N-gramas basados en dependencias sintácticas como características de clasificación

Grigori Sidorov[1] , Francisco Velasquez[1] , Efstathios Stamatatos[2] , Alexander Gelbukh[1] , y Liliana Chanona-Hernández[3]

[1] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Ciudad de México, México
[2] Universidad del Egeo, Grecia
[3] ESIME, Instituto Politécnico Nacional (IPN), Ciudad de México, México
www.cic.ipn.mx/~sidorov

Resumen. En este artículo introducimos el concepto de n-gramas sintácticos (sn-grams). Los sn-gramas difieren de los n-gramas tradicionales en la forma en que los elementos se consideran vecinos. En el caso de los sn-gramas, los vecinos se toman siguiendo relaciones sintácticas en árboles sintácticos, y no tomando las palabras tal y como aparecen en el texto. Los árboles de dependencia se ajustan directamente a esta idea, mientras que en el caso de los árboles de constituyentes deben realizarse algunos pasos adicionales sencillos. Los sn-gramas pueden aplicarse a cualquier tarea de PLN en la que se utilicen n-gramas tradicionales. Describimos cómo se aplicaron los sn-gramas a la atribución de autoría. Se utilizó un clasificador SVM para varios tamaños de perfil. Se utilizaron como referencia n-gramas tradicionales de palabras, etiquetas POS y caracteres. Los resultados obtenidos son mejores cuando se aplican los sn-gramas.

Palabras clave: n-gramas sintácticos, sn-gramas, análisis sintáctico, características de clasificación, rutas sintácticas, atribución de autoría.

1 Introducción

Las técnicas basadas en n-gramas predominan en la PNL moderna y sus aplicaciones. Los n-gramas tradicionales son secuencias de elementos tal y como aparecen en los textos. Estos elementos pueden ser palabras, caracteres, etiquetas POS o cualquier otro elemento que se encuentre uno tras otro. La convención común es que "n" corresponde al número de elementos de la secuencia.

La idea principal de este trabajo es que los n-gramas pueden obtenerse a partir del orden en que se presentan los elementos en los árboles sintácticos. Es decir, seguimos un camino en el árbol y construimos n-gramas, en lugar de tomarlos tal y como aparecen en la representación superficial del texto. Así, consideramos vecinas las

palabras (u otros elementos como etiquetas POS, etc.) que se suceden en el recorrido del árbol sintáctico, y no en el texto. A estos n-gramas los llamamos **n-gramas sintácticos (sn-grams)**. La gran ventaja de los sn-gramas es que se basan en las relaciones sintácticas de las palabras y, por tanto,

I. Batyrshin y M. González Mendoza (Eds.): MICAI 2012, Parte II, LNAI 7630, pp. 1-11, 2013.

cada palabra está ligada a sus vecinas "reales", ignorando la arbitrariedad que introduce la estructura superficial.

Por ejemplo, consideremos dos frases: "*comer con cuchara de madera*" frente a "*comer con cuchara metálica*", véase la Fig. 1. Observe que podemos utilizar tanto las representaciones de dependencia como las de constituyente de las relaciones sintácticas. Son equivalentes si se conoce la cabeza de cada constituyente. En nuestro ejemplo de constituyentes, las cabezas están marcadas con líneas más gruesas. Es muy fácil añadir información sobre cabezas en la gramática basada en constituyentes, es decir, uno de los componentes debe marcarse como cabeza en las reglas.

En caso de dependencias, seguimos el camino marcado por las flechas y obtenemos sn- gramos. En el caso de los constituyentes, primero "promocionamos" los nodos de cabeza para que ocupen los lugares de las bifurcaciones, como se muestra en la Fig. 2. Luego obtenemos los sn-gramas partiendo de los constituyentes dependientes y tomando las cabezas de las bifurcaciones. A continuación, obtenemos sn-gramas partiendo de los constituyentes dependientes y tomando las cabezas de las bifurcaciones.

Consideremos el caso de los bigramas para este ejemplo. Si extraemos los bigramas tradicionales de las frases, sólo tienen un bigrama en común: "*comer con*". En cambio, si consideramos los sn-gramas, encontramos dos bigramas comunes: "*comer con*", "*con cuchara*".

