Do PDF ao TXT: Desafios na extração de informação em textos técnico-científicos

Aline Silveira
Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
silveira26aline@gmail.com

Elvis de Souza
Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
elvis.desouza99@gmail.com

Tatiana Cavalcanti
Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
tatiana.shc@hotmail.com

Cláudia Freitas
Departamento de Letras
PUC-Rio
Rio de Janeiro, Brasil
claudiafreitas@puc-rio.br

***Index Terms*—Extração de informação, Documentos técnico-científicos, Pré-processamento, Processamento de Linguagem Natural**

I. Introdução

Extração de Informação (EI) é o processo de examinar um texto automaticamente visando a capturar informações relevantes para determinado interesse. Em outras palavras, o objetivo da extração de informação é transformar textos, que são fonte de dados não estruturados, em informação organizada de acordo com certos critérios de importância, deixando o que não interessa para trás [1]. No tocante ao modelo de texto científico, tal técnica se mostra relevante porque pode revelar tendências de pesquisa científica ao longo do tempo, concatenar informações de fontes muito vastas em volume e diversidade, evidenciar citações ausentes e proximidade temática entre os autores, e auxiliar na construção de bancos de dados a partir de grandes corpora textuais, entre outras aplicações. Nosso objetivo é construir um grande corpus de textos técnico-científicos de domínio específico com suas informações estruturadas e prontas para processamento futuro, tendo em vista facilitar buscas semânticas no conteúdo; para isso, EI se torna fundamental.

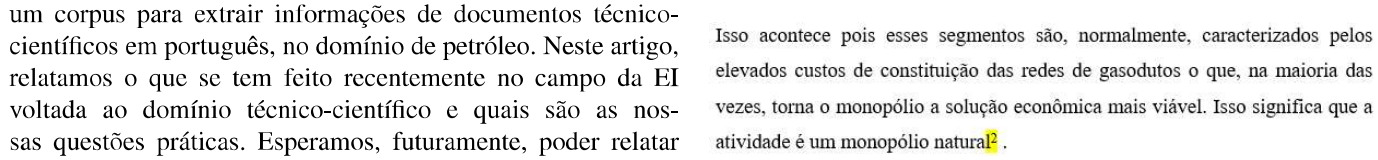
O Processamento de Linguagem Natural (PLN) viabiliza a conversão de um texto cru, essencialmente uma sequência aleatória de bits digitais, em uma sequência bem definida de unidades linguísticas detentoras de sentido [2]. Essa definição de caracteres, palavras e sentenças é fundamental porque, uma vez feita, serve de entrada para uma outra tarefa crucial dentro do PLN, a anotação morfossintática. A anotação morfossintática, por sua vez, pode facilitar a Extração de Informação, porque oferece um guia prévio de generalizações para o sistema ao dar o primeiro passo de reconhecimento de padrões linguísticos.

Nosso interesse na EI é resultado da necessidade de criar um corpus para extrair informações de documentos técnico-científicos em português, no domínio de petróleo. Neste artigo, relatamos o que se tem feito recentemente no campo da EI voltada ao domínio técnico-científico e quais são as nossas questões práticas. Esperamos, futuramente, poder relatar nossos resultados empíricos e, assim, contribuir de forma relevante para o estado da arte em extração de informação em textos técnico-científicos.

II. Do PDF ao TXT: desafios do pré-processamento

Para que qualquer texto de língua natural seja legível por uma máquina, são necessários alguns passos iniciais. O primeiro é garantir que todos os textos de que dispomos têm seus caracteres codificados da mesma forma (em Latin-1, UTF-8, CP1252, etc.). A essa primeira tarefa dá-se o nome de identificação de codificação de caractere. O segundo passo é a identificação da língua em que o texto está escrito. E o último grande passo é descartar informações indesejáveis como imagens, tabelas, cabeçalhos, links e notas de rodapé. Ao final desse processo, o que se tem é um texto bem definido, organizado pela sua língua, pronto para ser segmentado e analisado em níveis mais complexos [3].

