

Processament del Llenguatge Humà 5. Semàntica Lèxica

Semàntica

WordNet

SentiWordNet

Análisis del
sentimiento



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat d'Informàtica de Barcelona



Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Semántica

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

La semántica trata el significado en diferentes niveles:

- Semántica léxica: palabras individuales
- Semántica composicional: secuencias de palabras con significado (sintagmas, oraciones). Alta concordancia con sintaxis

Nos enfocaremos en la semántica léxica

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

Motivación de la
semántica léxica

WordNet

SentiWordNet

Análisis del
sentimiento

Motivación de la semántica léxica

Ejemplos de utilidad:

- Descubrimiento de patrones semánticos

P.e.: USA **bombed** Hiroshima

They began to **bombard** the defenses

→ A **sense_12533** B

- Reconocimiento de relaciones discursivas

Ex: [Anna will show up **later.**] [She has **missed the train.**] →
explicación

Ex: [Mathew is good cooking.] [Albert fails making every dish] →
contraste

- Análisis de la polaridad de opinión

Ex: @vooda1: CNN Declines to Air White House Press Conference
Live YES! THANK YOU @CNN FOR NOT LEGITIMI...
positivo

Ex: @Slate: Donald Trump's administration: "Government by the
worst men."
negativo

Semántica

Motivación de la
semántica léxica

WordNet

SentiWordNet

Análisis del
sentimiento

Índice

Semántica

Recursos

WordNet

SentiWordNet

Análisis del
sentimiento

1 Semántica

- Motivación de la semántica léxica
- Recursos

2 WordNet

- Definición
- Similaridad

3 SentiWordNet

4 Análisis del sentimiento

- Definición
- Ejemplos de métodos

Recursos léxico-semánticos

- Recursos basados en conocimiento: representados como grafos
 - Ex: **WordNet** (ontología léxica en inglés)
 - SentiWordNet** (polaridad del sentimiento en WordNet)
 - BabelNet** (Wikipedia+WordNet)
 - VerbNet** (comportamiento sintáctico-semántico verbal)
 - FrameNet** (comportamiento conceptual –representación fina de eventos–)
 - ConceptNet** (conocimiento sobre el sentido común)
- Recursos basados en corpus: uso contextual de palabras
 - Ex: Latent Semantic Analysis (LSA)
 - Word embeddings**
 - Los introduciremos en las últimas sesiones**

Semántica

Recursos

WordNet

SentiWordNet

Análisis del
sentimiento

Recursos léxico-semánticos

Semántica

Recursos

WordNet

SentiWordNet

Análisis del
sentimiento

WordNet	https://wordnet.princeton.edu/
SentiWordNet	https://github.com/aesuli/SentiWordNet
BabelNet	https://babelnet.org/
VerbNet	https://verbs.colorado.edu/verbnet/
FrameNet	https://framenet.icsi.berkeley.edu/fndrupal/
LSA	accesibles en
Word embeddings	https://radimrehurek.com/gensim/

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

Definición

SentiWordNet

Análisis del
sentimiento

WordNet

Semántica

WordNet

Definición

SentiWordNet

Análisis del
sentimiento

- Base de datos léxica para el inglés (otros - EuroWordNet)
- Contiene solo clases abiertas de palabras
- Define grupos de palabras sinónimas (*synsets*)
- Cada *synset* tiene una glosa y ejemplos
- Define relaciones léxicas entre *synsets*

<http://wordnetweb.princeton.edu/perl/webwn>

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

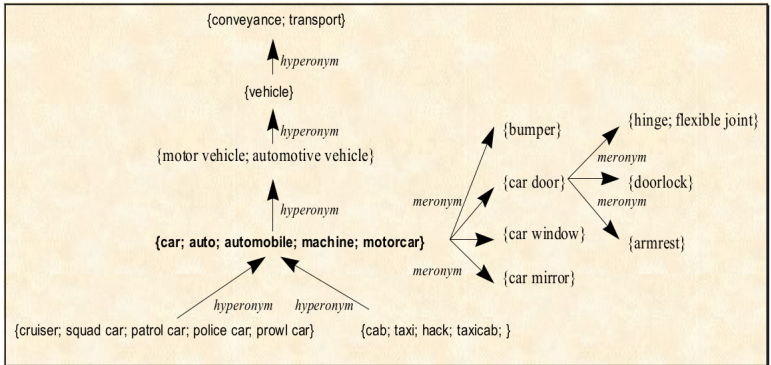
Display options for sense: (gloss) "an example sentence"

