Procesamiento de Lenguaje Natural

Teoría

Alcance

- Definir un proceso completo de NLP.

Recordatorio

- **Análisis sintáctico**: forma en que se combinan las palabras.
- **Análisis semántico**: significado de las palabras.
- **Análisis morfológico**: estructura interna de las palabras.

Recordatorio

- **Análisis sintáctico**: forma en que se combinan las palabras.
- **Análisis semántico**: significado de las palabras.
- **Análisis morfológico**: estructura interna de las palabras.

Tokenizacion

- Es el proceso de dividir una cadena en una lista de tokens (palabras).
- Clases:
  - Oraciones.
  - Palabras.

Tokenizing: Ejemplos

- Separación en oraciones.

  nltk.sent_tokenize()

- Separación en palabras.

  nltk.word_tokenize()

- Separación en palabras mediante expresiones.

  nltk.tokenize.regexp_tokenize()

Stopwords

- Son palabras comunes que generalmente no contribuyen al significado de una frase.

```
nltk.corpus.stopwords.words('english')
```

WordNet

- Es una base de datos léxica para el idioma inglés, es un diccionario diseñado para el procesamiento del lenguaje natural.
- Sesgo semántico.

Synsets

- Son grupos de palabras que expresan el mismo concepto.

```
nltk.corpus.wordnet.synsets('motorcar')

[Synset('car.n.01')]. hyponyms()

nltk.corpus.wordnet.synset('car.n.01')

[Synset('ambulance.n.01'),Synset('beach_wagon.n.01'),
   Synset('bus.n.04'), …
```

Campo Semántico

- Vehículos de transporte: autobús, bicicleta, tren, barco,...
- Familia: padre, madre, hijo, abuelo, nieto, tío, sobrino...
- Arboles: pino, ciprés, naranjo, abeto,...

Monosemia - Polisemia

- **Monosemia:** significado único de una palabra.

Casa.

- **Polisemia:** una misma palabra o signo lingüístico tiene varias acepciones o significados.

Cabo:

1. (masculino) Punta de tierra que penetra en el mar.
2. (masculino/femenino) Escalafón militar.
3. (masculino) Cuerda en jerga náutica.

Sinónimos - Antónimos

- Sinónimos Parciales

  alterado / nervioso

  alterado / modificado

  pesado / cansino.

- Sinónimos de Grado

  miedo / fobia / terror / pánico

  pena / tristeza / depresión

Hiponimia

- Consiste en que determinadas palabras (Hipónimos) poseen todos los rasgos semánticos de otra más general (Hiperónimo), pero que añade en su definición otros rasgos semánticos que la diferencian. Los Cohipónimos son hipónimos que comparten un mismo hiperónimo.

Hiperónimo: Día.

Hipónimos: Lunes, Martes, Miércoles, etc. o también Mañana, Tarde y Noche, etc.

Synset similarity

- cb = wordnet.synset('cookbook.n.01')
- ib = wordnet.synset('instruction_book.n.01')
- cb.wup_similarity(ib)

