Geocodificação de Documentos Textuais com Classificadores Hierárquicos Baseados em Modelos de Linguagem

Duarte Dias
IST
dcd@ist.utl.pt

Ivo Anastácio
INESC-ID Lisboa / IST
ivo.anastacio@ist.utl.pt

Bruno Martins
INESC-ID Lisboa / IST
bruno.g.martins@ist.utl.pt

Resumo

A maioria dos documentos textuais, produzidos no contexto das mais diversas aplicações, encontra-se relacionado com algum tipo de contexto geográfico. Contudo, os métodos tradicionais para a prospecção de informação em colecções de documentos vêem os textos como conjuntos de termos, ignorando outros aspectos. Mais recentemente, a recuperação de informação com suporte ao contexto geográfico tem capturado a atenção de diversos investigadores em áreas relacionadas com a prospecção de informação e o processamento de linguagem natural, envisionando o suporte para tarefas como a pesquisa e visualização de informação textual, com base em representações cartográficas. Neste trabalho, comparamos experimentalmente diferentes técnicas automáticas, as quais utilizam classificadores baseados em modelos de linguagem, para a atribuição de coordenadas geoespaciais de latitude e longitude a novos documentos, usando apenas o texto dos documentos como evidência de suporte. Medimos os resultados obtidos com modelos de linguagem baseados em n-gramas de caracteres ou de termos, usando colecções de artigos georreferenciados da Wikipédia em três línguas distintas, nomeadamente em Inglês, Espanhol e Português. Experimentamos também diferentes métodos de pós-processamento para atribuir as coordenadas geoespaciais com base nas classificações. O melhor método utiliza modelos de linguagem baseados em n-gramas de caracteres, em conjunto com uma técnica de pós-processamento que utiliza as coordenadas dos knn documentos mais similares, obtendo um erro de previsão médio de 265 Kilómetros, e um erro mediano de apenas 22 Kilómetros, para o caso da colecção da Wikipédia Inglesa. Para as colecções Portuguesa e Espanhola, as quais são significativamente mais pequenas, o mesmo método obteve um erro de previsão médio de 278 e 273 Kilómetros, respectivamente, e um erro de previsão mediano de 28 e de 45 Kilómetros.

Palavras chave

Processamento de Texto, Recuperação de Informação Geográfica, Geocodificação de Documentos

Abstract

Most text documents can be said to be related to some form of geographic context, although traditional text mining methods simply model documents as bags of tokens, ignoring other aspects of the encoded information. Recently, geographic information retrieval has captured the attention of many different researchers from fields related to text mining and data retrieval, envisioning the support for tasks such as map-based document indexing, retrieval and visualization. In this paper, we empirically compare automated techniques, based on language model classifiers, for assigning geospatial coordinates of latitude and longitude to previously unseen textual documents, using only the raw text of the documents as input evidence. We measured the results obtained with character-based or token-based language models over collections of georeferenced Wikipedia articles in four different languages, namelly English, Spanish and Portuguese. We also experimented with different post-processing methods for assigning the geospatial coordinates with basis on the resulting classifications. The best performing method combines character-based language models with a post-processing technique that uses the coordinates from the k most similar documents, obtaining an average prediction error of 265 Kilometers, and a median prediction error of just 22 Kilometers, in the case of the English Wikipedia collection. For the Spanish, and Portuguese collections, which are significantly smaller, the same method obtain an average prediction error of 273 and 278 Kilometers, respectively, and a median prediction error of 45 or 28 Kilometers.

Keywords

Text Mining; Geographic Information Retrieval; Document Geocoding;

1 Introdução

A maioria dos documentos textuais, produzidos no contexto das mais diversas aplicações, encontra-se relacionado com algum tipo de contexto geográfico. Recentemente, a Recuperação de Informação (RI) com base no contexto geográfico tem capturado a atenção de muitos investigadores em áreas relacio-

nadas com o processamento de língua natural e a prospecção de informação em grandes colecções de documentos textuais. Temos, por exemplo, que a tarefa de resolver referências a nomes de locais, apresentadas em documentos de texto, tem sido abordada em diversos trabalhos anteriores, com o objectivo de apoiar tarefas subsequentes em sistemas de RI geográficos, tais como a recuperação de documentos ou a visualização através de representações cartográficas (Lieberman e Samet, 2011). No entanto, a resolução de referências a nomes de locais apresenta vários desafios não-triviais (Leidner, 2007; Martins, Anastácio e Calado, 2010; Amitay et al., 2004), devido à ambiguidade inerente ao discurso em linguagem natural. Temos, por exemplo, que os nomes de locais são muitas vezes usados com outros significados não geográficos. Temos ainda que locais distintos são muitas vezes referidos pelo mesmo nome, ou que locais únicos são referidos por nomes diferentes. Além disso, existem muitos termos do vocabulário de uma dada língua, além dos nomes de locais, que podem surgir frequentemente associados com áreas geográficas específicas. Em lugar de tentar resolver corretamente as referências individuais a locais, que sejam apresentadas em documentos textuais, pode ser bastante interessante estudar métodos para a atribuição de âmbitos geográficos à totalidade dos conteúdos dos documentos (Wing e Baldridge, 2011; Adams e Janowicz, 2012). Os resultados poderão posteriormente ser aplicados em tarefas como a sumarização de colecções de documentos em representações baseadas em mapas (Bär e Hurni, 2011; Mehler et al., 2006; Erdmann, 2011).

Neste trabalho, é feita uma comparação de diferentes técnicas automáticas para a atribuição de coordenadas geoespaciais de latitude e longitude a novos documentos textuais, usando apenas o texto dos documentos como fonte de evidência, e utilizando uma representação discreta para superfície da Terra baseada numa decomposição em regiões triangulares de igual área. As diferentes regiões usadas na representação da Terra são inicialmente associadas aos documentos textuais que lhes pertencem (i.e., usamos todos os documentos presentes num conjunto de treino, que sejam conhecidos por se referir a cada uma das regiões em particular). De seguida, são construídas representações compactas (e.g., baseadas em modelos de linguagem suportados em n-gramas de caracteres ou de termos) a partir desses conjuntos de documentos georeferenciados, capturando as suas principais propriedades estatísticas. Novos documentos são então atribuídos à(s) região(ões) mais semelhante(s). Finalmente, são atribuídas as respectivas coordenadas geospaciais de latitude e longitude aos documentos, com base nas coordenadas centroides associadas à(s) região(ões). Foram ainda realizadas experiências com diferentes técnicas de pós-processamento para atribuir as coordenadas na etapa final, usando (i) as coordenadas centroide da região mais provável, (ii) uma média ponderada com as coordenadas das regiões mais prováveis, (iii) uma média ponderada com as coordenadas das regiões vizinhas da mais provável, e (iv) uma média ponderada com as coordenadas dos knn documentos de treino mais semelhantes (Shakhnarovich, Darrell e Indyk, 2006), que estejam contidos dentro da região mais provável para o documento.

Experiências com coleções de artigos da Wikipédia, contendo documentos em Inglês, Espanhol e Português, apresentaram bons resultados para a abordagem geral de geocodificação. O melhor método combina classificadores hierárquicos baseados em modelos de linguagem, usando n-gramas de caracteres, com a técnica de pós-processamento que utiliza a média ponderada das coordenadas dos knn documentos mais similares. Este método obteve um erro de previsão médio de 265 Kilómetros, e um erro de previsão mediano de apenas 22 Kilómetros, para o caso da colecção da Wikipédia Inglesa. Para as colecções Portuguesa e Espanhola, o mesmo método obteve um erro médio de 278 e 273 Kilómetros, e um erro mediano de 28 e 45 Kilómetros, respectivamente em cada uma das colecções.