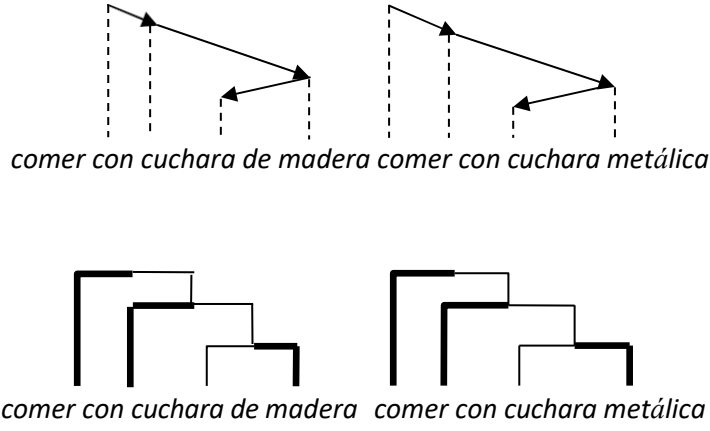

Fig. 1. Representaciones de las relaciones sintácticas

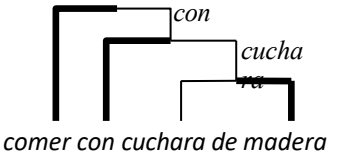

Fig. 2. Nodos cabeza promovidos

Lo mismo ocurre con los trigramas. En el caso de los n-gramas tradicionales, no hay n-gramas comunes, pero si utilizamos sn-gramas, entonces hay un trigrama común: "*comer con cuchara*".

En este caso, los sn-gramas permiten ignorar el fenómeno superficial de la lengua inglesa que añade un adjetivo antes del sustantivo y, de este modo, estropea los bigramas/trigramas tradicionales. Lo mismo ocurre, por ejemplo, con las oraciones subordinadas y, en general, con cualquier tipo de estructura sintáctica.

Otra posibilidad bastante obvia es construir sn-gramas ignorando las palabras auxiliares (stop words). Seguiremos los caminos del árbol y sólo pasaremos por los nodos de las palabras de parada. En el caso de nuestros ejemplos, tienen el sn-bigrama común "*comer cuchara*" que no se obtendrá utilizando los n-gramas tradicionales.

Los sn-gramas pueden utilizarse como características de clasificación del mismo modo que los n-gramas tradicionales. Así, en cualquier tarea en la que utilicemos n-gramas tradicionales, también podemos aplicar sn-gramas.

El problema obvio de los sn-gramas es que requieren un análisis sintáctico. El análisis sintáctico puede llevar un tiempo considerable y existe el problema de la disponibilidad de analizadores sintácticos para determinadas lenguas, aunque para lenguas bien estudiadas como el inglés o el español esta consideración no es un problema.

En este artículo, aplicamos los sn-gramas al problema de la atribución de autoría. Realizamos experimentos en que obtienen mejores resultados para esta tarea con los sn-gramas que con los n-gramas tradicionales. Además, creemos que los sn-gramas tienen una interpretación lingüística real en lo que se refiere al estilo de escritura de los autores porque reflejan relaciones sintácticas reales.

En este artículo, primero analizamos los n-gramas sintácticos (Sección 2), presentamos trabajos relevantes para la atribución de autoría (Sección 3) y, a continuación, presentamos resultados experimentales que comparan el enfoque propuesto con los métodos tradicionales (Sección 4). Por último, resumimos las conclusiones extraídas de este estudio.

2 N-gramas sintácticos (sn-grams)

La ventaja de los n-gramas sintácticos (sn-grams), es decir, los n-gramas que se construyen utilizando caminos en árboles sintácticos, es que son menos arbitrarios que los n-gramas tradicionales. Por tanto, su número es inferior al de los n-gramas tradicionales. Además, p u e d e n interpretarse como un fenómeno lingüístico, mientras que los n-gramas tradicionales no tienen una explicación plausible.

interpretación lingüística—son meros artefactos estadísticos.

La justificación de la idea de los sn-gramas está relacionada con la introducción de la lingüística

información en métodos basados en estadísticas. Creemos que esta idea ayuda a superar la principal desventaja de los n-gramas tradicionales—que contienen muchos elementos arbitrarios, es decir, mucho ruido.

