Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Um novo corpo e os seus desafios

Diana Santos[1]

[1]ILOS, Universidade de Oslo
Postboks 1003 Blindern, N-0315 Oslo, NORUEGA

d.s.m.santos@ilos.uio.no

Abstract. *This paper describes the Mariano Gago corpus, a text corpus created after this brilliant Portuguese scientist and politician died, with the aim to create a testbed for question-answering challenges, time-line depictions, summarization, sentiment analysis, reputation anlytics and media studies, as will be detailed in the paper.*

Resumo. *Este artigo apresenta um novo corpo eletrónico publicamente acessível, construído para homenagear um grande professor e político português, José Mariano Gago. Através de uma rica anotação, pretende-se potenciar o desenvolvimento de aplicações inovadoras.*

Pareceu-nos que, do ponto de vista da área do processamento computacional da língua portuguesa, a melhor homenagem a Mariano Gago seria precisamente criar um conjunto de textos que permitissem a avaliação – e o consequente progresso – de várias técnicas e aplicações relevantes, para o português e em geral.

1. O conteúdo

O corpo Mariano Gago inclui presentemente (agosto de 2015) cerca de 350 mil palavras, todas obtidas de fontes na internete, divididas grosso modo em cinco categorias: notícias provocadas pelo falecimento (143 mil palavras), discurso (12 mil), entrevista (31 mil), outras notícias (75 mil), e conteúdo do sítio de homenagem (43 mil), mas prevê-se o seu alargamento com o tempo.

Os seguintes tipos de textos constam do corpo:

- obituário: notícia da morte com um resumo da vida
- testemunho e/ou apreciação: quer em primeira mão, quer noticiado como "reações à morte de"; tanto em jornais, como em blogues ou simplesmente em páginas da internete de instituições ou pessoais
- notícias de acções ou ocorrências provocadas pela morte: no caso em questão, além da notícias do velório e do funeral, informações sobre variadas homenagens (quer anúncio, antes, quer reportagem, depois)
- notícias relacionadas com acontecimentos associados (em particular, discussão sobre se a forma de condolências do primeiro ministro foi apropriada ou não)
- textos da autoria do próprio Mariano Gago (de variadas índoles)
- entrevistas feitas e publicadas
- textos noticiosos sobre atuação ou declarações de Mariano Gago
- textos de crítica ou elogio a ações de Mariano Gago

- textos, por exemplo entrevistas, que mencionam Mariano Gago

É possível levantar ("download") o corpo, ou coleção, na sua totalidade de várias formas, acessíveis de `http://www.linguateca.pt/CorpoJMG/`: (i) na sua forma mais crua, como cinco arquivos em formato textual simples, concatenando sequencialmente cada texto individual, com o título na primeira linha e o URL na última; (ii) numa versão anotada com informação sintática e semântica, pelo PALAVRAS [Bick 2000] e pelos anotadores da Linguateca; (iii) em formato CWB[1].

Além disso, existe um ficheiro separado com informação sobre as fontes (URL) de cada texto; outro com o género ou géneros do texto, a data de publicação e a data a que se refere a notícia (no caso de ser uma notícia e ser possível identificar a data). Prevê-se que ao longo do tempo mais informação irá sendo tornada pública.

2. Metodologia da sua construção

Este corpo foi criado manualmente através da cópia dos resultados obtidas no Google pela pesquisa "José Mariano Gago" ou "Mariano Gago", todos os dias de 17 a 30 de abril. Só as notícias em português, e que não fossem indicadas como oriundas de outro sítio, foram usadas (embora muitas fossem claramente repetidas). Evidentemente que apenas as 30 ou 40 páginas de resultados apresentados puderam ser analisadas (correspondendo a cerca de 400 resultados diariamente), e não toda a Web.

No dia 27 de abril também foi feita a procura "Mariano Gago homenagem", o que produziu bastantes ocorrências da participação deste em homenagens a outras personalidades. A partir do dia 1 de maio e até ao fim desse mês, por considerarmos que o instantâneo da Web a que tínhamos acesso com as procuras iniciais não mudava, as procuras foram outras e mais espaçadas, tais como "Mariano Gago visita" e "Mariano Gago entrevista", praticamente todas elas correspondendo a notícias anteriores ao seu falecimento.

Quando as notícias não eram sobre Mariano Gago mas apenas o mencionavam, escolhemos apenas dois ou três parágrafos das mesmas (incluindo a referência). Se o artigo ou notícia continha três ou mais referências, ou se o nome de Mariano Gago se encontrava no título, usámo-lo todo.