O restante conteúdo deste artigo está organizado da seguinte forma: a Secção 2 apresenta trabalhos anteriores relacionados com a geocodificação de documentos. A Secção 3 apresenta a abordagem proposta, detalhando o uso dos classificadores baseados em modelos de linguagem, assim como as técnicas de pós-processamento propostas. A Secção 4 apresenta a validação experimental do método proposto, descrevendo os conjuntos de dados da Wikipédia que foram considerados, o protocolo experimental, e os resultados obtidos para as diferentes variações do método proposto. Finalmente, a Secção 5 sumariza as principais conclusões do trabalho, apontando ainda possíveis direções para trabalho futuro.

2 Trabalho Relacionado

A relação entre a linguagem e geografia tem sido um tema de interesse para os linguistas (Johnstone, 2010). Muitos estudos têm, por exemplo, mostrado que a geografia tem um impacto importante na relação entre termos do vocabulário e classes semânticas. Temos, por exemplo, que o termo *footbal*l, nos Estados Unidos, se refere ao desporto em particular de futebol americano. No entanto, em regiões como a Europa, o termo *football* é geralmente associado a diferentes modalidades desportivas (e.g., o futebol ou, menos frequentemente, rugby). Termos como *praia* ou *neve* também são mais propensos a serem associados a determinados locais. Neste estudo,

estamos interessados em ver se os termos do vocabulário, e se conteúdos textuais no geral, podem ser usados para prever localizações geográficas.

Overell (2009) investigou o uso da Wikipédia como fonte de dados para a geocodificação de artigos textuais, assim como para a classificação de artigos por categorias, ou para a resolução de referências individuais a nomes de locais. O objetivo principal de Overell era a resolução de referências a locais em documentos, tarefa para a qual a geocodificação de documentos global pode servir como fonte de evidência. Para a geocodificação de documentos, Overell propôs um modelo simples que usa apenas os metadados disponíveis (e.g., título do artigo, hiperligações de entrada e saída para com outros documentos, etc.), e não o próprio texto dos documentos.

Adams e Janowicz (2012) estudaram a relação entre os tópicos em documentos textuais e a sua distribuição geoespacial. Enquanto que a maioria dos trabalhos anteriores, focados na extração de informação geográfica desde documentos, se baseiam em palavras-chave específicas, tais como os nomes de locais, Adams e Janowicz propuseram uma abordagem que usa apenas termos e expressões não geográficas, aferindo sobre se os termos textuais comuns são também bons na previsão de localizações geográficas. A técnica proposta usa o modelo *Lattent Dirichlet Allocation* (LDA) para descobrir tópicos latentes na coleção de documentos. LDA é essencialmente um método não-supervisionado que permite modelar o processo de geração de documentos através de misturas probabilísticas de tópicos, os quais são por sua vez modelados como distribuições de probabilidade sobre um vocabulário de termos. Depois de ajustar o modelo LDA a uma colecção de documentos, os autores utilizam a técnica *Kernel Density Estimation* (KDE) para interpolar uma superfície de densidade, correspondendo a uma região geoespacial, ao longo de cada tópico do modelo LDA. Notando que cada documento pode ser visto como uma mistura de tópicos, os autores utilizam operações de álgebra de mapas para combinar as superfícies de densidade geradas com base em cada tópico, finalmente atribuindo aos documentos o local geoespacial de maior densidade.

Eisenstein et al. (2010) investigaram as diferenças dialetais e as variações em interesses regionais nos utilizadores do Twitter, utilizando uma coleção de *tweets* georreferenciados e uma técnica baseada em modelos probabilísticos. Especificamente, estes autores tentaram georeferenciar os utilizadores do Twitter localizados nos Estados Unidos, com base nos conteúdos por sí produzidos. Eles concatenaram todos os *tweets* de cada utilizador distinto, e usaram distribuições Gaussianas para modelar as localizações dos utilizadores. As abordagens propostas no nosso artigo usam, alternativamente, uma representação discreta para a superfície da Terra, em conjunto com modelos probabilísticos mais simples construídos sobre essa representação discreta.

Anastácio, Martins e Calado (2010) estudaram abordagens heurísticas para atribuir âmbitos geográficos a documentos textuais, com base no reconhecimento de referências a locais nos documentos, posteriormente combinando as referências reconhecidas. Os autores compararam especificamente abordagens com base (i) na frequência de ocorrência associada às referências a locais, (ii) na sobreposição geoespacial entre caixas delimitadoras associadas às referências, (iii) na distância hierárquica entre as referências, usando uma taxonomia geográfica de divisões administrativas, e (iv) na propagação de informação sobre um grafo codificando relações entre locais, usando novamente uma taxonomia geográfica com divisões administrativas. Experiências com uma coleção de páginas Web do *Open Directory Project*[1] mostraram que a técnica baseada na distância hierárquica consegue bons resultados. Neste trabalho, estamos também a estudar abordagens para a geocodificação do conteúdo de documentos textuais, mas neste caso usando directamente o texto como fonte de evidência, em alternativa à utilização de referências a locais nos textos.

Wing e Baldridge (2011), num estudo muito semelhante ao que é relatado no presente artigo, compararam abordagens diferentes para a geocodificação automática de documentos, usando também como base modelos estatísticos derivados de um conjunto vasto de documentos já geocodificados, como a Wikipédia. Os autores utilizaram a divergência de Kullback-Leibler entre um modelo de linguagem construído sobre um documento de teste, e modelos de linguagem para cada célula de uma representação discreta para a superfície da Terra, como forma de prever a célula mais provável de conter o documento de teste. Uma abordagem semelhante foi posteriormente proposta para a resolução temporal de documentos, sendo capaz de determinar a data da publicação de um dado artigo, com base no texto (Kumar, Lease e Baldridge, 2011). Novamente neste trabalho, os autores construíram histogramas que codificam a probabilidade de diferentes períodos temporais para um documento, mais tarde usando a divergência de Kullback-Leibler para fazer as previsões. O trabalho relatado neste artigo é muito semelhante ao de Wing e Baldridge, mas nós propomos utilizar (i) um esquema diferente para particionar o conjunto de documentos em regiões de igual área, de acordo com sua localização geoespacial, (ii) uma abordagem diferente para a classificação de documentos através de modelos de linguagem, (iii) uma abordagem de

[1] `http://www.dmoz.org/`

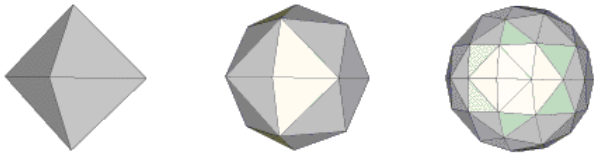


Figura 1: Decomposição da superfície terrestre com grelhas triangulares de resolução zero, um e dois.

decomposição hierárquica para melhorar o desempenho computacional do método de classificação, e (iv) diferentes técnicas de pós-processamento para atribuir as coordenadas geoespaciais com base na classificação obtida, i.e. com base nas pontuações associadas a cada célula da decomposição da Terra.