La desventaja obvia de los n-gramas sintácticos es que es necesario un procesamiento sintáctico previo. Esto consume un tiempo considerable y no es fácil de aplicar a algunas lenguas, porque se necesita un analizador sintáctico y un conjunto de recursos léxicos que son utilizados por el analizador sintáctico y no para cualquier lengua estos recursos están disponibles.

Anteriormente, ideas similares se relacionaban con algunas tareas específicas como

el uso de información sintáctica adicional en la traducción automática [1] o la generación en la traducción automática

[2], sin la generalización y taxonomía que proponemos en este trabajo. El término n-grama sintáctico no es muy común y se subestima su importancia. Se utiliza, por ejemplo, en [3] para la extracción de la polaridad de los constituyentes sintácticos (chunks) como elemento completo.

Obsérvese que hay intentos de superar las desventajas de los n-gramas tradicionales utilizando enfoques puramente estadísticos. Entre ellos, cabe citar las secuencias de salto y las secuencias máximas frecuentes (SMF).

Los n-gramas son muy similares a los n-gramas, pero durante su construcción se ignoran (omiten) algunos elementos de la secuencia correspondiente. También se intenta evitar el posible ruido, es decir, considerar las variaciones aleatorias de los textos. Puede haber omisiones con varios pasos de omisión.

Un ejemplo de bigramas salteados: digamos que para la secuencia ABCDE obtenemos los bigramas tradicionales AB, BC, CD y DE. Los bigramas de salto con un paso de salto de 1 serán AC, DB y CE. Se pueden utilizar varios pasos de salto. Normalmente, los bigramas de salto también incluyen n-gramas tradicionales, en cuyo caso el paso de salto es cero. El problema de los programas de salto es que su número crece muy rápido.

Las Secuencias Máximamente Frecuentes (SMF) [4] son saltosgramas con una frecuencia mayor, es decir, sólo se tienen en cuenta los saltosgramas cuya frecuencia es superior a un determinado umbral. El problema de las MFS es que para su construcción deben utilizarse algoritmos complejos y requiere un tiempo de procesamiento considerable. Otra desventaja de los MSF es que, a diferencia de los sn-gramas, dependen de la recopilación de texto. Y, al igual que ocurre con los sn-gramas, en general no es posible una interpretación lingüística de los MFS.

Según los **tipos de elementos** que forman los n-gramas sintácticos, puede haber varios tipos de sn-gramas:

- Sn-gramas de palabras: los elementos del sn-grama son palabras,
- sn-gramas POS: los elementos del sn-grama son etiquetas POS,
- Sn-gramas de etiquetas de relaciones sintácticas: Etiquetas SR, los elementos de sn-gram son nombres de relaciones sintácticas,
- sn-gramas mixtos: los sn-gramas están compuestos por elementos mixtos como palabras (unidades léxicas), etiquetas POS y/o etiquetas SR. Dado que un sn-grama sigue un enlace sintáctico, hay razones para utilizar sn-gramas mixtos, por ejemplo, pueden reflejar marcos de subcategorización. Hay muchas combinaciones posibles
  con respecto a qué parte—de la palabra principal o de la palabra dependiente en la relación—debe ser representada por unidad léxica, etiqueta POS o etiqueta SR. Estos
  combinaciones deberían explorarse experimentalmente en el futuro.

Tenga en cuenta que los sn-gramas de caracteres son imposibles.

En cuanto al **tratamiento de las stop words**, pueden ignorarse o tenerse en cuenta, como ya hemos dicho.

3 Sn-gramas de etiquetas SR

Para los experimentos descritos en este artículo, utilizamos relaciones sintácticas (etiquetas SR) como elementos de los sn-gramas. Para determinar las etiquetas SR se utilizó el analizador sintáctico Stanford, las etiquetas POS,

y para la construcción de árboles sintácticos basados en dependencias. Aunque el proceso de análisis sintáctico requirió mucho tiempo para un corpus grande, sólo se realizó una vez, por lo que los experimentos posteriores no llevan un tiempo considerable.

