XVI Congreso Nacional de Tecnologías de la Información Geográfica
25, 26 y 27 de Junio de 2014. Alicante.

Clasificación geográfica de textos informales

Fernando S. Peregrino[a*], David Tomás[a], Fernando Llopis[a]

[a]*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Alicante*

Resumen

La clasificación automática de textos es ampliamente conocida y usada en multitud de situaciones. Dicha clasificación puede ser afrontada desde distintos puntos de vista, siendo uno de los más usados la clasificación por ámbito geográfico.

De sobra son conocidas las clasificaciones geográficas realizadas por los buscadores de Internet y por los periódicos, las cuales agrupan un conjunto de páginas web o noticias acorde al ámbito geográfico que cubren. Por otro lado, la vigorosa aparición de las redes sociales con su lenguaje altamente informal ha hecho que las técnicas utilizadas para la clasificación geográfica automática de textos formales tengan que ser readaptadas con el propósito de obtener unos mejores resultados.

En este artículo presentamos el estado actual en este campo. Así como un estudio realizado sobre la utilización de las técnicas más empleadas en la clasificación de textos informales combinando dichas técnicas con recursos de distinta índole y formalidad.

Palabras clave: Clasificación de textos; Detección del foco geográfico en textos; Procesamiento del Lenguaje Natural; Recuperación de información geográfica; Modelos de lenguaje; Redes sociales.

* E-mail : fsperegrino@dlsi.ua.es

1. Introducción

Debido a la abrumadora cantidad de información que se procesa hoy en día a través de los medios digitales, hay un gran interés en la clasificación de dicha información acorde a un determinado tipo de características, tales como los términos que aparecen en el texto dado.

La clasificación de textos en distintas categorías es una tarea ampliamente tratada en la literatura del Procesamiento del Lenguaje Natural (PLN). Dichas categorías suelen obedecer a diversas necesidades, dependiendo de la utilidad que se le pretenda dar y de la naturaleza del corpus, siendo una de las más extendidas la clasificación de noticias de periódicos, webs, etc., en sus diversas secciones acorde a la temática tratada (política, sucesos, deporte, etc.).

Junto a este tipo de clasificación puede venir asociada una clasificación geográfica, pudiendo así clasificar una noticia como relevante para una determinada zona geográfica mientras que para otra carecería de interés. Por ejemplo, imaginémonos que un municipio ha aprobado la creación de una nueva línea de transporte público para cubrir un recorrido urbano. Dicha noticia sería claramente relevante en el ámbito geográfico de dicho municipio, e incluso en el de los municipios próximos a éste, ya que podría haber vecinos de las localidades próximas que se beneficiasen de dicho transporte, pero la noticia perdería relevancia conforme nos alejamos del foco geográfico de la misma.

Uno de los claros beneficiados de estas clasificaciones son los motores de búsqueda, por ejemplo los que podemos encontrar por internet, Google, Yahoo!, Bing, etc., ya que en cierta medida basan la recuperación de textos relacionados con la consulta formulada en una clasificación o etiquetación previa de los textos (páginas web en este caso).

Dado que la clasificación de este tipo de textos sería inviable si se realizase por seres humanos, se han desarrollado un amplio número de técnicas para tratar con esta problemática de una manera automática o semiautomática.

La manera más común de abordar dicha problemática de forma automática es mediante técnicas de aprendizaje automático (ML: Machine Learning), y dentro de estas técnicas nos encontramos con los modelos de lenguaje (LM: Language Modeling) como una de las técnicas más extendidas.

1.1. Foco geográfico

El foco geográfico de un documento identifica el lugar o lugares en los que se centra el contenido del texto. Cuando se trata de encontrar el foco geográfico en un texto, hay varias aproximaciones diferentes. Una de las aproximaciones predominantes es el uso de técnicas de aprendizaje automático (ML), tal y como se describe en (F. S. Peregrino et al., 2012).

En (E. Amitay et al., 2004) también se puede ver como los autores hacen uso de recursos externos, en este caso diccionarios geográficos (gazetteers).

Si nos centramos en lenguajes más informales, en (T. Qin et al., 2010) podemos ver como los autores utilizaron técnicas de LM para localizar los lugares mencionados en un conjunto de blogs de viaje.