3 Geocodificação de Documentos

A abordagem de geocodificação de documentos textuais, proposta neste artigo, baseia-se na discretização da superfície da Terra num conjunto de células triangulares, o que nos permite prever os locais, a associar aos documentos, com abordagens estatísticas padrão para a modelação de atributos discretos. No entanto, ao contrário de autores anteriores como Serdyukov, Murdock e van Zwol (2009) ou Wing e Baldridge (2011), os quais utilizaram uma grelha de células rectangulares, nós utilizamos uma grelha triangular, obtida através de um método de decomposição da superfície da Terra conhecido pela designação de *Hierarchical Triangular Mesh*[2] (Dutton, 1996; Szalay et al., 2005). Esta estratégia resulta numa grelha triangular que preserva uma área aproximadamente igual para cada célula, em lugar de resultar em células de tamanho variável, com regiões que se encolhem de acordo com a latitude, tornando-se progressivamente menores e alongadas à medida que se aproximam dos polos. Importa aqui referir que a nossa representação ignora todas as regiões geográficas de nível semanticamente superior, como os estados, países ou continentes. No entanto, esta representação é apropriada para o propósito de geocodificar documentos textuais, uma vez que os mesmos podem estar relacionados com regiões geográficas que não se encaixam numa divisão administrativa da superfície da Terra.

A *Hierarchical Triangular Mesh* (HTM) oferece uma decomposição multi-nível recursiva para uma aproximação esférica da superfície da Terra – ver as

Figura 2: Decomposição de triângulos esféricos.

Figuras 1 e 2, ambas adaptadas de imagens originais do sítio Web descrevendo a abordagem HTM. A decomposição começa num nível zero com um octaedro e, projectando as arestas do octaedro sobre a esfera terrestre, criam-se 8 triângulos esféricos (i.e., triângulos projectados sobre a esfera terrestre), 4 no hemisfério Norte e 4 sobre o hemisfério Sul. Quatro destes triângulos partilham um vértice nos polos, e os lados opostos formam o Equador. Cada um dos 8 triângulos esféricos pode ser dividido em quatro triângulos menores, através da introdução de novos vértices nos pontos médios de cada aresta, e adicionando um segmento de arco de grande círculo para conectar os novos vértices. Este processo de subdivisão pode repetir-se recursivamente, até chegarmos ao nível de resolução desejado. Os triângulos nesta grelha são as células utilizadas na nossa representação da Terra, e cada triângulo, em qualquer resolução, é representado por um identificador único. Para cada localização dada por um par de coordenadas sobre a superfície da esfera terrestre, existe um identificador que representa o triângulo, considerando uma resolução em particular, que contém o ponto correspondente.

Note-se que o mecanismo de representação proposto contém um parâmetro k que controla a resolução, i.e., a área das células. Na nossa aplicação de classificação de documentos, a utilização de células de granularidade elevada pode levar a estimativas muito grosseiras, muito embora a precisão da classificação, com uma resolução mais fina, também possa diminuir substancialmente, devido a termos dados insuficientes para construir os modelos de linguagem associados a cada célula. Nas nossas experiências, o parâmetro k variou entre os valores de 4 e 10, com 0 correspondendo a uma divisão de primeiro nível. A Tabela 1 apresenta o número de células gerado em cada um dos níveis de resolução considerados. O número de células n para uma resolução k é dado por $n = 8 * 4^k$. A Tabela 1 mostra também a área, em Kilómetros quadrados, correspondente a cada célula da representação.

Com base na representação em células para a superfície da Terra, dada pelo método HTM, foi depois utilizado o software *Alias-I LingPipe*[3] para construir modelos de linguagem baseados em n-gramas de caracteres ou de termos, usando-se segui-

[2] `http://www.skyserver.org/htm/Old_default.aspx`

[3] `http://alias-i.com/lingpipe`

Resolução	4	6	8	10
Número total de células	2,048	32,768	524,288	2,097,152
Área apróximada de cada célula (km^2)	28,774.215	17,157.570	1,041.710	261.675

Tabela 1: Número de células e a sua área aproximada em grelhas triangulares de diferentes resoluções.

damente estes modelos para associar, a cada célula da representação, a probabilidade de a mesma ser a melhor classe para um novo documento textual.

De forma resumida, temos que os classificadores desenvolvidos com o ***LingPipe*** tomam as suas decisões com base na probabilidade conjunta de documentos textuais e categorias, usando modelos de linguagem baseados em n-gramas de caracteres ou de termos textuais (i.e., nas nossas experiências, testamos estas duas abordagens diferentes de classificação). A ideia geral envolve estimar uma probabilidade $P(txt|cat)$ para cada categoria cat, estimar uma distribuição multinomial $P(cat)$ sobre as categorias e calcular o logaritmo das probabilidades conjuntas para as categorias e documentos, de acordo com regra de Bayes, produzindo-se assim:

$$\log_2 P(cat, txt) \propto \log_2 P(txt|cat) + \log_2 P(cat) \quad (1)$$

Na fórmula, $P(txt|cat)$ é a probabilidade de ver um determinado texto txt no modelo de linguagem para a categoria cat, e $P(cat)$ é a probabilidade marginal atribuída pela distribuição multinomial sobre as categorias. O livro de Carpenter e Baldwin (2011) apresenta mais detalhes sobre os modelos de linguagem usados para estimar $P(txt|cat)$, e sobre a distribuição multinomial $P(cat)$ sobre as categorias (ou seja, sobre as células da nossa representação da Terra). Esta última multinomial é basicamente estimada utilizando o critério MAP (i.e., *maximum a posteriori probability*) com hipóteses *a priori* aditivas (i.e., *priors* de Dirichlet).

No que diz respeito aos modelos de linguagem baseados em n-gramas de caracteres, temos essencialmente modelos de linguagem generativos com base na regra da cadeia, em que as estimativas são suavizadas através de interpolação linear com modelos de ordem inferior, e onde há uma probabilidade de 1.0 para a soma das probabilidades de todas as sequências de um comprimento especificado. Os nossos modelos baseados em n-gramas de caracteres consideram sequências de 8 caracteres. Quanto aos modelos de linguagem baseados em termos, são capturadas as sequências de termos com um modelo de bi-gramas, e modelados os espaços em branco e os símbolos desconhecidos separadamente. A segmentação dos textos em termos é feita através do um método disponível no software ***LingPipe***, o qual utiliza regras comuns a diferentes línguas indo-europeias, semelhantes às regras consideradas no MUC-6[4]. Um termo é assim definido como uma sequência de caracteres satisfazendo um dos seguintes padrões, enquanto que os espaços em branco (i.e., os separadores entre termos) correspondem a sequências de símbolos onde se incluem os espaços, tabulações e mudanças de linha:

- Termos alfa-numéricos. i.e. sequências de letras ou de dígitos;
- Termos numéricos. i.e. sequências de números, virgulas, e pontos;
- Hífens, i.e. sequências de um ou mais hífens;
- Igualdades, i.e. sequências de um ou mais símbolos de igualdade;
- Duplas-aspas, i.e. diferentes formas de representar duplas-aspas nos textos.

O leitor pode consultar o livro de Carpenter e Baldwin (2011) para obter informações mais detalhadas sobre o método de classificação que é aqui usado.