Consideremos una frase de ejemplo "*Las noticias económicas tienen poco efecto en los mercados financieros*" que se utiliza en algunos cursos de PNL. El resultado del analizador sintáctico de Stanford para las relaciones de dependencia es:

nn (news-2, Económico-1)
nsubj (have-3, news-2)
root (ROOT-0, have-3)
amod (effect-5, little-4)
dobj (have-3, effect-5)
prep (effect-5, on-6)
amod (mercados-8, financiero-7)
pobj (on-6, mercados-8)

Esta representación contiene la siguiente información: nombre de la relación (que aparece al principio de cada línea) y palabras relacionadas entre paréntesis, donde el primer argumento dentro del paréntesis representa la cabeza y el segundo representa el elemento dependiente. Así, *amod (efecto-5, pequeño-4) significa* que existe una relación de modificador adjetival (*amod)* de *efecto* a *pequeño*. Los números corresponden a los números de palabra de la frase. El analizador sintáctico de Stanford maneja 53 relaciones.

Las etiquetas SR se presentan en cuadrados sobre el árbol sintáctico de la Fig. 3.

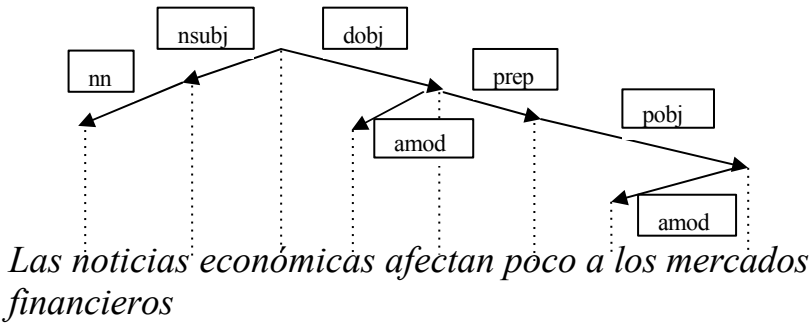

Fig. 3. Árbol de dependencias del ejemplo

A partir de las relaciones establecidas por el analizador sintáctico, obtenemos sn-gramas siguiendo las flechas. Por ejemplo, de este árbol podemos extraer los siguientes sn-gramas:

nsubj → nn
dobj →
amod
dobj → prep → pobj → amod

Nótese que utilizamos todos los caminos posibles, por ejemplo, si consideramos

trigramas, entonces se extraerán los caminos *dobj*→*prep*→*pobj* y *prep* → *pobj* → *amod*, etc.

4 Trabajos relevantes sobre la atribución de autoría

La atribución de autoría aborda una cuestión antigua y difícil: cómo asignar un texto de autoría desconocida o discutida a un miembro de un conjunto de autores candidatos de los que se dispone de muestras indiscutibles de textos [8]. A pesar de su aplicación a las obras literarias, la rápida expansión de los textos en línea en los medios de Internet (por ejemplo, blogs, mensajes de correo electrónico, publicaciones en redes sociales, etc.) puso de manifiesto aplicaciones prácticas de la atribución de autoría normalmente asociadas a tareas forenses [11].

Los enfoques automatizados de este problema implican el uso de métodos estadísticos o de aprendizaje automático [6]. Desde el punto de vista del aprendizaje automático, la atribución de autoría puede considerarse una tarea de clasificación multiclase de una sola etiqueta [10]. Hay dos pasos básicos: en primer lugar, los textos deben representarse adecuadamente como vectores de valores numéricos y, a continuación, un algoritmo de clasificación puede utilizar estos vectores para estimar la probabilidad de pertenencia a una clase para cualquier texto dado.

Hasta ahora se han propuesto miles de características estilométricas. Estos rasgos pueden distinguirse en las siguientes categorías principales según el análisis textual que requieran [6]: rasgos léxicos (por ejemplo, frecuencias de palabras funcionales, frecuencias de n-gramas de palabras, medidas de riqueza de vocabulario, etc.), rasgos de carácter (por ejemplo, frecuencias de letras y n-gramas de caracteres), rasgos sintácticos (por ejemplo, frecuencias de etiquetas de parte de voz, medidas de estructura de frases y oraciones, frecuencias de reglas de reescritura, etc.), características semánticas (por ejemplo, medidas de sinónimos, medidas de dependencia semántica, etc.) y características específicas de la aplicación (por ejemplo, tamaño de letra, color de letra, frecuencias de palabras específicas, etc.). Hasta ahora, varios estudios [13, 14, 18] han demostrado que las medidas más eficaces son las características léxicas y de caracteres.