Noun

- **S: (n) age** (how long something has existed) *"it was replaced because of its age"*
- **S: (n) historic period, age** (an era of history having some distinctive feature) *"we live in a litigious age"*

Relaciones Léxicas

Ejemplo de relaciones entre *synsets*



Semántica

WordNet

Definición

SentiWordNet

Análisis del sentimiento

Relaciones léxicas

- **Sinonimia**: mismo significado. P.e.Ex: age - historic_period
- **Antonimia**: significado opuesto. P.e.: dark - light
- **Homofonía**: mismo sonido. P.e.: son - sun
- **Homografía**: misma forma escrita. P.e.: lead (nombre - verbo)
- **Polisemia**: diferentes sentidos relacionados. P.e.: newspaper (periodico - firma)
- **Homonimia**: diferentes sentidos no relacionados. P.e.: position (lugar - estatus)
- **Hiperonimia**: padre. P.e.: cat - feline
- **Hiponimia**: hijo. P.e.: feline - cat
- **Holonimia**: grupo, todo. P.e.: student - class
- **Meronimia**: miembro, parte. P.e.: class - student
- **Metonimia**: sustitución de la entidad. P.e.: We ordered many delicious dishes at the restaurant.

Semántica

WordNet

Definición

SentiWordNet

Análisis del
sentimiento

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

Similaridad

SentiWordNet

Análisis del
sentimiento

Similaridad en WordNet

- Shortest Path Length: $Sim(s_1, s_2) = \frac{1}{SPL(s_1, s_2)}$
donde $SPL(s_1, s_2)$ = longitud del camino mínimo de s_1 a s_2 en número de vértices

Semántica

WordNet

Similaridad

SentiWordNet

Análisis del
sentimiento

Similaridad en WordNet

- Shortest Path Length: $Sim(s_1, s_2) = \frac{1}{SPL(s_1, s_2)}$
donde $SPL(s_1, s_2)$ = longitud del camino mínimo de s_1 a s_2 en número de vértices
- Leacock & Chodorow: $Sim(s_1, s_2) = -\log_2 \frac{SPL(s_1, s_2)}{2 \cdot MaxDepth}$
donde $depth(s) = SPL(TopSynset, s)$
 $MaxDepth = \max_{s \in WN} depth(s)$

Semántica

WordNet

Similaridad

SentiWordNet

Análisis del
sentimiento

Similaridad en WordNet

- Shortest Path Length: $Sim(s_1, s_2) = \frac{1}{SPL(s_1, s_2)}$
donde $SPL(s_1, s_2)$ = longitud del camino mínimo de s_1 a s_2 en número de vértices
- Leacock & Chodorow: $Sim(s_1, s_2) = -\log_2 \frac{SPL(s_1, s_2)}{2 \cdot MaxDepth}$
donde $depth(s) = SPL(TopSynset, s)$
 $MaxDepth = \max_{s \in WN} depth(s)$
- Wu & Palmer:
 $Sim(s_1, s_2) = \frac{2 \cdot depth(LCS(s_1, s_2))}{depth_{LCS(s_1, s_2)}(s_1) + depth_{LCS(s_1, s_2)}(s_2)}$
donde $LCS(s_1, s_2)$ = ancestro común más cercano a s_1 y s_2
 $depth_{s'}(s) = SPL(TopSynset, s)$ a través de s'