3. Usos deste recurso

A construção deste corpo teve em vista um conjunto de aplicações para os quais poderia servir de teste e de montra ou demonstração, tais como a caracterização do comportamento dos meis de comunicação social com presença na rede, a remoção de duplicados e demais limpeza, a construção de linhas temporais e de outras formas de visualizar um conjunto de documentos relacionados, a identificação e classificação de entidades mencionadas, a resposta automática a perguntas e a geração automática destas para efeitos de compreensão de português como língua estrangeira, a análise de sentimentos e opiniões, e a da reputação, a classificação automática de géneros textuais, e a identificação das fontes de uma notícia.

Por limitações de espaço, apenas discutiremos algumas destas aqui, veja-se o sítio da internete consagrado a este corpo para mais áreas.

[1] Veja-se `openCwb`.

3.1. Panorama dos meios de comunicação portugueses na rede

Quais os atores mais "publicadores"? Quais os mais citados? Quais os mais rápidos? Citam-se entre eles? Qual o panorama de reuso de informação, quer da Agência Lusa, quer de outros materiais? (Veja-se [Clough et al. 2002] sobre a questão do reuso em meios jornalísticos.) Quantos sítios da Internete indicam de onde vem o material publicado?

É possível, a partir desta notícia ou grupo de notícias, ter alguma ideia sobre os atores e a sua forma de atuação? Não estamos evidentemente a afirmar que o estudo da propagação e reuso de uma notícia (ou conjunto delas) pode caracterizar só por si os meios de informação portugueses, mas que a sua análise detalhada pode dar pistas para hipóteses a confirmar em posteriores estudos, assim como desenvolver sistemas (semi?)automáticos que calculam e mostram essa propagação para notícias futuras.

Um trabalho em progresso é a identificação do reuso ou da citação de diferentes partes dos textos ao longo do tempo, de forma a criar uma ilustração do fluxo da informação no tempo e a eventual diferença entre os subtópicos mencionados.

3.2. Construção de uma linha temporal

Uma tarefa relevante para jornalistas ou analistas de informação é, a partir de um conjunto de notícias, estabelecer uma linha temporal, e sistemas que a construam a partir de um conjunto de textos são uma aplicação interessante e útil para permitir lidar com o excesso de informação que nos rodeia.

Em relação ao corpo em questão, podemos estabelecer de facto duas ou três linhas temporais (as quais também são fornecidas a partir de uma anotação humana, para o treino e avaliação de sistemas):

- Dos acontecimentos relatados
- Da publicação das notícias
- Da atuação de Mariano Gago na sua vida

Além disso, constitui material excelente para desenvolver e testar o reconhecimento de datas e marcadores temporais, assim como para investigar a possível diferença na forma da citação à medida que o tempo passa, passando de "hoje", "ontem", "na passada sexta" e "no passado dia 17" a "a 17 de abril", etc.

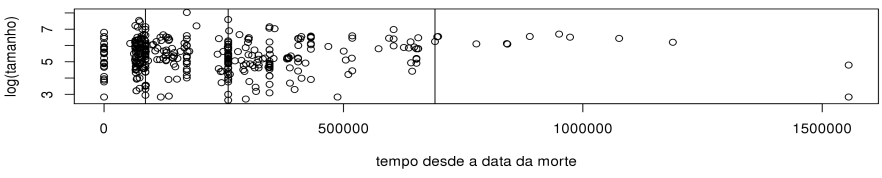

Figura 1. Conteúdo da parte do corpo das notícias que pode ser datada: as linhas verticais indicam 1, 3 e 8 dias respetivamente

Tabela 1. Distribuição das EM referentes a JMG

Mariano Gago	1835	José Mariano	37
José Mariano Gago	566	Prof. José Mariano Gago	25
Professor Mariano Gago	132	Professor Doutor José Mariano Gago	53
José Mariano Rebelo Pires Gago	81	Gago	49
Professor José Mariano Gago	48	Zé Mariano	24

3.3. Identificação e reconhecimento de entidades mencionadas

No caso de um corpo dedicado a uma personalidade, é obviamente interessante identificar TODAS as formas que a ele se referem, e mesmo separá-las por "familiaridade" ou distância em relação ao autor da notícia; opinião positiva ou negativa, etc. Veja-se, a título de exemplo, as diferentes designações usadas para referir Mariano Gago (antes de uma revisão cabal do sistema de REM): Este corpo é além disso ideal para estudar recuperação anafórica e cadeias de referência em português, algo que é possivelmente distinto na nossa língua em comparação com outras [Frankenberg-Garcia 1999].