Nótese que la información sobre relaciones sintácticas no suele utilizarse en esta tarea, siendo una de las excepciones [1], donde los autores analizan reglas de reescritura sintáctica. En este trabajo mostramos que las relaciones sintácticas tomadas como sn-gramas representan una medida muy eficaz.

En este trabajo, utilizamos como métodos de referencia características de caracteres, características léxicas (palabras) y etiquetas POS obtenidas mediante n-gramas tradicionales, es decir, según la aparición de los elementos en los textos.

5 Resultados experimentales y debate

Los experimentos se realizaron sobre los datos del corpus. El corpus utilizado en nuestro estudio incluye textos descargados del Proyecto Gutenberg. Seleccionamos libros de autores nativos de habla inglesa que tienen una producción literaria en un periodo similar. En este trabajo, todos los experimentos se realizan para el corpus de 39 documentos de tres autores.

Para la evaluación de los experimentos, utilizamos el 60% de los datos para el entrenamiento, y el 40% restante para la clasificación, como se presenta en la Tabla 1.

Utilizamos la implementación de WEKA del clasificador Support Vector Machines (SVM).

Se sabe que SVM produce muy buenos resultados en la tarea de atribución de autoría.

Utilizamos varias características como referencia: características basadas en palabras, características basadas en caracteres y etiquetas POS. Para las características

de referencia, utilizamos la técnica tradicional de n-gramas, es decir, los elementos se toman tal y como aparecen en los textos. Aplicamos la técnica sn-grams al mismo corpus y los resultados superaron a los métodos de referencia.

Tabla 1. Datos de entrenamiento y clasificación

Autor	Formación		Clasificación	
	ño (MB)	NovelasTama	ño (MB)	NovelasTama
Tarkington	8	3.6	5	1.8
George Vaizey	8	3.8	5	2.1
Louis Tracy	8	3.6	5	2.2
Total	**24**	**11**	**15**	**6.1**

Utilizamos el término "tamaño del perfil" para representar los primeros n-gramas/sn-gramas más frecuentes, por ejemplo, para un tamaño de perfil de 400 sólo se utilizan los primeros 400 n-gramas más frecuentes. Probamos varios umbrales para el tamaño del perfil y seleccionamos cinco umbrales para el tamaño del perfil, como se presenta en todas las tablas con los resultados.

Cuando la celda de la tabla contiene *NA (no disponible),* significa que nuestros datos eran insuficientes para obtener el número correspondiente de n-gramas. Esto ocurre sólo con los bigramas, porque en general hay menos bigramas que trigramas, etc. En estos casos, el número total de todos los bigramas es inferior al tamaño del perfil dado.

En las Tablas 2, 3 y 4 se presentan los resultados de los métodos de referencia seleccionados. La Tabla 5 contiene los resultados obtenidos utilizando sn-grams. Para apreciar mejor la comparación de los resultados, presentamos las Tablas 6 a 9, donde los resultados se agrupan por el tamaño de los n-gramas/sn-gramas.

Tabla 2. N-gramas basados en palabras (línea de base)

Tamaño del perfil	Tamaño del n-grama			
	2	**3**	**4**	**5**
400	86%	81%	67%	45%
1,000	86%	71%	71%	48%
4,000	86%	95%	67%	48%
7,000	86%	90%	71%	45%
11,000	89%	90%	75%	33%

Tabla 3. N-gramas basados en caracteres (línea de base)

Tamaño del perfil	Tamaño del n-grama			
	2	**3**	**4**	**5**
400	90%	76%	81%	81%
1,000	95%	86%	86%	76%
4,000	90%	95%	90%	86%
7,000	NA	90%	86%	86%
11,000	NA	100%	90%	86%

Tabla 4. n-gramas basados en etiquetas POS (línea de base)

Tamaño del perfil	Tamaño del n-grama			
	2	**3**	**4**	**5**
400	90%	90%	76%	62%
1,000	95%	90%	86%	67%
4,000	NA	100%	86%	86%
7,000	NA	100%	90%	86%
11,000	NA	95%	90%	86%

Tabla 5. sn-gramas basados en etiquetas SR

Tamaño del perfil	Tamaño del n-grama			
	2	**3**	**4**	**5**
400	100%	100%	87%	93%
1,000	100%	100%	87%	93%
4,000	100%	100%	93%	73%
7,000	100%	100%	87%	87%
11,000	100%	100%	93%	87%

En las tablas siguientes presentamos los datos agrupados por tamaños de n-gramas.

