ELiRF-UPV en TweetLID: Identificación del idioma en Twitter*

ELiRF-UPV at TweetLID: Twitter Language Identification

Lluís-F. Hurtado, Ferran Pla, Mayte Giménez y Emilio Sanchis
Universitat Politècnica de València
Camí de Vera s/n
46022 València
{lhurtado, fpla, mgimenez, esanchis}@dsic.upv.es

Resumen: En este trabajo se describe la participación del equipo del grupo de investigación ELiRF de la Universitat Politècnica de València en el Taller sobre Identificación del Idioma. Este taller es un evento enmarcado dentro de la XXX edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural. Este trabajo presenta las aproximaciones utilizadas para las dos tareas del taller, los resultados obtenidos y una discusión de los mismos.
Palabras clave: Twitter, Identificación del Idioma.

Abstract: This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València in the Twitter Language Identification Workshop (tweetLID 2014). This workshop is a satellite event of the XXX edition of the Annual Conference of the Spanish Society for Natural Language Processing. This work describes the approaches used for the two tasks of the workshop, the results obtained and a discussion of these results.
Keywords: Twitter, Identificación del Idioma.

1. Introducción y Objetivos

Las redes sociales se han convertido en una excelente plataforma para analizar el comportamiento y los contenidos de sus usuarios. En la actualidad, existe un gran interés en el procesamiento de los contenidos textuales generados en la web con el fin de determinar, entre otras cosas, las opiniones y sentimientos expresadas por los usuarios sobre una gran cantidad de temas como política, economía, productos comerciales, servicios, etc. En ese sentido, Twitter se ha convertido en una plataforma muy popular para determinar en tiempo real el comportamiento y las opiniones de los usuarios en la red sobre diferentes temas de interés. Muchas de las aproximaciones de análisis de sentimientos utilizan técnicas de procesamiento del lenguaje natural. Este proceso se puede realizar más adecuadamente si se dispone de un sistema de identificación del idioma con el fin de poder utilizar los recursos lingüísticos y las técnicas más adecuadas para el idioma que se pretende procesar.

La tarea primordial de identificación del idioma consiste en decidir el idioma o idiomas, entre los idiomas candidatos, que aparecen en un determinado texto. En la literatura se han utilizado diferentes métodos para la identificación de idioma, incluyendo n-gramas, palabras de alta frecuencia, y otros enfoques estadísticos y de aprendizaje automático (Lui, Lau, y Baldwin, 2014) que obtienen buenas prestaciones para documentos de la web. Para el tratamiento de los textos de Twitter (textos cortos, agramaticales, lenguaje específico, emoticonos, abreviaturas, símbolos y expresiones universales, etc.) el problema es más complejo y en consecuencia, las técnicas de determinación del idioma utilizadas en documentos de textos normativos y de tamaños muy superiores a los empleadas en Twitter, reducen sus prestaciones en textos cortos como los de Twitter (Goldszmidt, Najork, y Paparizos, 2013) (Lui y Baldwin, 2014).

Es en este contexto en el que surge el taller tweetLID 2014 como un evento satélite del Congreso de la Sociedad Española

* Este trabajo ha sido parcialmente subvencionado por los proyectos DIANA: DIscourse ANAlysis for knowledge understanding (MEC TIN2012-38603-C02-01) y Tímpano: Technology for complex Human-Machine conversational interaction with dynamic learning (MEC TIN2011-28169-C05-01)

para el Procesamiento del Lenguaje Natural (SEPLN'14). La idea principal del taller es centrarse en los 5 idiomas principales de la Península Ibérica (español, portugués, catalán, euskera y gallego), además del inglés. Se proporciona un conjunto de entrenamiento de tweets, anotados con el idioma o idiomas que contienen y un conjunto de test para la evaluación de las diferentes aproximaciones de los participantes. Esta tarea se puede abordar sólo utilizando el corpus de entrenamiento (restringida) o utilizando todos los recursos que se consideren adecuados (no restringida).