En (I. Anastácio, 2009) se hizo un estudio comparativo entre 4 sistemas de detección del foco geográfico: Yahoo! Placemaker, Web-a-Where (E. Amitay et al., 2004), GIPSY (A. Woodruff, 1994) y GREASE (B. Martins & M. J. Silva, 2005). El sistema ganador fue el desarrollado en (E. Amitay et al., 2004).

Por otro lado, la creciente popularidad de los servicios de "microblogging", representados por redes sociales tales como Twitter, Flickr o Foursquare, ha creado un nuevo escenario para las tecnologías del PLN,

emergiendo así un gran interés entre la comunidad científica en el campo de la detección geográfica en las redes sociales (principalmente en Twitter). En este nuevo escenario, miles de millones de comentarios (conocidos como 'tweets' en el caso de Twitter) son lanzados cada semana, mayoritariamente en un lenguaje informal.

1.2. Twitter

Twitter es la red social y servicio de microblogging más extendido que permite a sus usuarios enviar y leer mensajes de texto de hasta 140 caracteres, conocidos como tweets.

De acuerdo con su presidente ejecutivo, Dick Costolo, desde julio de 2006, después del lanzamiento de la red social que preside, el número de usuarios de dicha comunidad ha crecido exponencialmente hasta alcanzar actualmente (marzo de 2014) 241 millones de usuarios activos alrededor del mundo, los cuales lanzan unos 500 millones de tweets por día, expresando sus opiniones, dudas y necesidades. Alrededor del 76% de estos usuarios activos utilizan su móvil para conectarse a Twitter, haciendo posible, siempre que el usuario lo autorice, obtener las coordenadas geográficas del usuario y de los tweets enviados.

El uso de los dispositivos móviles en las redes sociales abre un amplio abanico de posibilidades de negocio relacionadas con la ubicación del usuario, sus gustos, etc. Dado que los usuarios de las redes sociales normalmente no habilitan los dispositivos GPS en sus teléfonos móviles, o simplemente no autorizan a las aplicaciones de las propias redes sociales a obtener su ubicación, resulta de crucial importancia el análisis de los comentarios que vierten con el fin de detectar la ubicación de estos usuarios.

Dado que el lenguaje empleado en este tipo de redes sociales es altamente informal, difiriendo así del que nos podemos encontrar en otro tipo de textos como el de noticias de periódicos, es necesario el uso de nuevas técnicas que nos permitan clasificar correctamente estos textos.

La inmensa mayoría del trabajo llevado a cabo en la detección del foco geográfico en Twitter se ha centrado en la información facilitada por los usuarios con sus tweets y con los lugares que han indicado en el campo de "Ubicación" de su perfil.

Así pues en (B. Hecht et al., 2011; J. Mahmud et al., 2012; S. Kinsella et al., 2011) se trató de inferir la ubicación de los usuarios mediantes técnicas de ML.

A diferencia de las aproximaciones previas, en el sistema descrito en este artículo hemos hecho uso tanto de lenguaje informal (Twitter) como del formal (Wikipedia) para la detección de la ubicación de los usuarios de Twitter.

El resto de este artículo está estructurado de la siguiente forma: En la sección 2 se describe el conjunto de datos sobre el que se ha trabajado así como la manera de la que se ha obtenido. En esta sección también se expondrán los experimentos que se han llevado a cabo. En la sección 3 se mostrarán los resultados obtenidos de los experimentos expuestos en la sección anterior. Posteriormente, en la sección 4 se discutirán los resultados de los experimentos expuestos. Finalmente, en la sección 5 se mostrarán las conclusiones que se han podido extraer de esta investigación.

2. Metodología, materiales, datos y herramientas

En esta sección se detallará el corpus utilizado para los experimentos llevados a cabo, y se describirán los experimentos que se realizaron.

2.1. Conjunto de datos

Desde el 20 de abril de 2013 hasta el 10 de junio de ese mismo año, mediante la SEARCH API de Twitter (actualmente integrada en la REST API aunque con algunas limitaciones) obtuvimos tweets georreferenciados en la ciudad más grande de cada una de las 50 provincias españolas, más las dos ciudades autónomas, lo que dio como resultado un total de 4.672.420 tweets provenientes de alrededor de 200.000 usuarios distintos. Puesto que los usuarios pueden enviar tweets desde más de una ciudad, se agruparon todos los tweets realizados por un usuario en una ciudad, es decir, se creó un conjunto distinto de tweets por cada usuario en cada ciudad en la que había "twiteado", obteniendo así un total de 203.495 conjuntos distintos de tweets.