Semántica

WordNet

Similaridad

SentiWordNet

Análisis del
sentimiento

Similaridad en WordNet

- Shortest Path Length: $Sim(s_1, s_2) = \frac{1}{SPL(s_1, s_2)}$
donde $SPL(s_1, s_2)$ = longitud del camino mínimo de s_1 a s_2 en número de vértices
- Leacock & Chodorow: $Sim(s_1, s_2) = -\log_2 \frac{SPL(s_1, s_2)}{2 \cdot MaxDepth}$
donde $depth(s) = SPL(TopSynset, s)$
 $MaxDepth = \max_{s \in WN} depth(s)$
- Wu & Palmer:
 $Sim(s_1, s_2) = \frac{2 \cdot depth(LCS(s_1, s_2))}{depth_{LCS(s_1, s_2)}(s_1) + depth_{LCS(s_1, s_2)}(s_2)}$
donde $LCS(s_1, s_2)$ = ancestro común más cercano a s_1 y s_2
 $depth_{s'}(s) = SPL(TopSynset, s)$ a través de s'
- Lin: $Sim(s_1, s_2) = \frac{2 \cdot IC(LCS(s_1, s_2))}{IC(s_1) + IC(s_2)}$
donde $IC(s) = -\log_2 P(s)$ = contenido de información de s (a partir de frecuencias en un corpus)

Semántica

WordNet

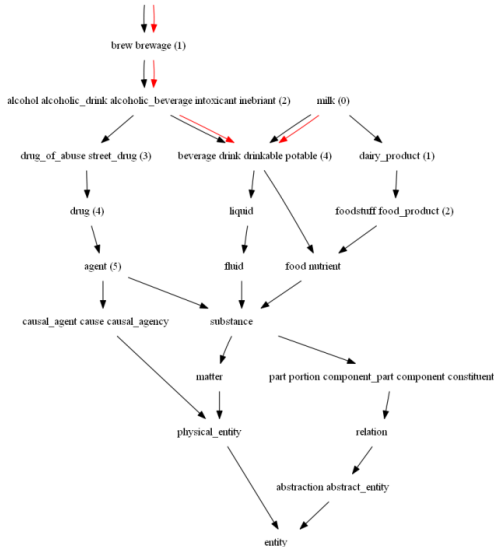
Similaridad

SentiWordNet

Análisis del
sentimiento

Ejemplo / ejercicio

Sandipan Dey (UMBC) beer (0)



$$spl(beer, milk) = 5$$

$$Sim_{spl}(beer, milk) = 0.2$$

$$Sim_{wp}(beer, milk) = 0.75$$

$$Sim_{spl}(drug, milk)?$$

$$Sim_{wp}(drug, milk)?$$

Semántica

WordNet

Similaridad

SentiWordNet

Análisis del sentimiento

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Definición

Extensión de WordNet con 3 datos añadidos:

- Score_positivo
- Score_negativo
- $\text{Score_objetivo} = 1 - \text{Score_positivo} - \text{Score_negativo}$

Wordnet		SentiWordnet		
Synsets antónimos	Glosa	obj	pos	neg
bad.a.01	having undesirable or negative qualities	0.375	0.0	0.625
good.a.01	having desirable or positive qualities. . .	0.25	0.75	0.0
bad.n.01	that which is below standard or expectations as of ethics or decency	0.125	0.0	0.875
good.n.03	that which is pleasing, valuable, useful	0.375	0.625	0.0

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Definición

Definición

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Definición

Subtareas:

- **Detección de opinión/subjetividad**: dado un texto ¿es un texto objetivo o subjetivo? Grado de subjetividad

Definición

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Definición

Subtareas:

- **Detección de opinión/subjetividad**: dado un texto ¿es un texto objetivo o subjetivo? Grado de subjetividad
- **Clasificación de polaridad**: dado un texto subjetivo, ¿expresa una opinión positiva o negativa? Grado de positividad/negatividad

Definición

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Definición

Subtareas:

- **Detección de opinión/subjetividad**: dado un texto ¿es un texto objetivo o subjetivo? Grado de subjetividad
- **Clasificación de polaridad**: dado un texto subjetivo, ¿expresa una opinión positiva o negativa? Grado de positividad/negatividad
- **Extracción de la opinión**: dado un texto subjetivo, extraer el foco de la opinión (tripletas <entidad, aspecto, polaridad>).

Índice

- 1 Semántica
 - Motivación de la semántica léxica
 - Recursos
- 2 WordNet
 - Definición
 - Similaridad
- 3 SentiWordNet
- 4 Análisis del sentimiento
 - Definición
 - Ejemplos de métodos

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Ejemplos de métodos

Análisis del sentimiento no supervisado

Posible método simple:

$$h(D) = \sum_{s \in \hat{D}} score(s)$$

\hat{D} puede ser el conjunto de *synsets* de adjetivos, o de nombres y adjetivos, o de nombres, verbos, adjetivos y adverbios.