0.9166666666666666

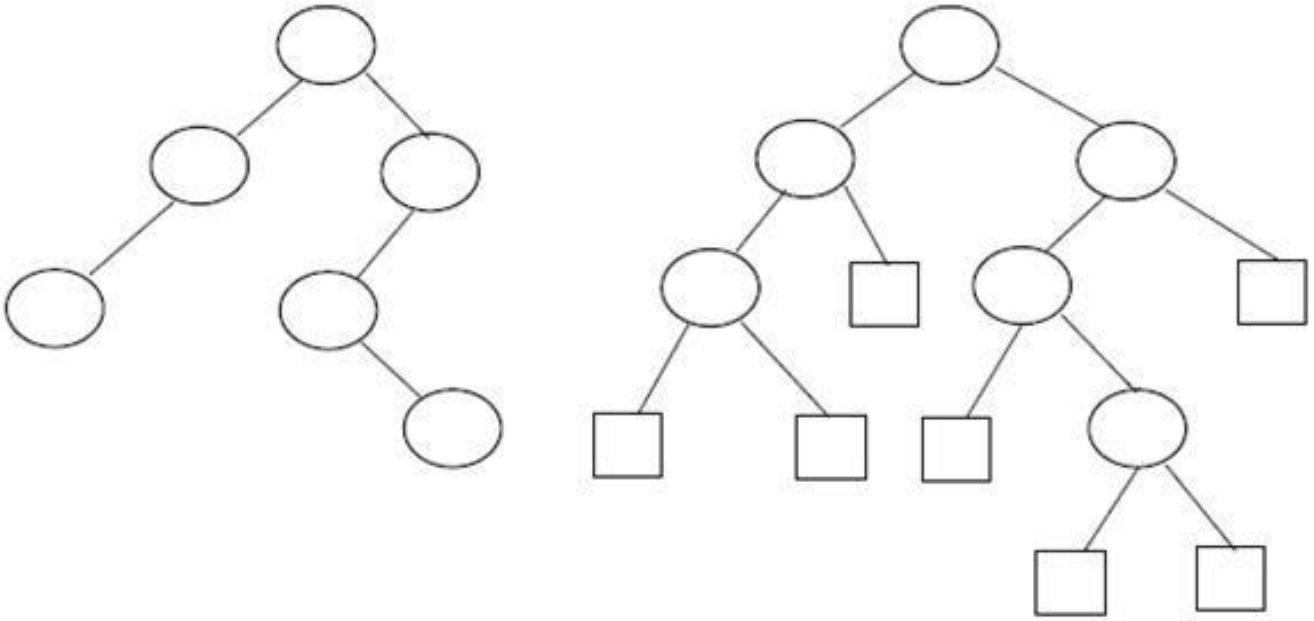
Wu-Palmer Similarity

Leacock Chordorow

Concepto: Annotations

- El agregado de relaciones a un diccionario.
- Linguistic Annotation Framework (LAF).

WordNet Abstracción



Collocations

- Son dos o más palabras que tienden a aparecer frecuentemente juntas. Sagrada Biblia.

nltk.collocations.BigramCollocationFinder.from_words()

[("'", 's'), ('arthur', ':'), ('#', '1'), ("'", 't')]

Part of speech (POS)

- Clasificación de una palabra.
  - Noun n
  - Adjective a
  - Adverb r
  - Verb v

len(wordnet.synsets('great')) = 7

len(wordnet.synsets('great', pos='n')) = 1

len(wordnet.synsets('great', pos='a')) = 6

Clasificación en español

- Sustantivos.
- Adjetivos.
- Artículos.
- Verbos.
- Adverbios.
- Preposiciones.
- Conjunciones.

POS Clasificación Completa

- VERB All verbs
- NOUN Common and proper nouns
- PRON Pronouns
- ADJ Adjectives
- ADV Adverbs
- ADP Prepositions and postpositions
- CONJ Conjunctions
- DET Determiners
- NUM Cardinal numbers
- PRT Participles
- X Other
- . Punctuation

Relacionar POS con

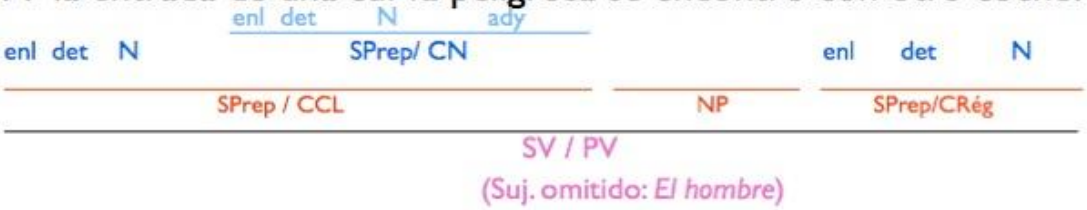
Oración 30:

Un hombre viajaba tranquilo en su coche.

Oración simple, predicativa, activa, intransitiva, enunciativa, afirmativa.

Oración 31:

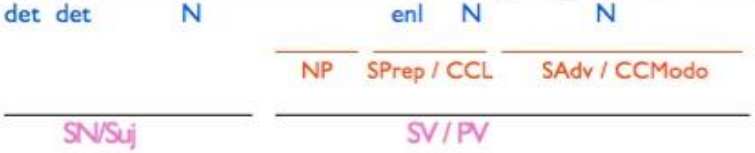
A la entrada de una curva peligrosa se encontró con otro coche.