Este corpus fue procesado, pasándolo todo a minúsculas, eliminando los símbolos de puntuación y URLs. Se mantuvieron los símbolos '#' y '@' ya que representa tópicos (conversaciones) y usuarios respectivamente dentro de Twitter, los cuales resultan ser unas características intrínsecas del lenguaje de Twitter.

Por otro lado, cada uno de los conjuntos de tweets del corpus estaba clasificado acorde a la ciudad a la que pertenecía, dando así como resultado 52 categorías distintas, las cuales se dividieron en 10 particiones de tamaño semejante, utilizando una de estas particiones como conjunto de evaluación y las restantes como conjunto de entrenamiento. En la tabla 1 se muestra el conjunto de evaluación obtenido, donde:

- En la primera columna se indica el número mínimo o máximo de *tweets* que tiene cada uno de los conjuntos de *tweets* que se utilizaron en la evaluación.
- En la segunda columna se indican el número de usuarios existentes que cumplen los requisitos de la primera columna.
- En la tercera columna se indica el número total de *tweets* existentes en el conjunto utilizado para la prueba que cumplen los requisitos de la primera columna.

Tabla 1. Conjunto de prueba.

Tweets por usuario	Nº de usuarios	Nº de tweets
Cualquier número de *tweets*	20.326	463.717
Menos de 10 *tweets*	14.107	40.238
Menos de 100 *tweets*	19.287	202.789
Más de 99 *tweets*	1.039	260.928

Con las 9 particiones restantes se han hecho experimentos con distintos tamaños de corpus de entrenamiento, ya que cada una de las particiones fue progresivamente añadiéndosele a la anterior hasta poder entrenar con el corpus completo. El objetivo de hacer estos tamaños distintos en las particiones de entrenamiento es el de poder comprobar cómo varía el rendimiento del algoritmo en función de la cantidad de datos de entrenamiento.

En la tabla 2 se puede ver en la primera columna el número de particiones utilizadas en el entrenamiento, en la segunda columna el número de usuarios que tiene cada uno de esos corpus de entrenamiento y, finalmente, en la tercera columna el número total de *tweets*.

Tabla 2. Conjunto de entrenamiento.

Nº de particiones utilizadas	Nº de usuarios	Nº de *tweets*
1	20.373	456.747
2	40.741	944.783
3	61.104	1.434.277
4	81.464	1.891.823
5	101.819	2.338.837
6	122.165	2.829.805
7	142.507	3.296.387
8	162.843	3.744.898
9	183.169	4.209.703

Puesto que el objetivo principal, y aportación más innovadora de este trabajo, es el comprobar si textos escritos en un lenguaje formal aportan alguna mejoría a la hora de clasificar textos informales, adicionalmente a este corpus de entrenamiento también se ha empleado en algunos experimentos términos procedentes de los artículos de la *Wikipedia* que representan a cada una de las 52 ciudades existentes.

2.2. Experimentos

Principalmente, los experimentos podrían ser clasificados dentro de 3 categorías: los realizados con textos informales (solamente Twitter), los realizados con textos formales (sólo Wikipedia) y los realizados con textos formales e informales (Twitter y Wikipedia).

Para estos experimentos se utilizó la herramienta Lemur Toolkit, la cual utiliza una combinación de modelos de lenguaje, tal y como se describe en (J. M. Ponte & W. B. Croft, 1998), utilizando suavizado Dirichlet (C. Zhai & J. Lafferty, 2004) y clasificado mediante KL-divergence (S. Kinsella et al., 2011), sin la utilización de técnicas de normalización, tales como dejar sólo la raíz de las palabras (stemming), ya que no resulta muy efectivo para textos informales de un tamaño corto.

3. Resultados

En la tabla 3 se pueden apreciar los resultados obtenidos para los distintos tamaños de conjuntos de entrenamiento, así como los distintos conjuntos de pruebas utilizados. En dicha tabla se puede apreciar cómo el número de tweets emitido por los usuarios afecta notablemente al rendimiento del sistema, siendo mejor cuanto más prolífico es el usuario, tal y como se puede apreciar observando el resultado obtenido con los usuarios más activos (100 o más tweets) usando el conjunto de entrenamiento completo (casilla resaltada en negrita).