Depois de termos probabilidades atribuídas a cada uma das células na nossa representação da Terra, calculamos as coordenadas geoespaciais de latitude e longitude, com base nas coordenadas centroide para a(s) célula(s) mais provável(eis). Nesta fase em particular, testámos quatro diferentes técnicas de pós-processamento dos resultados:

1. Atribuir coordenadas geoespaciais com base no centroide da célula mais provável.
2. Atribuir coordenadas geoespaciais de acordo com uma média ponderada das coordenadas centroide para todas as células possíveis, em que os pesos são as probabilidades atribuídas a cada uma das células pelo classificador.
3. Atribuir coordenadas geoespaciais de acordo com uma média ponderada das coordenadas centroide para a célula mais provável e para as suas vizinhas adjacentes na grelha triangular, novamente usando como pesos as probabilidades atribuídas a cada uma das células.

[4] `http://cs.nyu.edu/faculty/grishman/muc6.html`

4. Atribuir coordenadas geoespaciais de acordo com uma média ponderada das coordenadas associadas aos knn documentos mais semelhantes nos dados de treino, filtrados de acordo com a pertença das suas coordenadas à célula mais provável descoberta pelo classificador.

Os métodos dois e três da enumeração anterior exigem que o classificador retorne probabilidades bem calibradas sob as classes possíveis, enquanto que a abordagem baseada em modelos de linguagem, utilizada nas nossas experiências, é conhecida por produzir estimativas de probabilidade distorcidas e muito extremas. Na literatura de aprendizagem automática, existem muitos métodos para calibrar as probabilidades retornadas por métodos de classificação, mas a maioria desses métodos são definidos apenas para problemas de classificação binários (Gebel e Weihs, 2007). No nosso problema particular de classificação multi-classe, optamos por processar os valores retornados pelos classificadores baseados em modelos de linguagem através de uma função sigmoide da forma $(\sigma \times score)/(\sigma - score + 1)$, onde o parâmetro σ que controla o gradiente da curva foi ajustado empiricamente.

No que diz respeito ao quarto método de pós-processamento, nós medimos a semelhança entre os documentos de acordo com semelhança do cosseno, entre os vectores de características que os representam. As características correspondem à frequência de ocorrência de uni-gramas de termos. Nas nossas experiências, variou-se o parâmetro knn entre os valores de cinco e vinte documentos.

Embora os classificadores baseados em modelos de linguagem possam ser usados directamente para atribuir documentos às células mais prováveis, eles na prática são muito ineficientes quando se considera uma resolução fina, devido ao número elevado de classes – ver a Tabela 1 – e devido à necessidade de estimar, para cada documento, a sua probabilidade de ter sido gerado pelo modelo de linguagem correspondente a cada classe. Neste trabalho, propomos usar uma abordagem de classificação hierárquica, onde em vez de um classificador único considerando todas as células de uma grelha triangular detalhada, codificando a superfície da Terra, usamos uma hierarquia de classificadores com dois níveis. O primeiro nível corresponde a um modelo único de classificação utilizando células geradas com uma divisão grosseira da superfície terrestre. O segundo nível corresponde a classificadores diferentes, um para cada classe do primeiro nível, codificando diferentes partes da Terra com uma resolução mais elevada. Com este esquema hierárquico, a classificação pode ser feita com muito mais eficiência, uma vez que os documentos precisam de ser avaliados com menos modelos de linguagem. Ainda no que diz respeito a classificação hierárquica, nós também tiramos partido das propriedades da técnica HTM, de modo a reduzir o número de classes em cada um dos modelos gerados no último nível de hierarquia de classificação. Recursivamente, verificamos se uma dada célula não contém quaisquer documentos de treino atribuídos na resolução actualmente considerada, e se apenas uma das células vizinhas na grelha triangular contém documentos. Nestes casos, usamos uma única classe com base na grelha triangular com a resolução imediatamente menor, como forma de representar a região no modelo de classificação.

Num trabalho relacionado anterior focado na língua Inglesa, Wing e Baldridge (2011) relataram resultados muito precisos (i.e., um erro de previsão mediano de apenas 11.8 km, e um erro médio de 221 km), com uma abordagem de classificação semelhante, embora não hierárquica, baseada na divergência de Kullback-Leibler entre modelos de linguagem. No entanto, estes autores também afirmam que uma execução completa de todas as suas experiências (i.e., seis estratégias diferentes) necessitou de cerca de 4 meses em tempo de computação num processador Intel Xeon E5540 de 64-bit, utilizando cerca de 10-16 GB de RAM. A abordagem de classificação hierárquica permite reduzir substancialmente o esforço computacional exigido, tendo-se que as nossas experiências se realizaram em hardware semelhante durante apenas alguns dias.

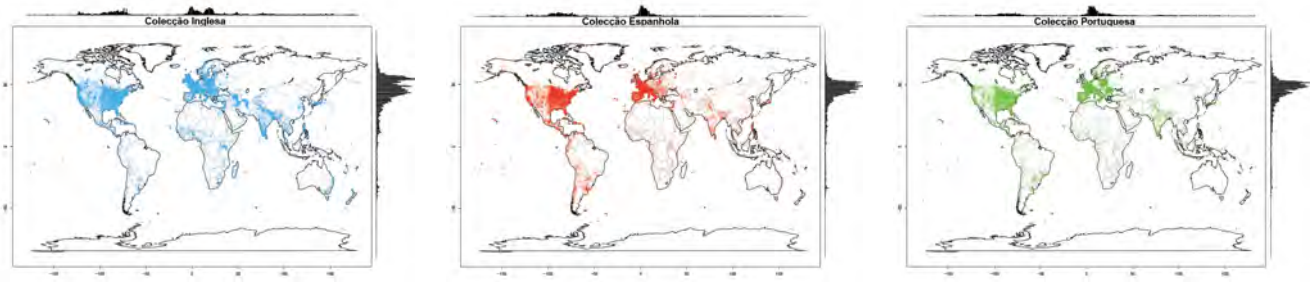
4 Avaliação Experimental

Vamos agora descrever a metodologia de avaliação experimental utilizada para comparar os métodos propostos, discutindo depois os resultados obtidos. Nas experiências aqui relatadas, foram utilizados artigos textuais das versões Inglesa, Espanhola e Portuguesa da Wikipédia, extraídas de *dumps* produzidos em 2012 (i.e., os *dumps* de 2012-06-01 no caso das Wikipédias Inglesa e Portuguesa, e o *dump* de 2012-05-15 no caso da Wikipédia Espanhola). Incluem-se nestas amostras um total de 393,294, 119,572 e 96,643 artigos, respectivamente em Inglês, Espanhol e Português, os quais se encontram associados a coordenadas de latitude e longitude. Estudos anteriores já demostraram que os artigos da Wikipédia são uma fonte adequada de conteúdos textuais georreferenciados para este tipo de testes (Overell, 2009; Wing e Baldridge, 2011).

Temos especificamente que foram processados todos os documentos dos *dumps* da Wikipédia usando o software *dmir-wiki-parser*[5], por forma a extrair o texto dos artigos, e por forma a extrair também as coordenadas geo-espaciais, usando padrões manual-

[5] `http://code.google.com/p/dmir-wiki-parser/`

Figura 3: Mapas tematicos representativos das distribuições geográficas dos documentos da Wikipédia.