Como já foi visto, alguns detalhes podem atrapalhar a leitura da ferramenta computacional utilizada e, por conseguinte, influenciar no resultado final da análise do texto. Alguns deles, por exemplo, no caso dos textos técnicos-científicos da categoria dissertação ou tese, são o sumário, os agradecimentos, a folha de aprovação, listas em geral, figuras e tabelas. Esses elementos, juntamente com paginação e notas de rodapé, quando transformados em texto plano (formato .txt), apresentam deformação e rompem com a linearidade dos textos. A figura 1, obtida a partir de um dos documentos de nosso corpus, ilustra a questão das notas de rodapé. Como podemos observar, a palavra "natural" se transforma em outra palavra ("natural2"). Ou seja, se "natural" for uma palavra relevante, perderíamos essa ocorrência e esse contexto. A figura 2 (oriunda do corpus Brasileiro e acessada pelo serviço AC/DC [4]), ilustra que, sem o cuidado necessário no pré-processamento, o que seria uma frase se transforma em blocos de texto contendo números e tabelas que, mais tarde, serão erroneamente processados como períodos únicos, dificultando etapas posteriores do processamento automático de texto.


Isso acontece pois esses segmentos são, normalmente, caracterizados pelos elevados custos de constituição das redes de gasodutos o que, na maioria das vezes, torna o monopólio a solução econômica mais viável. Isso significa que a atividade é um monopólio natural[2] .

Isso acontece pois esses segmentos são, normalmente, caracterizados pelos elevados custos de constituição das redes de gasodutos o que, na maioria das vezes, torna o monopólio a solução econômica mais viável. Isso significa que a atividade é um monopólio natural2

Figura 1. Exemplo de distorção relativa a notas de rodapé.

<p>: Gêneros Nºde ocorrências 1 Freqüência 2 Penicillium spp 50 53,161 % Aspergillus spp 19 41,333 % Cladosporium spp 10 4,534 % Rhizopus spp 1 0,343 % Fusarium spp 5 0,215 % Aureobasidium spp 1 0,143 % Chrysosporium spp 5 0,100 % Alternaria spp 4 0,057 % Epicoccum spp 1 0,029 % Helminthosporium spp 1 0,029 % Geotrichum spp 1 0,014 % Gliocladium spp 1 0,014 % Nigrospora spp 1 0,014 % Rhizomucor spp 1 0,014 % 1 Refere-se ao número de vezes que um determinado **gênero** foi identificado independentemente das contagens de UFC .

Figura 2. Exemplo de distorção relativa a tabelas.

O trabalho de [5], também voltado para a constituição de um corpus de documentos técnico-científicos, enfrentou - e resolveu - vários dos problemas mencionados aqui, ainda que nem todas as soluções utilizadas se adequem aos nossos propósitos. Reconhecida como a etapa "mais importante e mais trabalhosa"na constituição de um corpus que, posteriormente, será anotado morfossintaticamente, as autoras utilizaram editores de texto e expressões regulares, associados à supervisão humana, para eliminar ruídos associados a representações numéricas, o uso de ponto final em nomes próprios como *Dra.*, e asteriscos associados a palavras. Por outro lado, elementos que nos interessa manter, como títulos e referências bibliográficas ao longo do texto, foram excluídas. E um dos nossos principais problemas, em funçao do tipo de texto, nao parece ter sido um problema: as enumerações itemizadas, retratadas na figura 2.

Em suma, não faltam desafios ao pré-processamento de textos. Tendo essa etapa sido concluída corretamente, ou seja, uma vez que tenhamos um corpus de qualidade, as etapas subsequentes do processamento automático podem acontecer da melhor maneira possível.