Tabla 3. Resultados utilizando solamente *Twitter* como entrenamiento.

	Tweets por usuario en una ciudad dada			
Nº particiones	Cualquiera	Menos de 10	Menos de 100	100 o más
1	0.2917	0.2516	0.2854	0.4071
2	0.3322	0.2787	0.3234	0.4928
3	0.3618	0.2984	0.3497	0.5813
4	0.3832	0.3147	0.3695	0.6323
5	0.3978	0.3235	0.3832	0.6660
6	0.4131	0.3371	0.3986	0.6776
7	0.4266	0.3466	0.4118	0.6968
8	0.4343	0.3523	0.4193	0.7084
9	0.4411	0.3571	0.4252	**0.7315**

En la tabla 4 se puede apreciar los resultados obtenidos para los distintos tamaños de conjunto de entrenamiento utilizados junto a los artículos de la *Wikipedia* de las ciudades estudiadas, así como los distintos conjuntos de pruebas utilizados. En dicha tabla se puede apreciar (resaltado en negrita las mejoras más significativas) como la *Wikipedia* aporta una mejora cuando no disponemos de un gran número de *tweets* para evaluar, o cuando no disponemos de un conjunto de entrenamiento muy amplio, tal y como se discute en la siguiente sección.

Tabla 4. Resultados utilizando *Twitter* y *Wikipedia* como entrenamiento.

	Tweets por usuario en una ciudad dada			
Nº particiones	Cualquiera	Menos de 10	Menos de 100	100 o más
1	**0.3112**	**0.2743**	**0.3071**	0.3859
2	0.3375	0.2867	0.3299	0.4764
3	0.3652	0.3027	0.3542	0.5669
4	0.3848	0.3164	0.3717	0.6256
5	0.3993	0.3269	0.3851	0.6583
6	0.4141	0.3378	0.3997	0.6776
7	0.4273	0.3478	0.4125	0.6968
8	0.4364	0.3550	0.4215	0.7084
9	**0.4414**	0.3580	0.4261	0.7209

Los experimentos realizados solamente con los artículos de la *Wikipedia* dieron como resultado: 0.0862, 0.1018, 0.0903, 0.0135, respectivamente para cada una de las 4 categorías mostradas en las tablas 3 y 4. Estos resultados distan de los ya expuestos en las tablas 3 y 4, lo que muestra la escasa aportación que hace la *Wikipedia* por sí sola.

4. Discusión

La pequeña cantidad de información geográfica que se puede encontrar normalmente en un simple tweet hace que la tarea de identificar su foco geográfico sea extremadamente compleja, incluso para los seres humanos (S. Kinsella et al., 2011). Esto explica los pobres resultados obtenidos cuando disponemos de menos de 10 tweets por usuario.

A través de los experimentos llevados a cabo con Twitter y Wikipedia conjuntamente (tabla 4), se ha pretendido mostrar las aportaciones que pueden llegar a hacer los recursos de textos formales a la hora de clasificar geográficamente textos informales cuando disponemos de conjuntos de entrenamiento y/o prueba de distinto tamaño, mostrando una mejora significativa cuando nos encontramos con corpus de entrenamiento y/o de prueba no muy amplios. Conforme crece el tamaño del corpus se puede observar como la mejora obtenida con la inclusión de la Wikipedia disminuye.

5. Conclusiones

Se ha implementado un sistema que ha sido capaz de obtener una precisión del 44,11% a la hora de localizar a nivel de ciudad cualquier conjunto de tweets de un usuario, y un notable 73,15% cuando se trata de localizar a usuarios de Twitter más activos (100 o más tweets).

Los experimentos revelan que cuando se dispone de un corpus grande de entrenamiento, tanto para entrenar como para probar, la ganancia aportada por recursos formales no es significativa, sin embargo, esta tendencia cambia cuando el tamaño del corpus es limitado. En esta situación, Wikipedia se convierte en un interesante recurso para minimizar el impacto negativo de disponer de un conjunto de entrenamiento no muy extenso. Así pues, se consiguen las siguientes mejoras en los siguientes casos:

1. Cuando el conjunto de entrenamiento es pequeño. Normalmente esto ocurre en ciudades no muy grandes. En la tabla 4 se puede ver como la precisión aumenta de un 29,17% utilizando sólo *Twitter* como entrenamiento, a un 31,12% añadiendo la *Wikipedia* al entrenamiento (alrededor de un 6% de mejora).
2. Cuando el número de tweets enviados por un usuario no es muy grande. En los experimentos mostrados en le tabla 4 se puede apreciar como la precisión pasa de un 25,16% utilizando sólo *Twitter*, a un 27,43% utilizando *Twitter* y *Wikipedia* con usuarios poco activos, es decir, que han enviado menos de 10 *tweets* (alrededor de un 8% de mejora).