Oración simple, predicativa, activa, intransitiva pronominal, enunciativa, afirmativa.

Oración 32:

El otro vehículo venía hacia él peligrosamente.



Oración simple, predicativa, activa, intransitiva, enunciativa afirmativa.

Stemming

- Es una técnica para eliminar los prefijos o sufijos de una palabra.
- El stem de cooking es cook.

Lancaster.

SnowballStemmer.

Análisis Morfológico

- **Familia Léxica** (Palabras Derivadas): palabras que derivan de otra a la que se le añaden Morfemas Derivativos (Prefijos, Sufijos e Interfijos).

Agua: aguacero, aguardiente, aguafiestas

Árbol: arboleda, arbolista, desarbolar

Barco: barquero, barquilla, embarcar

Análisis Morfológico

- **Los Monemas:** son las partes más pequeñas de una palabra que poseen significado. Son los Lexemas y los Morfemas.

Lexemas y Morfemas

- **Lexemas**: tienen significado léxico y constituyen la parte invariable de la palabra.

gat-o; niñ-a.

cant-ábamos.

Lexemas y Morfemas

- **Morfema**: es la unidad mínima capaz de expresar un significado gramatical.

gato: gat (lexema) + o (morfema con significado de género masculino)

niñas: niñ (lexema) + a (morfema de género femenino) + s (morf. de plural)

cantaba: cant (lexema) + aba (morfema de modo indicativo y tiempo imperfecto)

Lemmatizing

- Lemmatización es muy similar a la derivación, pero es más similar a sustitución de sinónimo.

```
nltk.corpus.wordnet.synsets('car')
['auto', 'automóvil', 'carro', 'coche', 'máquina',
 'turismo', 'vehículo']
```

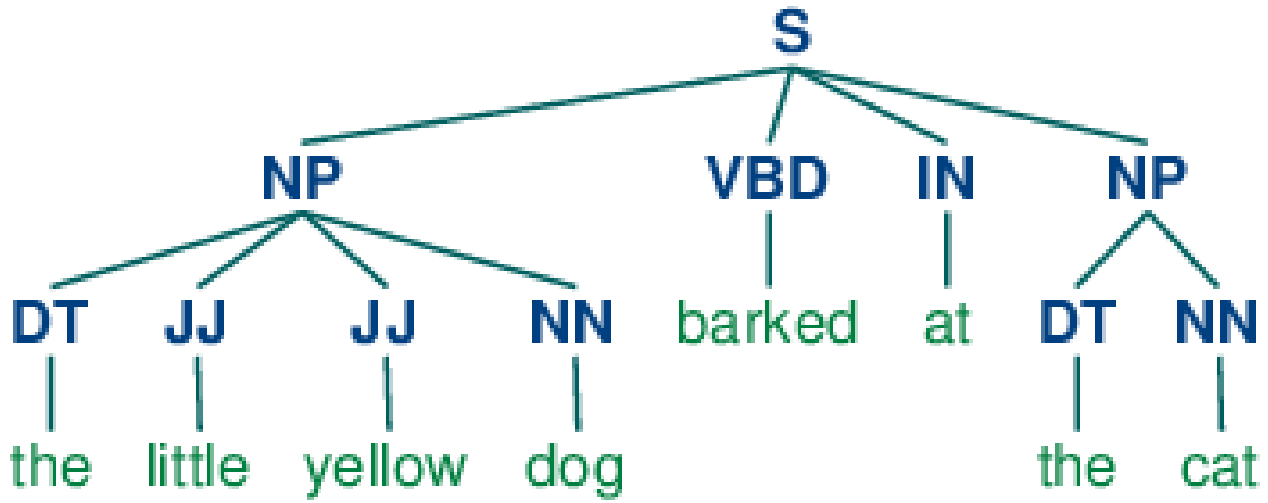
Corpora - Corpus

- Un corpus es una colección de documentos de texto.
  - Carga "Lazy".
  - Vistas Propias.
  - MongoDB.