III. EI em documentos técnico-científicos: SemEval 2017 e 2018

Para otimizar a eficiência em qualquer trabalho de investigação, é imprescindível fazer uma busca para que se conheça o que já foi feito no campo e que tipo de conclusões já foram tiradas. Considerando que o nosso interesse de pesquisa reside em textos técnico-científicos, fez sentido analisar, especificamente, as tarefas apresentadas nas avaliações SemEval de 2017 e 2018. Por meio dessas competições e a partir de suas várias tarefas e subtarefas, que abarcam diversos temas e objetivos específicos, os participantes têm a oportunidade de testar seus sistemas e avaliar sua posição em relação aos outros da mesma área. Percebemos que, ao observarmos os métodos, as escolhas, as conclusões e os resultados obtidos na avaliação, ganhamos um melhor entendimento acerca de nosso próprio trabalho, podendo aproveitar o que deu bons frutos e descartar o que falhou.

Em particular, vale descrever as tarefas de 2017 e 2018 individualmente. No SemEval de 2017, a tarefa 10 (ScienceIE - Extracting Keyphrases and Relations from Scientific Publications [6]) foi a de identificar termos-chave, classificá-los e relacionar os de mesma classificação a partir de um único parágrafo de publicações científicas das áreas de Ciência da Computação, Ciência de Materiais e Física. Um dos maiores objetivos dos desenvolvedores da tarefa era o de atender aos editores de publicações científicas, tendo em vista que com essa tarefa eles poderiam recomendar artigos aos leitores, identificar revisores em potencial para submissão e analisar tendências de pesquisa ao longo do tempo.

Para realizar tal tarefa foi necessário subdividi-la em três subtarefas: a primeira foi a de identificação dos termos-chave - subtarefa A; a segunda, de classificação desses termos em 3 categorias diferentes (PROCESS, TASK e MATERIAL) - subtarefa B, e a terceira, de classificação do tipo de relação entre os termos de mesma categoria em sinônimo ou hiperônimo - subtarefa C. É importante ressaltar que havia três cenários de avaliação diferentes, um com o texto cru, simplesmente – necessitando que as três subtarefas fossem realizadas sem nenhum ponto de partida; outro com a identificação manual prévia dos termos-chave - ou seja, com a subtarefa A já realizada previamente; e um terceiro cenário com a identificação manual prévia dos termos-chave e suas classificações – ou seja, com as subtarefas A e B dadas previamente. Os sistemas que participaram da competição variaram muito em técnicas, desde redes neurais até métodos baseados em regras. Isso mostra a diversidade de maneiras de resolver esta tarefa. Para a anotação do corpus que serviu de treinamento e avaliação aos sistemas foram recrutados estudantes de graduação dos mesmos domínios dos textos e professores dessas mesmas áreas. A concordância entre eles oscilou entre 45% e 85%, sendo metade dos casos com concordância maior ou igual a 60%. Tal fato mostra a dificuldade humana em detectar termos-chave, suas classificações e relações, mesmo quando os anotadores são especialistas no domínio. Do mesmo modo, a observação da concordância entre os anotadores nos informa o grau de dificuldade que a máquina enfrenta ao realizar essas tarefas, evidentemente complexas.

Vejamos os 3 cenários diferentes em que as avaliações ocorreram. No primeiro cenário de avaliação, com o texto cru, participaram 17 sistemas, dos quais o maior pontuador teve um desempenho de 43% de acerto (F1) e utilizou o método de rede neural recorrente. Já no segundo cenário de avaliação, em que apenas as subtarefas B e C deveriam ser resolvidas, somente 4 competidores participaram. O ganhador teve um acerto (F1) de 64%, e fez uso de classificadores com características lexicais, características ortográficas e n-gramas para a subtarefa B. Para a subtarefa C, o ganhador fez uso de um sistema baseado em regras na parte de sinônimos e de padrões da Hearst [7] para detectar hiperônimos. No último cenário, no qual a única subtarefa a ser realizada foi a C - relacionar os itens de mesma classificação -, a melhor atuação - dentre as 5 participantes - foi do sistema de redes neurais convolucionais que alcançou a marca de 64%, tal como no cenário anterior. Realça-se aqui

o salto de desempenho entre os cenários, que se relaciona diretamente com a importância da precisão da subtarefa A para o sucesso das subtarefas consecutivas: isto é, identificar quais são os termos-chave parece ser a parte mais difícil de toda a tarefa.