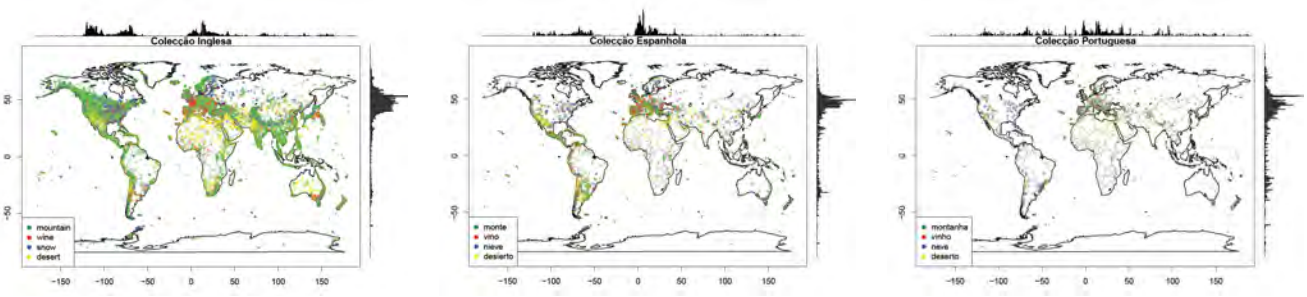
mente definidos para capturar alguns dos múltiplos modelos e formatos usados para expressar latitude e longitude na Wikipédia (i.e., valores situados na *infobox* de cada página). Considerando uma ordem aleatória para os artigos processados desta forma, cerca de 90% dos artigos em cada língua foram utilizados no treino de modelos de classificação (i.e., um total de 353,294 artigos em Inglês, 89,572 em Espanhol, e 77,314 em Português) e os outros 10% foram utilizados para a validação dos métodos propostos (isto é, um total de 40,000, 30,000 e 19,329 artigos, respectivamente em Inglês, Espanhol e Português). A Tabela 2 apresenta uma caracterização estatística para os conjuntos de documentos considerados, enquanto que a Figura 3 ilustra a distribuição geoespacial dos locais associados com os documentos nas três diferentes colecções. Pode-se observar que algumas regiões geográficas (por exemplo a América do Norte ou a Europa) são consideravelmente mais densas em termos de associações a documentos do que outras (por exemplo, África). Verificamos ainda que na colecção Portuguesa existe uma maior concentração de artigos na Europa e na América do Sul (i.e., no Brasil), e que na colecção Espanhola existe uma maior concentração de artigos na Europa e nos países latinos da América do Sul. Além disso, temos que os oceanos e outras grandes massas de água são escassos em associações a documentos da Wikipédia. Isto implica que o número de classes que têm de ser consideradas nos nossos modelos de classificação é muito menor do que os números teóricos de classes apresentados na Tabela 1. No nosso conjunto de dados em Inglês, existe um total de 1,123 células contendo associações para documentos numa resolução de nível 4, e um total de 8,320, 42,331 e 144,693 células, respectivamente quando considerando resoluções de 6, 8 e 10. Estes números são significativamente inferiores no caso das colecções em Espanhol e Português, com apenas 928 e 886 células contendo associações para documentos numa resolução de nível 4, respectivamente no caso das colecções Espanhola e Portuguesa. Ver Tabela 4 onde se apresenta o número de células diferentes em cada colecção, junto com o número médio de documentos de treino por cada célula.

Importa referir que a associação de coordenadas geoespaciais aos documentos das Wikipédias Portuguesa e Espanhola levantou alguns problemas, dado que apenas um número de reduzido destas páginas contém menções explícitas a coordenadas nas suas *infoboxes*. Como forma de contornar esta limitação, utilizámos os links existentes entre as páginas nas várias línguas da Wikipédia, associando assim as coordenadas geoespaciais existentes para as páginas da Wikipédia Inglesa, às páginas equivalentes nas versões Portuguesa e Espanhola. Temos assim que muitos dos documentos usados nas 3 línguas diferentes se referem na prática aos mesmos conceitos e entidades do mundo real. Especificamente nas colecções referentes às Wikipédias Portuguesa e Espanhola, e no total dos documentos usados para treino e teste, temos respectivamente que 62,973 e 81,181 dos documentos são referentes a conceitos também existentes na colecção da Wikipédia Inglesa (i.e., as coordenadas geoespaciais são exactamente iguais, muito embora as descrições textuais sejam diferentes). No total, temos que 50,322 dos documentos considerados são referentes a conceitos partilhados entre as três colecções da Wikipédia.

Como forma inicial de validar a hipótese de que os termos textuais podem ser indicativos de localizações geográficas específicas, filtramos primeiro os documentos de acordo com a ocorrência de termos particulares. De seguida, representámos esses documentos num mapa. A Figura 4 mostra a incidência geográfica de termos textuais diferentes nas suas traduções para as três linguas consideradas, nomeadamente os termos, *montanha*, *vinho*, *neve*, e *deserto*. As figuras mostram que estes termos particulares são mais associados às regiões que seriam esperadas (i.e., termos como *vinho* estão mais associados a regiões como França, ou termos como *deserto* estão mais associados ao Norte de África).

Usando os três conjuntos de documentos da Wikipédia, fizemos experiências com modelos de classificação considerando diferentes níveis de resolução para as células. A Tabela 3 apresenta os resultados obtidos para alguns dos diferentes métodos em estudo (ou seja, para os dois tipos de classifica-

Figura 4: Mapas tematicos representativos das distribuições geográficas de certos termos.



Wikipédia EN	Treino	Teste
Num. Documentos	390,032	40,000
Num. Termos	160,508,876	16,696,639
Média Termos/Doc.	411	417
St.Dev. Termos/Doc.	875.517	901.215

Wikipédia ES	Treino	Teste
Num. Documentos	89,572	30,000
Num. Termos	29,633,769	9,788,169
Média Termos/Doc.	330	326
St.Dev. Termos/Doc.	1016.289	960.214

Wikipédia PT	Treino	Teste
Num. Documentos	77,314	19,329
Num. Termos	13,897,992	3,433,134
Média Termos/Doc.	179	179
St.Dev. Termos/Doc.	615.192	615.250

Tabela 2: Caracterização dos conjuntos de dados da Wikipédia usados na avaliação experimental.

dores, e considerando as três primeiras estratégias de pós-processamento, em que não se usam documentos similares), mostrando os valores de erro para cada tamanho de célula. Os erros de previsão apresentados na Tabela 3 correspondem à distância em Kilómetros, calculada através das formulas de Vincenty[6], com base nas coordenadas estimadas e nas coordenadas indicadas na Wikipédia. Os valores de exactidão correspondem ao rácio entre o número de classificações correctas (ou seja, aquelas onde a célula mais provável contém as verdadeiras coordenadas geoespaciais de latitude e longitude, tal como associadas ao documento) e o número de classificações efectuado. Os valores $k1$ e $k2$ correspondem à resolução usada na representação da Terra, para cada nível do classificador hierárquico.