El presente artículo resume la participación del equipo ELiRF-UPV de la Universitat Politècnica de València en este taller. A continuación se presentan las características de los sistemas desarrollados, los recursos utilizados, así como los resultados obtenidos.

2. *Sistemas desarrollados y recursos utilizados*

La identificación del idioma de una colección de documentos es un ejemplo paradigmático de una problema de clasificación multiclase. Si además, como es el caso de tweetLID, cada documento puede contener trozos de texto escritos en idiomas distintos, nos encontramos ante un problema de clasificación multietiqueta: a cada muestra a clasificar se le puede asociar una o varias etiquetas del conjunto de etiquetas disponibles.

2.1. Aproximaciones restringidas

Para las aproximaciones restringidas utilizamos modelos de clasificación basados en Máquinas de Vectores de Soporte (SVM) por su capacidad para manejar con éxito grandes cantidades de características. En concreto usamos dos librerías (*LibSVM*[1] y *LibLinear*[2]) que han demostrado ser eficientes implementaciones de SVM que igualan el estado del arte.

El software se ha desarrollado en *Python* y para acceder a las librerías de SVM se ha utilizado el toolkit *scikit-learn*[3].

Para tratar el problema de la clasificación multiclase se decidió utilizar una estrategia de uno-contra-todos (one-vs-all) donde se aprende un clasificador binario por cada clase capaz de discriminar entre *esa clase* y *no esa clase* (todas las demás). Se desarrollaron dos sistemas cuya mayor diferencia consiste en la técnica utilizada para el tratamiento de la clasificación multietiqueta.

2.1.1. Constrained-run1

Para el primer sistema restringido seleccionamos una técnica sencilla, pero habitual, en la clasificación multietiqueta. Consideramos como una nueva etiqueta cada conjunto de multietiquetas que aparece en el entrenamiento. Esta aproximación presenta dos limitaciones evidentes: a) las combinaciones de etiquetas no vistas en el entrenamiento no pueden ser predichas; y b) las combinaciones poco habituales tendrán pocas muestras de entrenamiento aunque las etiquetas individuales tengan muchas muestras. A pesar de sus limitaciones esta aproximación suele obtener buenos resultados experimentales. La etiqueta seleccionada es la que su clasificador one-vs-all obtiene mayor confianza.

Se utilizó la aproximación de bolsa de caracteres para representar cada tweet como un vector de características que contuviese los coeficientes tf-idf de los n-gramas de caracteres presentes en el tweet. Tanto el valor de n como los parámetros de los clasificadores se determinaron en la fase de ajuste de parámetros mediante validación cruzada.

2.1.2. Constrained-run2

Para el segundo sistema se eligió una estrategia distinta. No se crearon nuevas etiquetas, como resultado de las combinaciones vistas en el entrenamiento. En lugar de esto, se permitió seleccionar más de una etiqueta *básica*. Para lo cual, se sustituyó el criterio de elegir la etiqueta de mayor confianza por el de elegir todas las etiquetas cuyo modelo (one-vs-all) superase un umbral determinado experimentalmente.

En el caso de que, para un tweet, ninguna etiqueta superase el umbral, se seleccionaba como etiqueta la misma que seleccionó el sistema anterior, constrained-run1.

2.2. Aproximaciones no restringidas

A la hora de abordar la tarea no restringida el equipo ELiRF-UPV planteó el uso de aproximaciones clásicas para la detección del idioma en textos normativos y la utilización de recursos no directamente relacionados con Twitter. De esta forma se intentaba establecer una comparativa entre modelos aprendi-

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[3] http://scikit-learn.org/stable/

dos sólo con tweets (los sistemas de la tarea restringida) y modelos aprendidos exclusivamente utilizando textos normativos (nuestros sistemas para la tarea no restringida).

