Hacia la detección de ironía en textos cortos

Gabriela Jasso López
Facultad de Ingeniería UNAM/IIMAS

Ironía

> decir lo contrario a lo que se quiere decir (Quintiliano)

- lenguaje figurativo
- complicado para una computadora

Esfuerzos recientes

- **Utsumi (1999)** modelo basado en proposiciones irónicas, calcula un grado de ironía (inglés)
- **Veale et al. (2008)** sistema de generación de metáforas creativas (inglés)
- **Carvalho et al. (2009)** emoticones y puntuación para detectar ironía en comentarios de usuario (portugués)
- **González-Ibáñez et al. (2011)** reconoce sarcasmo en twitter (inglés)
- **Davidov et al. (2011)** reconoce sarcasmo en reseñas de Amazon y en tweets **(2010)** (inglés)

Esfuerzos recientes

- **Liebrecht et al. (2013)** reconoce sarcasmo en tweets (neerlandés)
- **Rosso et al. (2013)** reconoce ironía en tweets con un modelo basado en estilo, escenarios emocionales (inglés)
- **Barbieri et al. (2014)** reconoce ironía en tweets con un modelo basado en estilo, polaridades, entre otros (inglés)
- **Tungthamthiti et al. (2014)** reconoce sarcasmo en tweets con un modelo basado en factores de estilo y polaridades, entre otros (inglés)

Diseño del corpus irónico

- Twitter
- español
- extenso
- denotado por #ironía (Rosso, Barbieri)
- y #sarcasmo (Tungthamthiti, Liebrecht)

Sarcasmo e ironía

Ironía *o antífrasis*: Expresión en tono de burla de una significación contraria (o diferente) a la del enunciado que se pone de manifiesto por el contexto o la pronunciación, el gesto, etc.

Sarcasmo es la clase de ironía que se caracteriza por la intención cruel, hostil o maliciosa que expresa

Asteísmo es la expresión de una alabanza en forma de represión, o viceversa

García Barrientos, J. Las figuras retóricas. Págs. 56-57

Sarcasmo e ironía (ejemplos)

Ironía: "¡qué buena suerte tengo!", para entender que no tiene buena suerte

Sarcasmo: "¡qué listo eres!" para dar a entender que es tonto

Sarcasmo e ironía (Twitter)

- Nada más lindo que pasar un domingo por la tarde estudiando #sarcasmo
- Creo que ya lo dije hoy pero no me canso de triunfar en el amor #ironia
- @sopitas me siento como si el.país no estuviera en llamas #sarcasmo
- Que raro eso de ver a Higuaín fallar en los momentos decisivos ehh.. #Ironia

Ventajas del corpus irónico

- Basado en lo que la mayoría considera irónico/sarcástico
- Automáticamente etiquetado
- Fácil obtención
- Español
- Tamaño comparable a la mayoría de las fuentes

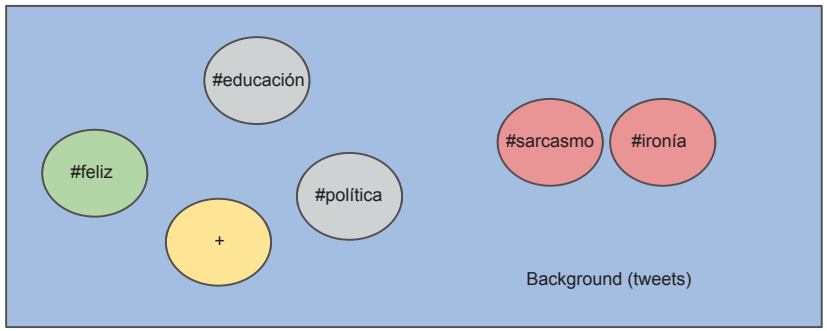
Desventajas del corpus irónico

- Errores ortográficos
- Tweets repetidos
- Nociones de ironía/sarcasmo muy fuera del promedio
- Mensajes que se transmiten con imágenes
- Al lector desconocido le falta el contexto del usuario

Clases no irónicas (comparación)

- Rosso
  - #humor
  - #política
  - #educación
- Davidov
  - reseñas positivas (imdb)
  - tweets clasificados manualmente
- González-Ibáñez
  - etiquetas positivas y negativas (#happy, #sad, ...)