Essa tarefa da competição SemEval de 2017 mostra a variação metodológica com que sistemas podem operar. Por outro lado, mostra também que ainda há um grande espaço para avanços, já que 64%, embora não seja um índice baixo, pode ser insatisfatório do ponto de vista prático, isto é, de quem irá tirar proveito dos resultados.

No ano seguinte, no SemEval 2018, a tarefa 7 (Semantic Relation Extraction and Classification in Scientific Papers [8]) foi a de identificar e classificar – entre 6 categorias pré-determinadas (USAGE, RESULT, MODEL, PART_WHOLE, TOPIC, COMPARISON) – as relações semânticas entre entidades presentes nos resumos de artigos científicos. Os artigos eram todos pertencentes ao domínio da Linguística Computacional. Segundo o artigo que descreve a tarefa e seus resultados, um de seus objetivos seria melhorar o acesso à literatura científica, atendendo a uma necessidade de informação que não estaria sendo suprida pelas ferramentas padrão de pesquisa, nem pelos humanos especialistas nos domínios específicos, que, muitas vezes, não dispõem do tempo necessário para se atualizar em relação aos avanços científicos nas suas áreas. Nesse caso, seriam de grande ajuda sistemas que estruturassem essas informações automaticamente.

A partir dessa tarefa principal, definiram-se três subtarefas: a classificação de relações em clean data (1); a classificação de relações em noisy data (2); e a identificação e a classificação de relações em material não previamente anotado (3), exceto pelas entidades, que estão anotadas em todos os experimentos. Ao todo, 32 times participaram de ao menos uma das subtarefas.

Na primeira subtarefa (clean data), as entidades cujas relações serão classificadas tinham sido previamente anotadas manualmente, enquanto na segunda (noisy data), isso havia sido feito de forma automática. Desse modo, a terceira subtarefa é aquela mais sujeita a dificuldades, já que as relações não só precisam ser categorizadas, mas também previamente identificadas pelos sistemas competidores, que não contam com as regalias das duas primeiras subtarefas.

Para a anotação do corpus, foram recrutados anotadores especialistas dentre os membros da organização da tarefa, assim como estudantes de PLN. A concordância inter-anotadores relativa à rotulação das classes semânticas foi de 90.8%, tendo sido calculada entre dois anotadores a partir de uma amostra de 150 resumos provenientes da primeira subtarefa. É importante acentuar aqui o valor inegável de uma alta concordância entre anotadores, uma vez que um "golden" construído com base em divergências está muito mais sujeito a inconsistências.

Ao final da competição, concluiu-se que o maior desafio dentre as etapas do processamento era a identificação adequada das relações semânticas. Isso fica mais claro ao observarmos os melhores valores de F1 para cada subtarefa (Tabela 1), já que observamos uma queda na performance dos sistemas na terceira subtarefa, em que a presença de relações entre entidades não era dada anteriormente.

Tabela I
RESUMO ENTRE OS RESULTADOS DAS COMPETIÇÕES DE EXTRAÇÃO DE INFORMAÇÃO.

Competição	Subtarefas	F1
SemEval 2017 Tarefa 10	A identificação dos termos-chave	43%
	B classificação dos termos-chave	64%
	C classificação das relações entre termos-chave	64%
SemEval 2018 Tarefa 7	1 classificação das relações em clean data	81.7%
	2 classificação das relações em noisy data	90.4%
	3 extração e classificação das relações	49.3%

Ademais, ao tratar dos tipos de relações semânticas, afirma-se que mais do que o significado dessas relações, a maior dificuldade é distribuição desigual das identificações de entidades nas subtarefas com clean e noisy data. As entidades anotadas manual e automaticamente foram de diferentes naturezas: os seres humanos eram orientados a anotar termos e expressões mais complexas, enquanto a máquina anotava termos menores e com um nível mais baixo de especificidade.