Os valores da Tabela 3 mostram que o método de classificação proposto obtém melhores resultados com o aumento do número de documentos de treino, tendo-se que os resultados são um pouco melhores no caso da colecção em Inglês, em comparação com os resultados para as colecções em Espanhol e Português. O método correspondente ao uso de modelos de linguagem baseados em n-gramas de caracteres, utilizando uma resolução do segundo nível de 8 (i.e., áreas de classificação de 1041 Km^2) obteve os melhores resultados, com uma exactidão de cerca de 0.4 na tarefa de encontrar a célula correta, no caso da colecção em Inglês, enquanto que a atribuição de coordenadas geoespaciais aos documentos teve um erro de 268 Kilómetros, em média, também no caso da colecção em Inglês. No caso concreto desse teste, os documentos que foram atribuídos à célula correta foram associados a coordenadas que se encontravam a uma distância média de 14 Kilómeteos para com as coordenadas corretas. Podemos também observar que para uma resolução 10 (i.e., para uma área de classificação de 262 Km^2), os resultados pioram substancialmente, provavelmente devido ao reduzido número de documentos de treino associado a cada célula do modelo, como demonstrado na Tabela 4. Os resultados da Tabela 3 mostram ainda que tanto a segunda como a terceira técnica de pós-processamento melhoram geralmente os resultados da geocodificação sobre o método base em que se atribuem as coordenadas do ponto centróide da célula mais provável. No entanto, os resultados mostram apenas uma ligeira melhoria para esta técnica, e acreditamos que isto se deve ao facto dos nossos classificadores, baseados em modelos de linguagem, não fornecerem estimativas de probabilidade precisas e bem calibradas, tendo-se que a nossa técnica de calibração baseada num pós-processamento dos valores, através de uma função sigmoide, continua a produzir resultados demasiado extremos.

A Tabela 5 apresenta os resultados obtidos com modelos de linguagem baseados em n-gramas de caracteres (ou seja, com o melhor método de acordo com a experiência anterior), quando se utiliza o quarto método de pós-processamento, em que se atribuem coordenadas de latitude e longitude através do ponto centroide das coordenadas associadas aos knn documentos mais semelhantes, contidos dentro da célula mais provável para cada documento. A pri-

[6] `http://en.wikipedia.org/wiki/Vincenty`

Método	Resolução k1	Resolução k2	Exactidão do Classificador 1º Nível	Exactidão do Classificador 2º Nível	Distância Geoespacial Centroide Média	Distância Geoespacial Centroide Mediana	Distância Geoespacial Todas as Células Média	Distância Geoespacial Todas as Células Mediana	Distância Geoespacial Células Vizinhas Média	Distância Geoespacial Células Vizinhas Mediana
Documentos da Wikipédia Inglesa										
N-gramas caracteres	0	4	**0.9609**	**0.8354**	405.214	240.017	438.762	228.829	386.271	219.379
	1	6	0.9411	0.6669	**254.846**	62.846	282.874	71.119	257.741	65.551
	2	8	0.9283	0.3989	268.480	**25.757**	283.761	48.039	269.493	28.569
	3	10	0.8912	0.1615	281.669	30.405	287.909	51.595	281.755	30.464
N-gramas termos	0	4	0.9444	0.5209	693.764	240.017	544.543	228.829	691.899	219.379
	1	6	0.9103	0.4244	433.224	92.770	729.546	303.055	444.766	120.561
	2	8	0.8909	0.2747	455.025	44.621	572.406	191.522	457.858	50.349
	3	10	0.8297	0.1305	547.760	42.162	591.111	123.457	547.996	41.982
Documentos da Wikipédia Espanhola										
N-gramas caracteres	0	4	**0.9526**	**0.7725**	405.379	233.200	403.748	232.134	404.222	232.099
	1	6	0.9162	0.5890	281.810	68.098	280.973	67.570	281.616	67.887
	2	8	0.9015	0.3339	280.293	30.629	**279.639**	**30.293**	280.241	30.480
	3	10	0.8495	0.0948	327.278	43.311	326.754	43.122	327.272	43.311
N-gramas termos	0	4	0.8878	0.6450	729.810	263.677	729.015	262.576	729.158	262.575
	1	6	0.8249	0.4867	559.329	82.117	558.906	81.726	559.198	81.940
	2	8	0.8049	0.3015	570.181	39.035	569.633	38.429	570.112	38.824
	3	10	0.7379	0.1082	692.899	36.376	692.416	35.611	692.874	36.317
Documentos da Wikipédia Portuguesa										
N-gramas caracteres	0	4	**0.9470**	**0.7680**	381.253	231.011	378.907	229.360	379.165	229.378
	1	6	0.8966	0.5172	264.155	74.424	**262.699**	73.361	263.714	73.918
	2	8	0.8768	0.2392	275.398	46.781	274.166	**45.917**	275.340	46.533
	3	10	0.8303	0.0748	286.686	52.098	286.050	51.472	286.681	52.057
N-gramas termos	0	4	0.8863	0.6674	584.522	252.107	582.190	250.372	582.847	250.485
	1	6	0.8164	0.4457	460.500	86.201	459.485	85.350	460.140	85.593
	2	8	0.8003	0.2317	454.941	50.158	453.643	48.862	454.853	50.032
	3	10	0.7438	0.0886	491.483	48.462	490.463	46.906	491.469	48.457

Tabela 3: Resultados obtidos na geocodificação de documentos com diferentes classificadores.

	Resolução HTM			
Wikipédia EN	4	6	8	10
Num. Células	1,107	8,404	42,907	146,498
Média Docs/Célula	352	46	9	3

	Resolução HTM			
Wikipédia ES	4	6	8	10
Num. Células	928	5,105	18,655	52,348
Média Docs/Célula	97	18	5	2

	Resolução HTM			
Wikipédia PT	4	6	8	10
Num. Células	886	4,638	17,772	48,692
Média Docs/Célula	96	18	5	2

Tabela 4: Número de documentos de treino disponível para cada célula do método HTM.

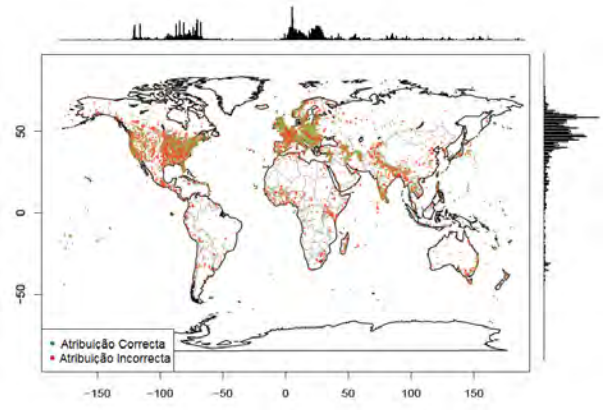


Figura 5: Localizações estimadas para os documentos de teste da Wikipédia Inglesa.

meira coluna da Tabela 5 indica o número de documentos semelhantes que foi considerado, enquanto que os valores de $k2$ correspondem à resolução para a representação da Terra, utilizada no segundo nível do classificador hierárquico.

A melhor configuração para as três linguas corresponde a uma resolução de 8. Os resultados mostram ainda que usando os 5 documentos mais similares se obtêm melhores resultados para as colecções Inglesa e Espanhola, com uma distância média de 265 e 278 Kilómetros, respectivamente, e uma distância mediana de 22 e 29 Kilómetros, respectivamente. Na colecção Portuguesa obtêm-se melhores resultados usando os 10 documents mais similares, com uma distância média de 273 Kilómetros e uma distância mediana de 45 Kilómetros.