- Detección de opinión/subjetividad:

$$score(s) = 1 - obj_s \quad \text{or} \quad score(s) = obj_s$$

- Clasificación de polaridad:

$$score(s) = pos_s - neg_s$$

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Ejemplos de métodos

Análisis del sentimiento no supervisado

Posible método simple:

$$h(D) = \sum_{s \in \hat{D}} score(s)$$

\hat{D} puede ser el conjunto de *synsets* de adjetivos, o de nombres y adjetivos, o de nombres, verbos, adjetivos y adverbios.

- Detección de opinión/subjetividad:

$$score(s) = 1 - obj_s \quad \text{or} \quad score(s) = obj_s$$

- Clasificación de polaridad:

$$score(s) = pos_s - neg_s$$

Pros:

- no necesita corpus de entrenamiento

Cons:

- puede producir resultados por debajo del estado del arte
- requiere etiquetado POS y desambiguación de sentidos

Análisis del sentimiento supervisado

Posible método simple:

Bag of words con Naïve Bayes

$$h(D) = h(w_1, \dots, w_n) = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^n P(w_i|y)$$

donde y es la categoría (pos/neg, subj/obj) y w_1, \dots, w_n es el *bag of words* relativo a D

- Dado un corpus de entrenamiento, C tal que

$$C = Y_1 \cup Y_2, Y_1 \cap Y_2 = \emptyset, Y_i = \{d_j | d_j \text{ de categoría } i\}$$

- $P(y) \approx P_{MLE}(y) = \frac{|Y_i|}{|C|}$

- $P(w_i|y) \approx P_{MLE}(w_i|Y_j) = \frac{c(w_i, Y_j)}{\sum_{w_i \in Y_j} c(w_i, Y_j)}$

Análisis del sentimiento supervisado

Posible método simple:

Bag of words con Naïve Bayes

$$h(D) = h(w_1, \dots, w_n) = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^n P(w_i|y)$$

donde y es la categoría (pos/neg, subj/obj) y w_1, \dots, w_n es el *bag of words* relativo a D

- Dado un corpus de entrenamiento, C tal que

$$C = Y_1 \cup Y_2, Y_1 \cap Y_2 = \emptyset, Y_i = \{d_j | d_j \text{ de categoría } i\}$$

- $P(y) \approx P_{MLE}(y) = \frac{|Y_i|}{|C|}$
- $P(w_i|y) \approx P_{MLE}(w_i|Y_j) = \frac{c(w_i, Y_j)}{\sum_{w_i \in Y_j} c(w_i, Y_j)}$

Pros:

- resultados mejores que la no supervisión
- no requiere etiquetado POS ni desambiguación de sentidos

Cons:

- requiere corpus de entrenamiento costoso de producir

Semántica

WordNet

SentiWordNet

Análisis del sentimiento

Ejemplos de métodos

Métodos híbridos para el análisis del sentimiento

Semántica

WordNet

SentiWordNet

Análisis del
sentimiento

Ejemplos de métodos

Posible solución simple:

- Combina método basado en el conocimiento (SentiWordnet) con dos métodos de aprendizaje automático supervisado
- Output = consenso via *voting*, por ejemplo:
 - si** al menos 2 métodos responden y_i , el consenso es y_i
 - sino** el consenso es la y_i respuesta del método con más precisión en el corpus de entrenamiento

La combinación suele mejorar los resultados de los métodos individuales

Anexo

- Base del teorema de Bayes:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

- Hipotesis de independencia *Naïve* entre atributos:

$$P(y|x_1, \dots, x_n) \approx P(y) \prod_{i=1}^n P(x_i|y)$$

- *Maximum likelihood estimation* de $P(y)$ y $P(x_i|y)$
- Predicción:

$$h(x_1, \dots, x_n) = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i|y)$$

- Necesita técnicas de suavizado para evitar ceros:
(NLTK descarta los atributos nunca vistos)