Análisis de la negación léxica:

Para la clasificación supervisada de la orientación semántica de opiniones

Alonso Palomino-Garibay, Sofía N. Galicia-Haro y Alexander Gelbukh
alonsop@ciencias.unam.mx, sngh@fciencias.unam.mx
Facultad de Ciencias, UNAM.
Centro de Investigación en Computación, IPN.

Universidad Nacional Autónoma de México

Resumen

Reportamos investigación sobre el efecto de negación léxica para la predicción de la orientación semántica. Las últimas metodologías para derivar la orientación semántica están basadas en clasificación automática. Analizamos el uso de bigramas de negación para métodos supervisados. Utilizamos un corpus de opiniones sobre lavadoras en idioma Español, usamos diversas palabras de negación como características de entrenamiento en máquinas de soporte vectorial.

Introducción

La enorme cantidad de comentarios de libre acceso en la Web, para productos y servicios, ha permitido que esas opiniones sean un recurso valioso para tomar decisiones. Esta subtarea está relacionada al procesamiento de lenguaje natural y a la minería de textos. Típicamente, las tareas en esta área van desde distinguir segmentos de texto subjetivo hasta determinar la polaridad de las palabras, de los enunciados y de los documentos.

Objetivos

1. Probar distintas características basadas en negación léxica.
2. Generar bigramas a partir de dicha negación léxica.
3. Analizar el resultado de cada bigrama de negación
4. Usar dichos bigramas como características de entrenamiento para un algoritmo de aprendizaje supervisado.
5. Conocer y explorar el alcance y comportamiento del algoritmo de aprendizaje supervisado con los bigramas antes mencionados.
6. Conocer y explorar el alcance y comportamiento de este algoritmo de aprendizaje supervisado con los bigramas antes mencionados.

Herramientas y métodos

Corpus y conocimiento lingüístico

Para este trabajo usamos la colección de opiniones de [2]. La colección fue compilada automáticamente del sitio `ciao.es` que consta de 2800 opiniones de lavadoras. El tamaño promedio por archivo en lexemas es de 345. El número total de lexemas de la colección es de 854,280. La colección total fue anotada con información de lema y categorías gramaticales utilizando *FreeLing*.

Configuración base

De la colección total de opiniones en español, extrajimos un subconjunto significativo de opiniones diferentes: 2598 opiniones. No eliminamos las opiniones que claramente son anuncios de empresas de mantenimiento (SPAM) ya que tanto este tipo de textos como las opiniones pagadas por fabricantes aparecen en cualquier colección de opiniones de productos. Utilizamos esta colección para entrenar un estimador cuyo objetivo es determinar qué tan bueno es un producto en base a la orientación semántica de las opiniones, y el puntaje de los usuarios que corresponden a: malo (una estrella), regular (dos estrellas), bueno (tres estrellas), muy bueno (cuatro estrellas) o excelente (5 estrellas). Las características de esta colección en cuanto a número de opiniones por puntaje se presentan en la Tabla 1. Como se observa y como podría esperarse de opiniones de aparatos electrodomésticos cuyo uso es tan generalizado por su gran utilidad, las opiniones positivas son mayores en una proporción de $6:1$. De acuerdo a [4] la inclusión de bigramas como características de entrenamiento mejora significativamente la tarea de minar opiniones. Por lo tanto, consideramos los siguientes bigramas morfosintácticos para métodos supervisados.

Polaridad	# de reseñas	Estrellas
Excelente	1190	5
Muy bueno	838	4
Bueno	239	3
Regular	127	2
Malo	204	1

Tabla 1: Corpus de Opiniones

Patrón	# de Bigramas	Opiniones
adjetivo-adverbio	504	401
adverbio-adjetivo	2,024	2,024
sustantivo-adjetivo	2,598	2,598
verbo-adverbio	2,006	2,006

Tabla 2: Distribucion de bigramas en el corpus

Negación

En [1] los autores señalan que la negación está presente en todos los lenguajes humanos y se usa para revertir la polaridad de partes de enunciados. La negación en Español fue dividida por [3] en negación total y parcial. La autora también analizó el efecto de la negación parcial en sintagmas, en adyacencia a sintagmas y en palabras de negación. Entre las palabras negativas consideradas están las siguientes: pronombres:*nadie, ninguno, nada* y los adverbios: *nunca, jamás, nada*. Este criterio lo seguimos en este trabajo, la negación fue considerada a nivel de secuencias morfosintácticas. Las formas negadas fueron obtenidas con patrones de búsqueda específicamente formados por secuencias de categorías gramaticales. Definimos los siguientes patrones:

1. $\text{ninguno}_{\text{LEMMA\_DET}}$-noun
2. $\text{nada}_{\text{PRONOUN}}$-adjective
3. [$\text{jamás}_{\text{ADVERB}}$ |$\text{nunca}_{\text{ADVERB}}$ |$\text{no}_{\text{ADVERB}}$]-verb
4. $\text{no}_{\text{ADVERB}}$-pronoun-verb

Resultados

Los resultados incluyendo todos los bigramas de negación para el método supervisado se muestran en la Figura 1. La primer columna muestra los patrones que corresponden a la configuración base.