Uma visualização dos resultados obtidos com o melhor método de geocodificação encontra-se apresentada na Figura 5, em que o mapa representa a distribuição geoespacial das localizações previstas para os documentos da colecção em Inglês. A figura mostra que os erros são uniformemente distribuídos, e mostra ainda que a Europa e a América do Norte

knn	k1	k2	Inglês Média	Inglês Mediana	Espanhol Média	Espanhol Mediana	Português Média	Português Mediana
5	0	4	289.750	90.515	290.021	77.583	260.805	76.665
	1	6	235.702	34.005	260.862	34.695	244.363	46.324
	2	8	265.734	**22.315**	277.895	**27.865**	273.218	44.612
	3	10	281.442	30.209	327.198	43.273	286.649	51.972
10	0	4	274.515	68.252	279.100	60.258	249.039	58.964
	1	6	233.982	32.092	**260.049**	33.824	**243.443**	44.839
	2	8	265.655	22.371	278.205	28.266	273.355	**44.587**
	3	10	281.460	30.208	327.215	43.324	286.685	51.960
15	0	4	271.045	64.008	277.652	59.384	247.373	55.599
	1	6	**233.928**	32.243	260.666	35.257	243.880	45.133
	2	8	265.744	22.480	278.511	28.723	273.600	44.896
	3	10	281.464	30.172	327.211	43.298	286.689	51.961
20	0	4	270.337	63.373	278.069	60.767	247.886	55.217
	1	6	234.197	32.687	261.367	36.155	244.437	45.485
	2	8	265.869	22.640	278.734	28.899	273.797	45.109
	3	10	281.466	30.170	327.213	43.298	286.689	51.961

Tabela 5: Resultados obtidos na geocodificação de documentos com o método de pós-processamento baseado na utilização dos knn documentos de treino mais similares.

continuam a ser as regiões de maior densidade.

A Figura 6 ilustra a distribuição para os erros produzidos pelos classificadores baseados em n-gramas de caracteres, em termos da distância entre as coordenadas estimadas e as coordenadas verdadeiras, quando se utiliza o método de pós-processamento considerado como *baseline*, e o método que utiliza as coordenadas dos knn documentos mais similares, para as três linguas. Estes gráficos apresentam o número de documentos em que o erro (i.e., a distância) é maior ou igual do que um dado valor, utilizando eixos logarítmicos. A Figura 6 mostra que o método de pós-processamento com base na análise dos documentos mais semelhantes atribui coordenadas à maioria dos exemplos com um pequeno erro em termos de distância, e com apenas cerca de 100 documentos correspondendo a um erro maior do que 10,000 Kilómetros, no caso da colecção em Inglês. Piores resultados são apresentados para o método base, com cerca de 200 documentos em que se observa um erro maior do que 10,000 Kilómetros nas coordenadas previstas, mais uma vez no caso da colecção de Wikipédia em Inglesa.

Finalmente, na Tabela 6 sumarizam-se os resultados para a melhor configuração em cada lingua, os quais foram sempre obtidos com classificadores baseados em n-gramas de caractéres. A Tabela 6 apresenta ainda os valores correspondentes a um intervalo de confiança de 95% para os erros médios e medianos obtidos com a melhor configuração.

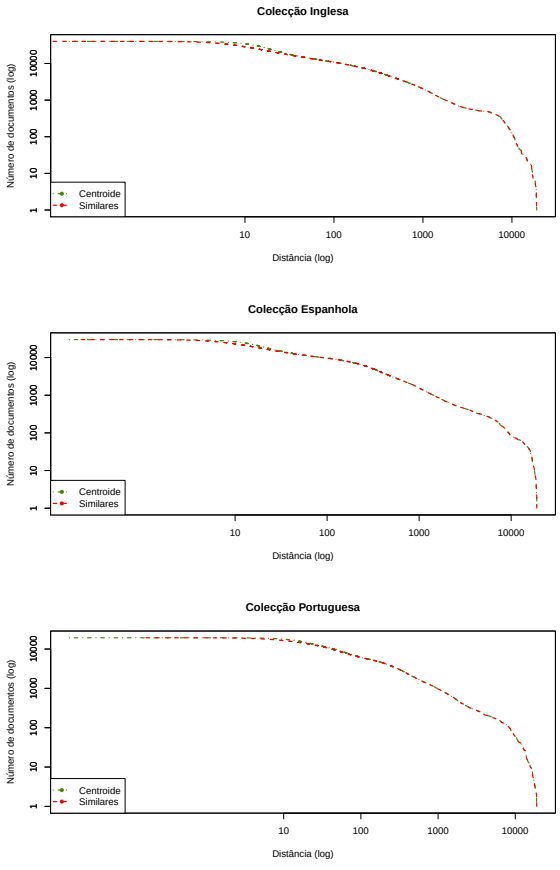


Figura 6: Distribuição dos valores de erro obtidos, em termos da distância geoepacial para com as coordenadas de latitude e longitude correctas.

5 Conclusões e Trabalho Futuro

Este trabalho avaliou diferentes métodos para a geocodificação de documentos textuais, os quais uti-

	Resolução			Distância			
Língua	k1	k2	knn	Média	Média (95%)	Mediana	Mediana (95%)
Inglês	2	8	5	265.734	255.230 - 276.237	22.315	21.859 - 22.771
Espanhol	2	8	5	277.895	265.725 - 290.065	27.865	26.996 - 28.734
Português	2	8	10	273.355	258.838 - 287.871	44.587	43.498 - 45.676

Tabela 6: Sumarização dos melhores resultados obtidos.

lizam classificadores baseados em modelos de linguagem para a atribuição de regiões geo-espaciais aos documentos, fazendo ainda um pós-processamento dos resultados da classificação por forma a atribuir as coordenadas geoespaciais de latitude e longitude mais prováveis. Mostramos que a identificação automática das coordenadas geoespaciais de um documento, com base apenas no seu texto, pode ser feita com alta precisão, utilizando métodos simples de classificação supervisionada, e usando uma representação discreta para a superfície da Terra, com base numa grelha triangular hierárquica. O método proposto é simples de implementar, e tanto o treino como os testes podem ser facilmente paralelizados por forma a processar grandes colecções de documentos. A nossa estratégia de geocodificação mais eficaz utiliza modelos de linguagem baseados em n-gramas de caracteres, e atribui as coordenadas de latitude e longitude através do centroide das coordenadas dos knn documentos de treino mais semelhantes ao documento sob análise, contidos dentro da região mais provável para cada documento.

Importa referir que as experiências relatadas neste artigo foram efectuadas separadamente com três colecções de documentos distintas, nomeadamente em Inglês, Espanhol, e Português. Para trabalho futuro, seria interessante realizar experiências com colecções de documentos ainda noutras linguas, e seria também interessante introduzir um terceiro nível de classificação no método proposto, por forma a que os documentos fossem inicialmente classificados de acordo com a língua, e posteriormente processados com os modelos de geocodificação correspondentes. Assim, e usando por exemplo os dados das várias Wikipédias, seria possível construir um sistema que automáticamente geocodificasse documentos, independentemente da sua língua.