Propuesta



Liebrecht et al (2013)

Recolección de textos no irónicos

- Todo lo no explícitamente etiquetado como irónico aquí se considera no-irónico

- Búsqueda de tweets con **quién/quien**, **cómo/como**, **cuándo/cuando**, **dónde/donde**, **por qué/porque**, **está/esta**, entre otras

Distribución del corpus

~14,000 tweets irónicos

~621,000 tweets no irónicos (en su mayoría)

- balanceada (igual cantidad por clase) (Liebrecht, Rosso, Barbieri, Davidov otros)
- 30% irónico / 70% no irónico (Rosso, Barbieri)
- 10% irónico / 90% no irónico (propuesta)

Preprocesado

- se elimina el # del hashtags
- minúsculas
- @usuario => @
- http://dominio.com/articulo... => http://link
- #ironía / #sarcasmo / variantes eliminados
- espacios excesivos
- se revuelve antes de realizar operaciones
- se eliminan duplicados

Experimento

- Random Forest / SVM
  - uni/bi/trigramas, tfidf
  - 10 folds (validación cruzada)

- Regresión logística
  - Word2vec(Mikolov et al, (2013))/doc2vec (Le, Mikolov (2014))
  - Para entrenar word2vec se usa la totalidad de los tweets

Resultados

balanceado	Acc	Prec	Rec	F-S
*RF	0.82	0.83	0.81	0.82
*SVM	0.86	**0.87**	**0.85**	**0.86**
Rosso	0.72	0.74	0.69	0.71
Davidov	**0.89**	0.79	0.86	0.82
Tungthamthiti	0.79	0.78	0.79	0.79
González	0.75	no reportados		
*LR(w2v)	0.81	0.82	0.81	0.81

Resultados

30%-70%	Acc	Prec	Rec	F-S
*RF	0.87	**0.91**	0.62	0.74
*SVM	**0.90**	0.89	**0.77**	**0.82**
Rosso	0.80	0.66	0.45	0.53
*LR(w2v)	0.81	0.70	0.63	0.66

Resultados

10%-90%	Acc	Prec	Rec	F-S
RF	0.93	**0.98**	0.38	0.54
SVM	**0.95**	0.90	**0.61**	**0.73**
LR(w2v)	0.84	0.29	0.35	0.32

Word2vec

Algunas analogías interesantes (basadas en distancias entre vectores de word2vec):

Nada más lindo que pasar un domingo por la tarde estudiando #sarcasmo

- No hay nada mas lindo que hacer un examen el domingo #sarcasmo
- Mañana a la tarde si esta lindo quiero hacer algo!
- Domingo. Lindo día para empezar a estudiar. #Ironía
- Lindo dia para estar despierta a las 5 de la mañana #Ironía #camavolve
- Que mejor que empezar un sábado por la mañana a estudiar #Sarcasmo
- Este fin de semana a dormir todo lo que aguante el cuerpo!!! http://t.co/RP6nsaplnl
- Hoy esta para dormirlo todo, pero no todo el día estudiando
- Que raro un sábado a estas horas yo sola en casa #ironía
- esta tarde a estudiar , q bien me lo voy a pasar #ironia

Word2vec

Que raro eso de ver a Higuaín fallar en los momentos decisivos ehh.. #Ironia

- @waltersafarian En estos momentos me parece que Agüero está con un rendimiento un poquito por encima de los otros 2.
- Argentina tiene jugadores pa mucho más pienso yo
- No tiene todos los jugadores La pobre 😓
- @KAKA @10Ronaldinho una cosa era Brasil cuando tenía estos CRACKS.... Y bueno, ahora... Es nada!
- El problema es que Ecuador juega sin delanteros.... así como berraco que hagan un gol....
- piojofail
- De momentos buenos y momentos malos está hecho el fútbol: Sergio Bueno

Conclusiones

- Resultados favorables a pesar de no usar diversas características de los antecedentes
- Resultados prometedores para word2vec
- A la par con la mayoría de las fuentes
- Consideración de un corpus más realista
- Precedente para corpus en español
- Posibilidad de mejorar métricas usando más características