Features	Metric	Value
Noun-Adjective	F1score	0.8419
	Recall	0.8287
	Precision	0.8556
Noun-Adjective Verb-Adverb	F1score	0.9287
	Recall	0.9266
	Precision	0.9309
Noun-Adjective Verb-Adverb Adverb-Adjective	F1score	0.9258
	Recall	0.9230
	Precision	0.9286
Noun-Adjective Verb-Adverb Adverb-Adjective Adjective-Adverb	F1score	0.9339
	Recall	0.9312
	Precision	0.9366

Figura 1: Clasificación con *SVM*

La segunda columna muestra en cada fila el tipo de medida y la tercera columna el valor obtenido para cada medida y para el patrón de negación correspondiente. En la Figura 2 se muestran los resultados obtenidos en la adición de cada uno de los patrones para negación léxica. Haciendo una comparación con los resultados base de la Figura 1, se observa que el único patrón que no aumentó los resultados es el que corresponde a $\text{no}_{\text{ADVERB}}$-pronoun-verb. El rango de mejoras pasó de $0{,}41\,\%$ a $1{,}85\,\%$. Dos patrones superaron el $1\,\%$: $\text{nunca}_{\text{ADVERB}}$-verb y $\text{nada}_{\text{PRONOUN}}$-adjective. $\text{Nunca}_{\text{ADVERB}}$-verbo apareció en 77 opiniones positivas y 17 opiniones negativas. $\text{Nada}_{\text{PRONOUN}}$-adjective apareció en 401 opiniones positivas y 90 opiniones negativas.

Features		Metric	Values
Noun-Adjective	+ $\text{no}_{\text{ADVERB}}$-$\text{verb}_{\text{AUX\_PAST PARTICIPLE}}$	F1score	0.9315
		Recall	0.9289
		Precision	0.9342
+ Verb-Adverb	+ $\text{ninguno}_{\text{LEMMA\_DET}}$ -noun	F1score	0.9406
		Recall	0.9382
		Precision	0.9431
	+ $\text{jamás}_{\text{ADVERB}}$-verb	F1score	0.9380
		Recall	0.9359
		Precision	0.9401
+ Adverb-Adjective	+ $\text{nunca}_{\text{ADVERB}}$ -verb	F1score	0.9524
		Recall	0.9510
		Precision	0.9537
	+ $\text{no}_{\text{ADVERB}}$ -verb	F1score	0.9407
		Recall	0.9382
		Precision	0.9432
+ Adjective-Adverb	+ $\text{nada}_{\text{PRONOUN}}$ -adjective	F1score	0.9501
		Recall	0.9487
		Precision	0.9515
	+ $\text{no}_{\text{ADVERB}}$-pronoun-verb	F1score	0.9395
		Recall	0.9370
		Precision	0.9420

Features		Metric	Value
	+		
Noun-Adjective	$\text{ninguno}_{\text{LEMMA\_DET}}$ -noun		
Verb-Adverb	$\text{jamás}_{\text{ADVERB}}$-verb	F1score	0.9448
Adverb-Adjective	$\text{nunca}_{\text{ADVERB}}$-verb		
Adjective-Adverb	$\text{no}_{\text{ADVERB}}$-verb	Recall	0.9429
	$\text{nada}_{\text{PRONOUN}}$ -adjective		
	$\text{no}_{\text{ADVERB}}$-pronoun-verb	Precision	0.9467

Figura 2: Rendimiento con bigramas de negación añadidos

Todos estos bigramas fueron usados como características de entrenamiento por el algoritmo SVM para el caso multiclase de la implementación de *scikit-learn*.

Conclusiones

- Mostramos que usar negación léxica para generar bigramas como características del método supervisado puede ser útil para derivar la polaridad de opiniones en Español.
- Presentamos un análisis del efecto de palabras específicas de negación para derivar la orientación semántica en opiniones en idioma Español.

Referencias

[1] E. Blanco and D. Moldovan. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 581–589. Association for Computational Linguistics, 2011.

[2] S. N. Galicia-Haro and A. Gelbukh. Extraction of semantic relations from opinion reviews in spanish. In *Human-Inspired Computing and Its Applications*, pages 175–190. Springer, 2014.

[3] B. Sanz Alonso. La negación en español. In *Actuales tendencias en la enseñanza del Español como lengua extranjera II: actas del VI Congreso Internacional de ASELE*, pages 379–384. Colegio de España, 1996.

[4] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012.