Existem muitas aplicações possíveis para o método de geocodificação de documentos descrito no presente documento. Uma aplicação em particular, que estamos a considerar num trabalho em curso, relaciona-se com o uso das distribuições de probabilidade sobre as células, da nossa representação da Terra, na construção de mapas temáticos que mostrem a incidência geográfica de determinadas construções extraídas dos textos (por exemplo, mapas mostrando a distribuição geográfica das opiniões expressas em relação a determinados temas). No entanto, importa reforçar que a abordagem de classificação proposta, com base em modelos de linguagem, não fornece estimativas de probabilidade precisas e bem calibradas para as diferentes classes envolvidas no problema, focando-se apenas na tarefa mais simples de prever qual a classe mais provável. Para trabalho futuro, em lugar de usarmos um método heurístico de calibração com base no pós-processamento dos valores retornados pelo classificador, gostaríamos de experimentar outras abordagens de classificação para a atribuição da(s) célula(s) mais provável aos documentos, tais como por exemplo modelos de máxima entropia (i.e., regressão logística). Também gostaríamos de experimentar com modelos de máxima entropia usando restrições nas expectativas especificando afinidades entre os termos e as classes (Druck, Mann e McCallum, 2008), ou com modelos utilizando regularização *à posteriori* (Ganchev et al., 2010), aproveitando o facto de que a presença de palavras correspondentes a nomes de locais deve ser vista como um forte indicador para que o documento pertença a uma determinada classe. Ainda no que se refere a nomes de locais, importa notar que, muito embora a identificação de coordenadas geoespaciais para a totalidade de um documento possa fornecer uma forma conveniente de ligar textos a locais específicos, útil para diferentes aplicações, existem muitas outras aplicações que poderiam beneficiar da resolução completa das referências a locais individuais nos documentos (Leidner, 2007). As distribuições de probabilidade para as células, fornecidas pelo método de classificação, podem por exemplo ser usadas para definir uma confiança prévia na resolução de nomes de locais.

Outra possibilidade de trabalho futuro relaciona-se com o uso de um meta-algoritmo originalmente proposto para problemas de regressão ordinal (Pang e Lee, 2005), ou seja, para problemas onde temos uma ordem natural entre os possíveis resultados, tais como o nosso problema em que temos uma noção de distância entre as células possíveis. A ideia básica deste método é a de que dados semelhantes devem receber classificações semelhantes. Usando esta premissa, podemos corrigir os resultados fornecidos pelos classificadores, considerando as classes reais dos knn documentos de treino mais semelhantes, obtidos através da semelhança do cosseno sobre os vec-

tores de características, e utilizando as fórmulas de Vincenty para medir a similaridade entre as classes (isto é, a distância entre as coordenadas centroide associadas às células). O classificador poderia ser usado como uma função inicial de preferência $\pi(x, l)$ que desse uma estimativa sobre a forma de classificar os documentos (i.e., que desse pontuações de classificação para um documento x e uma classe l). Essencialmente, iríamos utilizar uma métrica de distância entre as etiquetas d, e um conjunto de documentos $knn(x)$ com os knn exemplos mais próximos do documento x, de acordo com uma função de similaridade $sim(x, y)$ entre pares de documentos x e y. O problema da classificação de documentos pode ser resolvido através da escolha das classes que minimizem a fórmula abaixo, onde α representa um parâmetro de combinação ajustado empiricamente:

$$\sum_{x \in teste} \left[-\pi(x, l) + \alpha \sum_{y \in knn(x)} d(l_x, l_y) sim(x, y) \right] \quad (2)$$

Finalmente, gostaríamos também de experimentar com técnicas de expansão de documentos, especialmente para documentos pequenos, por forma a construir pseudo-documentos através da concatenação dos conteúdos relacionados com os documentos originais (por exemplo, utilizando as hiperligações entre documentos). Importa no entanto referir que estamos principalmente interessados na geocodificação de documentos usando apenas o texto, pois existe um grande número de situações (e.g., documentos históricos em bibliotecas digitais) em que outros tipos de informação simplesmente não estão disponíveis.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia (FCT), através dos projetos com referências PTDC/EIA-EIA/109840/2009 (SInteliGIS), UTA-Est/MAI/0006/2009 (REACTION), e PTDC/EIA-EIA/115346/2009 (SMARTIES). O autor Ivo Anastácio foi ainda suportado por uma bolsa de doutoramento com referência SFRH/BD/71163/2010.

Gostaríamos também de agradecer a Pável Calado, Luísa Coheur e Mário J. Silva, pelos seus comentários a versões preliminares deste trabalho.

Referências

Adams, B. e K. Janowicz. 2012. On the geo-indicativeness of non-georeferenced text. Em *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Amitay, E., N. Har'El, R. Sivan, e A. Soffer. 2004. Web-a-where: geotagging web content. Em *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*.

Anastácio, I., B. Martins, e P. Calado. 2010. A comparison of different approaches for assigning geographic scopes to documents. Em *Proceedings of the 1st Simpósio de Informática*.

Bär, Hans e Lorenz Hurni. 2011. Improved density estimation for the visualisation of literary spacesl. *The Cartographic Journal*, 48.

Carpenter, Bob e Breck Baldwin. 2011. *Natural Language Processing with LingPipe 4*. LingPipe Publishing, draft edition.

Druck, Gregory, Gideon Mann, e Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. Em *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in Information Retrieval*.

Dutton, G. 1996. Encoding and handling geospatial data with hierarchical triangular meshes. Em M. J. Kraak e M. Molenaar, editores, *Advances in GIS Research II*.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, e Eric P. Xing. 2010. A latent variable model for geographic lexical variation. Em *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Erdmann, Eva. 2011. Topographical fiction: A world map of international crime fiction. *The Cartographic Journal*, 48.

Ganchev, Kuzman, João Graça, Jennifer Gillenwater, e Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11.

Gebel, Martin e Claus Weihs. 2007. Calibrating classifier scores into probabilities. Em *Proceedings of Advances in Data Analysis*.

Johnstone, B. 2010. Language and place. Em R. Mesthrie e W. Wolfram, editores, *Cambridge Handbook of Sociolinguistics*.

Kumar, Abhimanu, Matthew Lease, e Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. Em *Proceeding of the 20th ACM conference on Information and Knowledge Management*.

Leidner, J. 2007. *Toponym Resolution in Text*. Tese de doutoramento, University of Edinburgh.

Lieberman, Michael D. e Hanan Samet. 2011. Multifaceted toponym recognition for streaming news. Em *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*.

Martins, B., I. Anastácio, e P. Calado. 2010. A machine learning approach for resolving place references in text. Em *Proceedings of the 13th AGILE International Conference on Geographic Information Science*.

Mehler, Andrew, Yunfan Bao, Xin Li, Yue Wang, e Steven Skiena. 2006. Spatial analysis of news sources. *IEEE Transactions Visualization in Computer Graphics*, 12.

Overell, Simon. 2009. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. Tese de doutoramento, Imperial College London.

Pang, Bo e Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL.

Serdyukov, Pavel, Vanessa Murdock, e Roelof van Zwol. 2009. Placing flickr photos on a map. Em *Proceedings of the 32nd international ACM SIGIR conference on Research and development in Information Retrieval*.

Shakhnarovich, Gregory, Trevor Darrell, e Piotr Indyk. 2006. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing)*. The MIT Press.

Szalay, Alexander S., Jim Gray, George Fekete, Peter Z. Kunszt, Peter Kukol, e Ani Thakar. 2005. Indexing the sphere with the hierarchical triangular mesh. Relatório Técnico MSR-TR-2005-123, Microsoft.

Wing, Benjamin e Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.