulart, R. and Quaresma, P.: Extracting XML Syntactic Chunks from Portuguese Corpora. Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages. Batz-sur-Mer, France (2003)
4. Müller, C. and Strube, M.: MMAX: A tool for the annotation of multi-modal corpora. Proceedings of the IJCAI, Washington, USA (2001)
5. Grosz, B. J., Joshi, A.K., Weinstein, S.: Centering: a framework for modeling the local coherence of discourse. Computational Linguistics, 12 (3) (1995) 203-226
6. Tetreault, Joel R.: A corpus-based evaluation of centering and pronoun resolution. Computational Linguistics, Vol. 7 Issue4. Special issue on computational anaphora resolution (2001) 507-520
7. Martínez-Barco, P. et al: Evaluation of Pronoun Resolution Algorithm for Spanish Dialogues. Lecture Notes in Proceedings of the Venezia per il Trattamento Automatico delle Lingue VEXTAL'99. Venice, Italy (1999)
8. Abraços, José, Lopes, José G.: Extending DRT with a focusing mechanism for pronominal anaphora and ellipsis resolution. Proceedings of the 15th conference on Computational linguistics, Vol. 2. Kyoto, Japan (1994) 1128-1132
9. Sidner, C.: Focusing in the comprehension of definite anaphora. Readings in Natural Language Processing. Morgan Kaufmann (1986) 363-394
10. Kamp, H., Reyle, U.: From discourse to logic. Klawer Academic Publishers (1993)
11. Lappin, S., Leass, H.: An algorithm for pronominal anaphora resolution. Computational Linguistics (1994) 535-561
12. Hobbs, J.: Resolving pronoun references. Lingua, 44 (1977) 311-338