Destaca-se, por fim, a atualidade dos trabalhos objetivados pelo SemEval, que acompanham as pesquisas correntes e espelham os tópicos que se sobressaem em determinado período. Competições como essa não só estimulam seus integrantes a aprimorarem seus sistemas e a desenvolverem suas pesquisas, como também nos ajudam a entender como diferentes grupos estão pensando o mesmo assunto.

IV. LIÇÕES APRENDIDAS

Tendo em vista tudo o que foi exposto aqui, nota-se o grau de dificuldade das tarefas a que as máquinas estão submetidas. Se para nós, humanos, mesmo diante do nosso vasto conhecimento sobre língua e *expertise* nas temáticas específicas, definir os limites de palavras e expressões e se elas podem ser consideradas relevantes ou não em determinado contexto é tarefa árdua, quiçá para sistemas de processamento automático de texto, que, segundo os resultados das competições analisadas, não chegam a 80%.

No entanto, é preciso retomar aqui que em nenhuma das tarefas apresentadas foi necessário um pré-processamento detalhado dos textos analisados. Ou seja, nosso objetivo de aprender sobre os desafios do pré-processamento de documentos técnico-científicos com a experiência de outros trabalhos foi parcialmente alcançado, já que no SemEval de 2017 o corpus é composto de parágrafos, e no de 2018, de resumos. Com essa decisão, os organizadores evitaram todo o trabalho

anterior de preparação dos textos. Em nosso caso, por outro lado, iremos trabalhar com documentos completos, o que não permite essa abordagem imediata. A exceção cabe ao trabalho de [5], que se debruçaram sobre algumas das questões que também nos inquietam. No entanto, as autoras salientam a necessidade de algum acompanhamento humano (trata-se de um processamento semi-automático).

Um possível caminho a ser tomado é descartar a ideia de um processamento de texto integral e trabalhar apenas com fragmentos, como fizeram as tarefas que descrevemos. O quanto se ganha ao processar documentos completos? O quanto se perde processando apenas resumos ou parágrafos? Aliás, perdemos alguma coisa? É um interesse nosso avaliar a diferença na qualidade da informação extraída automaticamente em documentos completos ou em fragmentos especialmente relevantes, como resumos. No entanto, a única maneira de verificar isso é comparando as duas situações, o que envolve o processamento dos documentos na íntegra. Ao que parece, precisaremos desenvolver nossas próprias soluções para o pré-processamento de documentos técnico-científicos.

AGRADECIMENTOS

Este projeto foi financiado com o apoio da ANP - Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associado ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio.

REFERÊNCIAS

[1] J. R. Hobbs and E. Riloff, "Information extraction," in *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau, Eds. Chapman & Hall/CRC, 2010.

[2] D. D. Palmer, "Text preprocessing," in *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau, Eds. Chapman & Hall/CRC, 2010.

[3] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, Jun. 1999, pp. 3–10. [Online]. Available: https://www.aclweb.org/anthology/P99-1001

[4] D. Santos and E. Bick, "Providing Internet access to Portuguese corpora: the AC/DC project," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf

[5] L. Lopes and R. Vieira, "Building domain specific parsed corpora in portuguese language," in *Proceedings of the 10th National Meeting on Artificial and Computational Intelligence (ENIAC)*, Fortaleza, Brasil, 2013.

[6] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, "SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 546–555. [Online]. Available: https://www.aclweb.org/anthology/S17-2091

[7] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, ser. COLING '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 539–545. [Online]. Available: https://doi.org/10.3115/992133.992154

[8] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, "SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 679–688. [Online]. Available: https://www.aclweb.org/anthology/S18-1111