5.1. Trabajo futuro

Como trabajo futuro a esta investigación se han planeado una serie de mejoras y experimentos enumerados a continuación:

1. Probar nuestro sistema con otras técnicas de ML y comprar los resultados con los actuales. También sería interesante el poder comparar nuestro sistema con otros que tratan de resolver la misma problemática.
2. Probar nuestro sistema con distinta granularidad. Actualmente, nuestro sistema trabaja únicamente con la ciudad más grande de cada una de las provincias españolas (más Ceuta y Melilla). Intuitivamente, si nos centrásemos en un provincia e hiciéramos el mismo experimento con las ciudad más grande de cada comarca en dicha provincia, la detección del foco geográfico variaría significativamente debido a la coincidencia de expresiones y términos usados por la gente en un área menor en comparación con la usada en los experimentos de este artículo.

3. Probar nuestro sistema en otros lenguajes. Con este fin, hemos recolectado un corpus de *tweets*, de similar tamaño al empleado en este artículo, de la ciudad más grande cada uno de los estados de EEUU.
4. Añadir más recursos externos de diferente naturaleza, tales como textos de las fotografías de *Flickr*, comentarios de *Trip Advisor*, artículos de periódicos, etc., y probar nuestro sistemas en distintos escenarios: localización de fotos en *Flickr*, comentarios en blogs, etc.
5. Automatizar completamente nuestro sistema recuperando automáticamente los artículos de la *Wikipedia* de cada una de las ciudades examinadas (actualmente es una labor manual).

Agradecimientos

Esta investigación ha sido parcialmente financiada por la Generalitat Valenciana bajo el proyecto GV/2012/110, y está incluida en el proyecto GRE12-44: *Tratamiento inteligente de la información para ayuda a la toma de decisiones*.

Referencias

Peregrino, Fernando S., Tomás, D., & Llopis, F. (2012). Una aproximación basada en corpus para la detección del foco geográfico en el texto. Procesamiento del Lenguaje Natural, 50(0). Retrieved from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4661

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 273-280. Retrieved from http://dl.acm.org/citation.cfm?id=1009040

Qin, T., Xiao, R., Fang, L., Xie, X., & Zhang, L. (2010). An efficient location extraction algorithm by leveraging web contextual information. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems GIS 10 (p. 53). ACM Press. Retrieved from http://portal.acm.org/citation.cfm?id=1869801

Anastácio, I., Martins, B., & Calado, P. (2009). A Comparison of Different Approaches for Assigning Geographic Scopes to Documents. INForum (pp. 285-296).

Woodruff, A. G., & Plaunt, C. (1994). GIPSY: Automated Geographic Indexing of Text Documents. Journal of the American Society for Information Science (JASIS), 45(9), 645-655. John Wiley \& Sons, Inc.

Martins, B., & Silva, M. J. (2005). A graph-ranking algorithm for geo-referencing documents. Fifth IEEE International Conference on Data Mining (ICDM'05).

Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. Proceedings of the third ACM international conference on Web search and data mining - WSDM '10 (p. 251). ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=1718487.1718519

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber ' s Heart : The Dynamics of the " Location " Field in User Profiles. Conference on Human Factors in Computing Systems (pp. 237-246).

Mahmud, J., Nichols, J., & Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM, 511-514. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4605/5045

Kinsella, S., Murdock, V., & O'Hare, N. (2011). "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11 (p. 61). ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2065023.2065039\nhttp://dx.doi.org/10.1145/2065023.2065039

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. (Array, Eds.)Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 98(3), 275-281. ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=290941.291008Kinsella, S., Murdock, V., & O'Hare, N. (2011). "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11 (p. 61). ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2065023.2065039\nhttp://dx.doi.org/10.1145/2065023.2065039