Tabla 6. Comparación de bigramas

Tamaño del perfil	Características			
	sn-gramas de etiquetas SR	**n-gramas de etiquetas POS**	**Basado en el carácter n-gramas**	**Basado en palabras n-gramas**
400	100%	90%	90%	86%
1,000	100%	95%	95%	86%
4,000	100%	NA	90%	86%
7,000	100%	NA	NA	86%
11,000	100%	NA	NA	89%

Tabla 7. Comparación de trigramas

Tamaño del perfil	Características			
	sn-gramas de etiquetas SR	**n-gramas de etiquetas POS**	**Basado en el carácter n-gramas**	**Basado en palabras n-gramas**
400	100%	90%	76%	81%
1,000	100%	90%	86%	71%
4,000	100%	100%	95%	95%
7,000	100%	100%	90%	90%
11,000	100%	95%	100%	90%

Tabla 8. Comparación para 4-gramas

Tamaño del perfil	Características			
	sn-gramas de etiquetas SR	**n-gramas de etiquetas POS**	**Basado en el carácter n-gramas**	**Basado en palabras n-gramas**
400	87%	76%	81%	67%
1,000	87%	86%	86%	71%
4,000	93%	86%	90%	67%
7,000	87%	90%	86%	71%
11,000	93%	90%	90%	75%

Tabla 9. Comparación para 5-gramas

Tamaño del perfil	Características			
	sn-gramas de etiquetas SR	**n-gramas de etiquetas POS**	**Basado en el carácter n-gramas**	**Basado en palabras n-gramas**
400	93%	62%	81%	45%
1,000	93%	67%	76%	48%
4,000	73%	86%	86%	48%
7,000	87%	86%	86%	45%

11,000	87%	86%	86%	33%

Se puede apreciar que en todos los casos la técnica sn-gram supera a la técnica basada en n-gramas tradicionales. Consideramos que las etiquetas SR y las etiquetas POS son similares a efectos de nuestra comparación; ambas son pequeños conjuntos de etiquetas: 36 y 53 elementos, asociados a palabras. Obsérvese que la mayoría de estos elementos tienen una frecuencia muy baja.

Como puede observarse, la puntuación máxima de nuestra tarea es muy alta: 100%. Está relacionado con el hecho de que utilizamos muchos datos y nuestra clasificación sólo distingue entre tres clases. En algunos casos, los métodos de referencia también alcanzan la línea superior. Aun así, sólo ocurre para un número reducido de tamaños de perfil específicos. Los mejores resultados los obtienen los sn-grams utilizando bigrams y trigrams para cualquier tamaño de perfil. Para cualquier combinación de parámetros, los métodos básicos obtienen peores resultados que los sn-grams.

Puede surgir la pregunta de si merece la pena trabajar con un número reducido de clases. En nuestra opinión, es útil e importante. En primer lugar, la atribución de autoría es a menudo una aplicación del mundo real en caso de disputa sobre la autoría de un documento, y en este caso el número de clases se reduce a dos o tres, es decir, es nuestra situación.

6 Conclusiones

En este artículo introducimos el concepto de n-gramas sintácticos (sn-grams). La diferencia entre los n-gramas tradicionales y los sn-gramas está relacionada con la forma en que los elementos se consideran vecinos. En el caso de los sn-gramas, los vecinos se toman siguiendo relaciones sintácticas en árboles sintácticos, mientras que los n-gramas tradicionales se forman tal y como aparecen en los textos.

Se puede utilizar cualquier representación sintáctica para aplicar la técnica de los sn-gramas: árboles de dependencia o árboles de constituyentes. En el caso de los árboles de dependencia, debemos seguir los enlaces sintácticos y obtener sn-gramas. En el caso de los árboles de constituyentes, hay que realizar algunos pasos adicionales, pero estos pasos son muy sencillos.

Los sn-gramas pueden aplicarse en cualquier tarea de PNL en la que se utilicen n-gramas tradicionales.