En concreto, se desarrolló un sistema (unconstrained-run1) que utiliza clasificadores multiclase basados en SVMs y otro sistema (unconstrained-run2) basado en el uso del módulo de identificación de idioma incorporado en la herramienta lingüística *Freeling*[4].

Una característica de los textos en redes sociales, que dificulta su tratamiento por modelos aprendidos con textos normativos, es su agramaticalidad. Para paliar ligeramente ese problema, se realiza un preprocesado de los tweets antes de clasificarlos. El preproceso consiste en la eliminación de los tokens no formados exclusivamente por letras y la reducción de los caracteres repetidos más de 2 veces en un token.

2.2.1. Unconstrained-run1

Para este primer sistema se optó por aprender modelos de clasificación basados en SVM utilizando como corpus de aprendizaje textos obtenidos de las distintas versiones idiomáticas de *Wikipedia*.

La *Wikipedia* dispone de versión en todos los idiomas del tweetLID. Además, existen versiones descargables como un único fichero en formato XML que hace más cómodo el procesado de los textos. El tamaño de estos ficheros depende del idioma; en nuestro caso oscilaba entre los 11GB de texto en inglés a los 565MB de la versión en gallego.

Se creó un lexicón por cada idioma que contenía las palabras más frecuentes en ese idioma. Se definieron umbrales distintos para cada idioma con el objetivo de intentar equilibrar la talla de los lexicones (más restrictivos para idiomas con más textos). Como contrapartida, esto supone una mayor *limpieza* en los lexicones de idiomas con mayor cantidad de texto.

A partir de los lexicones, se creó un corpus de entrenamiento considerando cada palabra como una muestra para su idioma. Utilizando ese corpus se aprendió un clasificador multiclase, pero monoetiqueta, siguiendo la misma estrategia descrita para el sistema constrainded-run1.

2.2.2. Unconstrained-run2

Para el segundo sistema no restringido se optó por utilizar el identificador de idioma presente en la herramienta *Freeling*. Este módulo de identificación de idioma utiliza un modelo de 4-gramas de caracteres por cada idioma que se debe considerar. La probabilidad asignada por cada modelo se divide por la longitud del texto a clasificar para obtener una probabilidad normalizada por carácter.

Freeling dispone de modelos para todos los idiomas que se deben considerar en el taller, excepto para el euskera. Para suplir esta carencia, a partir del léxico utilizado en el sistema anterior (unconstrained-run1) se aprendió un modelo de 4-gramas de caracteres para el euskera que se incorporó a *Freeling*.

Incluso con el nuevo modelo para el euskera, los resultados obtenidos utilizando directamente *Freeling* no fueron satisfactorios. Por este motivo se decidió implementar un sistema algo más sofisticado.

El sistema desarrollado etiqueta cada tweet con todos los idiomas para los que se obtiene una probabilidad mayor a un umbral. Esa probabilidad es el resultado de combinar linealmente la probabilidad de un modelo de palabras y la de un modelo de segmentos. La probabilidad de que en el tweet t aparezca el idioma L_i, $P(L_i|t)$, se calcula como:

$$P(L_i|t) = \lambda P_w(L_i|t) + (1-\lambda)P_s(L_i|t)$$

Donde, $P_w(L_i|t)$ es la probabilidad asignada por un modelo de palabras y $P_s(L_i|t)$ es la probabilidad del modelo de segmentos.

La probabilidad del modelo de palabras se calcula como el número de palabras de t que aparecen en el lexicón de L_i normalizado por la suma total para todos los idiomas.

Para calcular la probabilidad del modelo de segmentos se genera el conjunto, S_t, de todos los segmentos del tweet $t = t_1..t_{|t|}$ que dividan el tweet en dos partes, esto es:

$$S_t = \bigcup_{i=1}^{|t|} \{t_1^i, t_{i+1}^{|t|}\}$$

$P_s(L_i|t)$ se calcula como el número de segmentos de S_t para los que *Freeling* decide que el idioma es L_i, $C(S_t, L_i)$, normalizado por el número total de segmentos en S_t.