Chunked Phrase

- Un chunk una frase dentro de una oración.
- Es una representación similar a un árbol de frases dentro de una oración.

Chunked Phrase



Chunked Phrase

Parse

```
(ROOT
  (S
    (S
      (NP
        (NP (NNP American) (NNPS Airlines))
        (, ,)
        (NP
          (NP (DT a) (NN unit))
          (PP (IN of)
            (NP (NNP AMR))))
        (, ,))
      (ADVP (RB immediately))
      (VP (VBD matched)
        (NP (DT the) (NN move))))
    (, ,)
    (NP (NNP spokeman) (NNP Tim) (NNP Wagner))
    (VP (VBD said))
    (. .)))
```

POS Tagging

- Sin las etiquetas POS, el chunk no puede saber cómo extraer frases de una oración

Entrenamientos de unigram POS

- tagger.tag_sents([['Hello', 'world', '.'], ['How', 'are', 'you', '?']])
- [
- [('Hello', 'NN'), ('world', 'NN'), ('.', 'NN')],
- [('How', 'NN'), ('are', 'NN'), ('you', 'NN'), ('?', 'NN')]
-]

Taggers con Backoff Tagging

- Le permite encadenar taggers juntos para que si un tagger no sabe cómo etiquetar una palabra.

ngram taggers

- Secuencia de n items.

Modelo likely para tags

- Para encontrar las palabras más comunes: nltk.probability.FreqDist
- Cuenta la frecuencias de palabras en el corpus.

Tagging: Tipos

- regular expressions.
- prefix or the suffix.

Entrenamiento: Brill tagger

- Serie de reglas para corregir los resultados de un etiquetador inicial.
- Estas reglas son basadas en cuántos errores corrigen menos el número de nuevos errores produce.

Entrenamiento: TnT tagger

- Mantiene un número de
  - FreqDist
  - ConditionalFreqDist
- Estas distribuciones de frecuencia cuentan unigrams, bigrams y trigrams.

Tagging de nombres propios

- Se crea una lista con nombres propios.
  - Clase: NamesTagger

NLTK-Trainer

- http://nltk-trainer.readthedocs.io/en/latest/.

Extracting Chunks

- Proceso de extraer frases de una oración.
- La idea es que las frases significativas se pueden extraer de una oración buscando patrones particulares.

Chunking and chinking con Regex

- Usando expresiones regulares modificadas, se pueden definir patrones.

Extraccion de Named Entities

- Pieza de información de nuestro interés.
  - Nombres.
  - Lugares.

Transformación: Chunks

- Las transformaciones son correcciones gramaticales y reorganización de las frases sin pérdida de significado.

Filtrado

- Eliminar palabras sin perder el significado de la oración.

  Stopwords

Cambios en las oraciones

- Frases verbales.
- Cardinales.
- Frases en infinitivo.
- De plural a singular.

Text Classification

- Que es.
- Ejemplo simple.
- Concepto de "bag of words"

Clasificador: Naive Bayes

- Clasificador muy usado en NLP.

Características: Naive Bayes

- Si hay pocos datos de entrenamiento.
- High-bias: tipo de algoritmo que no sobre ajusta demasiado

Otros Clasificadores

- Support Vector Machines.
- Regularized logistic regression.
- Arboles de decisión.
- Scikit-learn.
- Maximum entropy o logistic regression.

Entropía

- La entropía es la incertidumbre del resultado.

Precision and Recall II

- Precisión:

  total de extracciones correctas / total de extracciones.

- Recall:

  total de extracciones correctas / total de relaciones existentes.