Realizamos experimentos para la tarea de atribución de autoría utilizando SVM para varios tamaños de perfil. Se utilizó un corpus relativamente grande de obras de tres autores.

Utilizamos como referencia los n-gramas tradicionales de palabras, etiquetas POS y caracteres. Los resultados muestran que la técnica sn-gram supera a la técnica de referencia.

Como trabajo futuro, aplicaremos la idea de los sn-gramas a otras tareas de PNL. En lo que respecta a nuestros datos concretos, realizaremos una comparación exhaustiva de todas las características utilizando la técnica de los sn-gramas. También tenemos previsto experimentar con varios tipos de sn-gramas mixtos.

Agradecimientos. Trabajo realizado con el apoyo parcial del gobierno mexicano (proyectos CONACYT 50206-H y 83270, SNI) y el Instituto Politécnico Nacional, México (proyectos SIP 20111146, 20113295, 20120418, COFAA, PIFI), el gobierno del Distrito Federal (ICYT-DF proyecto PICCO10-120) y FP7-PEOPLE-2010-IRSES: Web Information Quality - Evaluation Initiative (WIQ-EI) proyecto 269180 de la Comisión Europea. También agradecemos a Sabino Miranda y Francisco

Viveros sus valiosos y motivadores comentarios.

Referencias

1. Khalilov, M., Fonollosa, J.A.R.: N-gram-based Statistical Machine Translation versus Syntax Augmented Machine Translation: comparación y combinación de sistemas. En: Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 424-432 (2009)
2. Habash, N.: The Use of a Structural N-gram Language Model in Generation-Heavy Hybrid Machine Translation. En: Belz, A., Evans, R., Piwek, P. (eds.) INLG 2004. LNCS (LNAI), vol. 3123, pp. 61-69. Springer, Heidelberg (2004)
3. Agarwal, A., Biads, F., Mckeown, K.R.: Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams. En: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL), pp. 24-32 (2009)
4. García-Hernández, R.A., Martínez Trinidad, J.F., Carrasco-Ochoa, J.A.: Finding Maximal Sequential Patterns in Text Document Collections and Single Documents. Informatica 34(1), 93-101 (2010)
5. Baayen, H., Tweedie, F., Halteren, H.: Fuera de la cueva de las sombras: Using Syntactic Annotation to Enhance Authorship Attribution. Literary and Linguistic Computing, 121-131 (1996)
6. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538-556 (2009)
7. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing 13(3), 111-117 (1998)
8. Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval 1(3), 233-334 (2006)
9. Juola, P.: Concurso ad-hoc de atribución de autoría. En: Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, pp. 175-176 (2004)
10. Sebastiani, F.: Aprendizaje automático en la categorización automatizada de textos. ACM Computing Surveys 34(1) (2002)
11. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems 20(5), 67-75 (2005)
12. van Halteren, H.: Verificación de autores mediante perfiles lingüísticos: An exploration of the parameter space. ACM Transactions on Speech and Language Processing 4(1), 1-17 (2007)
13. Grieve, J.: Atribución cuantitativa de la autoría: An evaluation of techniques. Literary and Linguistic Computing 22(3), 251-270 (2007)
14. Luyckx, K.: Problemas de escalabilidad en la atribución de autoría. Tesis doctoral, Universidad de Amberes (2010)
15. Argamon, S., Juola, P.: Panorama del concurso internacional de identificación de autores en PAN-2011. En: 5th Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (2011)
16. Diederich, J., Kindermann, J., y otros: Atribución de autoría con máquinas de vectores soporte. Applied Intelligence 19(1), 109-123 (2003)
17. Escalante, H., Solorio, T., et al.: Histogramas locales de n-gramas de caracteres para la atribución de autoría. En: 49th Annual Meeting of the Association for Computational Linguistics, pp. 288-298 (2011).
18. Keselj, V., Peng, F., y otros: N-gram-based author profiles for authorship attribution. Computational Linguistics 3, 225-264 (2003)
19. Koppel, M., Schler, J., y otros: Atribución de autoría en la naturaleza. Recursos lingüísticos y evaluación 45(1), 83-94 (2011).
20. Koppel, M., Schler, J., y otros: Measuring differentiability: unmasking pseudonymous authors. Journal of Machine Learning Research, 1261-1276 (2007)