Questões como *ministro*, *malogrado ministro*[2] ou *ex-ministro* referindo-se à mesma personalidade podem ser muito interessantes de tratar quando o objetivo é uma sumarização ou visualização de um conjunto (incoerente) de textos. (Mariano Gago teve, aliás, vários títulos em governos diferentes...) De facto, e como realçado em [Stoyanov and Cardie 2006], a forma como uma pessoa é mencionada é por si só uma pista importante para mostrar a opinião do autor sobre ela.

A análise das várias relações confessadas ou afirmadas pelos autores dos testemunhos também permite, embora provavelmente muito parcialmente, estabelecer uma imagem de quais as personalidades relacionadas com Mariano Gago e em que relação o foram, através por exemplo de uma rede de personalidades, tal como a proposta por [Hoof 2013] ao estudar cartas antigas de dois mil anos atrás.

3.4. Análise de sentimentos e de opiniões

Outra área para cujo desenvolvimento o presente recurso pretende contribuir é a determinação automática de textos positivos e negativos sobre um dado assunto, ou mesmo de textos concebidos como factuais, por oposição aos que apresentam opiniões do seu autor, área tradicionalmente chamada análise de subjetividade pela comunidade do PLN.Neste caso, seria muito interessante conseguir determinar qual a emoção ou opinião predominante: tristeza, admiração, entusiasmo, gratidão, pena, irritação[3], etc.

Embora este corpo tenha sido automaticamente analisado em relação a emoções a partir de um léxico de emoções abrangente, a deteção da emoção total e das nuances de cada frase está longe de estar resolvida, seja em que língua for. Com este corpo pretendemos por exemplo investigar o campo da admiração, seguindo [Santos and Mota 2015].

É evidente que as pessoas que não apreciam uma personalidade acabada de morrer não lhe escrevem obituários, por isso em geral a opinião dos mesmos sobre o falecido é

[2]Convém indicar que *malogrado* aparece neste corpo apenas no sentido de ter morrido cedo...

[3]Por exemplo, interessante, porque provavelmente inesperada num corpo deste tipo, foi a irritação mencionada por vários autores em relação às condolências expressas pelo primeiro ministro português, que foram consideradas mal formuladas e deram origem não só a piadas como até a críticas ferozes.

positiva. Contudo, graus de distância entre o homenageado e o autor do texto, temas abordados, menção ou não de questões negativas, e a escolha dos termos apropriados, são áreas que seria de grande interesse estudar, para identificar atitudes consensuais e outras divergentes em relação à personalidade em questão.

Outra das áreas relevantes – e complexas – na deteção de sentimento e opiniões, ver [Pang and Lee 2008], é a atribuição correta do detentor da opinião. Com o objetivo de tentar automatizar essa tarefa para o português, marcámos todos os verbos de dizer presentes no material, inspirados por [Freitas 2015].

Finalmente, será que baseado num conjunto de textos deste tipo é possível desenvolver um sistema que tenta atribuir fidedignamente a origem de uma dada notícia ou informação? Quantos de nós não estamos cansados de ler informações contraditórias publicadas por diferentes jornalistas e/ou comentadores, e não termos maneira de saber em que são baseadas? Um sistema que tentasse averiguar, dada uma notícia sobre cujo conteúdo tivéssemos dúvidas, a razão e as fontes que estavam por detrás dela poderia ser muito útil para tornar a informação veiculada mais confiável.

Referências

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical An alysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Aarhus, Denmark.

Clough, P., Gaizauskas, R., and Piao, S. L. (2002). Building and annotating a corpus for the study of journalistic text reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC*, volume 2002), pages 1678–1691.

Frankenberg-Garcia, A. (1999). Crosslinguistic influence as a key to extracting second language teaching materials for monolingual classes from translation corpora. In Granger, S., editor, *Proceedings of the Workshop Contrastive Linguistics and Translation Studies: Empirical Approaches*.

Freitas, B. (2015). Discurso relatado: relatório parcial sobre a obtenção dos verbos do dizer. Technical report, PUC Rio.

Hoof, L. V. (2013). Dead languages and digital humanties: Social network analysis in the ancient world. What are Digital Humanitities? UiO, June 14-15, 2013.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Santos, D. and Mota, C. (2015). A admiração à luz dos corpos. In Simões, A., Barreiro, A., Santos, D., Sousa-Silva, R., and Tagnin, S. E. O., editors, *Linguística, Informática e Tradução: Mundos que se Cruzam. Homenagem a Belinda Maia*, pages 57–77.

Stoyanov, V. and Cardie, C. (2006). Partially supervised coreference resolution for opinion summarization th rough structured rule learning. In *Proceedings of EMNLP 2006, Sydney, July*, pages 336–344.