High Information Words

- Es una palabra fuertemente sesgada hacia una única etiqueta de clasificación.
- Método:
  - high_information_words()

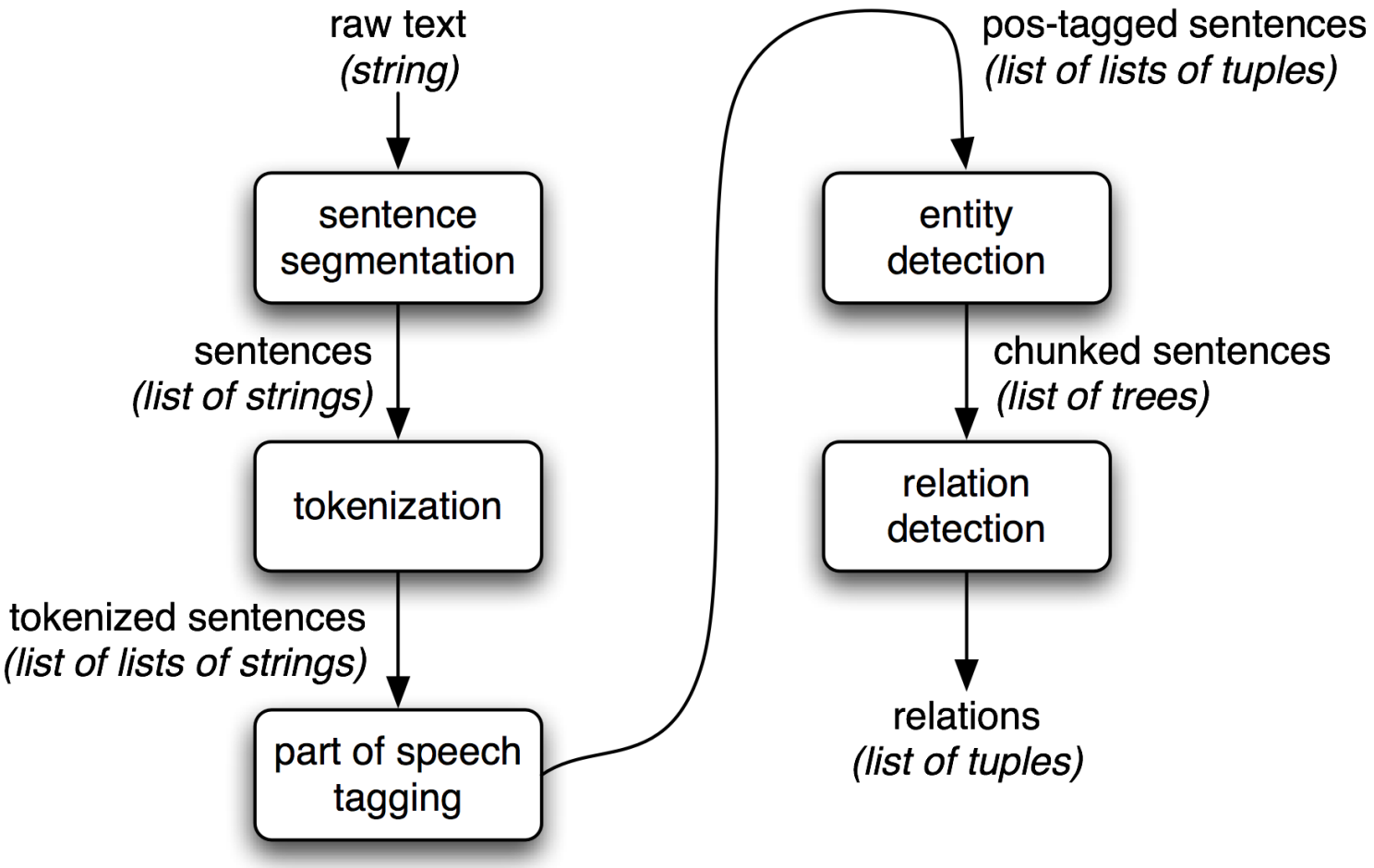
Combinación de Clasificadores

- Una forma de mejorar el rendimiento de la clasificación es combinar clasificadores.
- Esto significa combinar al menos tres clasificadores juntos

Distribucion de procesos

- Execnet.
- Apache Storm.
- Akka.io.

Proceso: NLP


raw text
(string)
sentence segmentation
sentences
(list of strings)
tokenization
tokenized sentences
(list of lists of strings)
part of speech tagging
pos-tagged sentences
(list of lists of tuples)
entity detection
chunked sentences
(list of trees)
relation detection
relations
(list of tuples)

Gracias