Si a un tweet no se le puede asignar ningún idioma, mediante el procedimiento descrito, el idioma asignado puede ser *indefinido* o el idioma más frecuente en el corpus de entrenamiento.

[4] http://nlp.lsi.upc.edu/freeling/

Sistema	P	R	F1	Posición
constrained-run2	0.825	0.744	0.752	1ª
constrained-run1	0.824	0.730	0.745	2ª
unconstrained-run2	0.737	0.723	0.697	2ª
unconstrained-run1	0.742	0.686	0.684	3ª

Tabla 1: Resultados obtenidos por los sistemas del grupo ELiRF-UPV para las dos tareas de la competición tweetLID-2014

3. Ajuste y Evaluación

La tarea consiste en asignar uno o varios idiomas a cada tweet. El conjunto de entrenamiento estaba formado teóricamente por 14991 tweets que cada grupo debía descargar él mismo (debido a restricciones de privacidad impuestas por Twitter). En el momento de la descarga solo encontramos disponibles 13919, que fue lo que constituyó nuestro corpus de entrenamiento y ajuste.

Para la tarea restringida -puesto que el conjunto de entrenamiento y ajuste era el mismo- se determinó el mejor conjunto de parámetros mediante un proceso de validación cruzado de 5 iteraciones (5-fold cross-validation). Durante este proceso de ajuste de parámetros también se probaron diferentes kernels para las SVM, los mejores resultados se obtuvieron utilizando un kernel lineal.

Para la tarea no restringida, se utilizó un corpus de 13919 tweets como conjunto de prueba y ajuste y los lexicones de Wikipedia como corpus de entrenamiento.

La Tabla 1 muestra los resultados obtenidos por los diferentes sistemas desarrollados por el equipo ELiRF-UPV en las dos tareas de la competición. Se muestran los valores de precisión (**P**), exhaustividad (**R**) y **F1** calculados como la media de los valores obtenidos para estas medidas en todos los idiomas considerados en la competición (referenciado habitualmente como macroaverage). También se incluye la posición alcanzada por cada sistema en la competición.

4. Conclusiones y trabajos futuros

En este trabajo se ha presentado la participación del equipo ELiRF-UPV en las 2 tareas planteadas en tweetLID-2014. Nuestro equipo ha utilizado técnicas de aprendizaje automático, en concreto, aproximaciones basadas en Máquinas de Vectores de Soporte y modelos de 4-gramas. Para ello, hemos utilizado la herramienta *Freeling*, la librería para Python *scikit-learn* y las librerías externas *LibSVM* y *LibLinear*. Los sistemas desarrollados por el equipo ELiRF-UPV han alcanzado los dos primeros puestos en la tarea restringida, y el segundo y tercer lugar en la tarea no restringida.

Nuestro grupo esta interesado en seguir trabajando en la minería de textos en redes sociales y especialmente en la identificación del idioma como paso previo a la correcta aplicación de técnicas de procesamiento del lenguaje natural a los documentos adaptadas a su idioma.

Como trabajo futuro cabe reseñar nuestra intención de unir las dos aproximaciones desarrolladas en este taller - la aproximación basada en corpus específico de tweets y la aproximación basada en léxico genérico extraído de *Wikipedia*. Esta unión podría realizarse mediante la combinación (por ejemplo, mediante interpolación) de los diferentes sistemas; o mediante la creación de un corpus de aprendizaje que adapte el corpus genérico a las características especificas de Twitter.

Bibliografía

Goldszmidt, Moises, Marc Najork, y Stelios Paparizos. 2013. Boot-strapping language identifiers for short colloquial postings. En *Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*. Springer Verlag.

Lui, Marco y Timothy Baldwin. 2014. Accurate language identification of twitter messages. En *Proceedings of the EACL 2014 Workshop on Language Analysis in Social Media (LASM 2014)*, páginas 17–25.

Lui, Marco, Jey Han Lau, y Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.