

DISEÑO E IMPLEMENTACIÓN DE UN ANALIZADOR DE DEPENDENCIAS PARA PROCESAMIENTO DE LENGUAJE NATURAL EN ESPAÑOL

Alejandro Alcalde ¹

¹Grado Ingeniería Informática
Universidad de Granada

19 de diciembre de 2016

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

- Falta de Software Español
- Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

- Algoritmo
- Resultados
- Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

■ Falta de Software Español

■ Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

■ Algoritmo

■ Resultados

■ Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

SOPORTE DE IDIOMAS EN PIPELINES ACTUALES

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

ANNOTATOR	AR	ZH	EN	FR	DE	ES
Tokenize/Segment	✓	✓	✓	✓		✓
Sentence Split	✓	✓	✓	✓	✓	✓
Part of Speech	✓	✓	✓	✓	✓	✓
Lemma			✓			
Named Entities		✓	✓		✓	✓
Constituency Parsing	✓	✓	✓	✓	✓	✓
Dependency Parsing		✓	✓	✓	✓	
Sentiment Analysis			✓			
Mention Detection		✓	✓			
Coreference		✓	✓			
Open IE			✓			

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

- Falta de Software Español

- Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

- Algoritmo

- Resultados

- Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

QUÉ ES EL NLP

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo
Resultados
Implementación

Conclusiones

DEFINICIÓN

Ciencia que estudia la computación lingüística.

- Resúmenes.
- Traducción automática.
- Reconocimiento de voz.
- Sistemas de Diálogo Hablado.
- Clasificación de documentos.
- Análisis de sentimientos.

NIVELES DE ANÁLISIS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo
Resultados
Implementación

Conclusiones

- Documento.
- Sentencia.
- Entidad.

OBJETIVOS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

- Revisión bibliográfica.
- Elección de un parseador y diseño para SCALA.
- Implementación y TDD.
- Evaluación y comparación de resultados.

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

- Falta de Software Español
- Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

- Algoritmo
- Resultados
- Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

ALGORITMO SELECCIONADO PARA ESPAÑOL

STATISTICAL DEPENDENCY ANALYSIS WITH SVMs

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español
Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo
Resultados
Implementación

Conclusiones

```
1: Input Sentence:  $(w_1, p_1), (w_2, p_2), \dots, (w_n, p_n)$ 
2: Initialize:
3:    $i \leftarrow 1$ 
4:    $\mathcal{T} \leftarrow \{(w_1, p_1), (w_2, p_2), \dots, (w_n, p_n)\}$ 
5:   no_construction  $\leftarrow$  true
6: while  $|\mathcal{T}| \geq 1$  do
7:   if  $i == |\mathcal{T}|$  then
8:     if no_construction == true then break
9:     end if
10:    no_construction  $\leftarrow$  true
11:     $i \leftarrow 1$ 
12:  else
13:     $\mathbf{x} \leftarrow$  getContextualFeatures( $\mathcal{T}, i$ )
14:     $y \leftarrow$  estimateAction(model,  $\mathbf{x}$ )
15:    construction( $\mathcal{T}, i, y$ )
16:    if  $y ==$  Left or Right then no_construction  $\leftarrow$  false
17:    end if
18:  end if
19: end while
```

SVMs

QUÉ ES UNA SVM

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

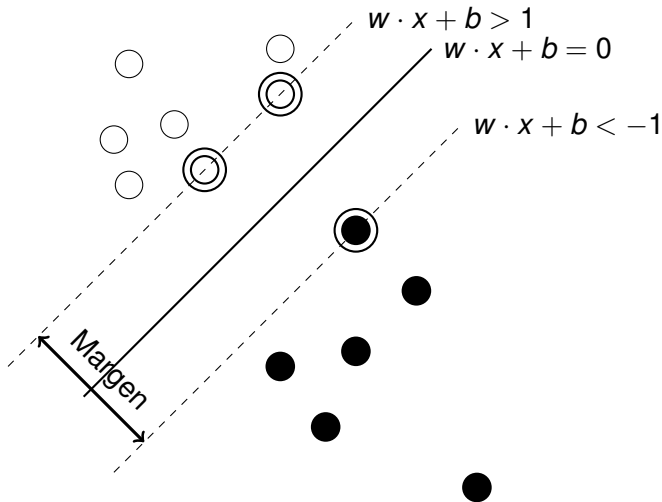
Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

- Gran poder de generalización.
- Con el *Kernel Trick* se combinan características.

ACCIÓN DESPLAZAR

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

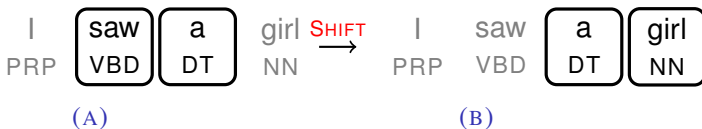


FIGURA 1: DESPLAZAR. (a) Antes. (b) Después

ACCIÓN DERECHA

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

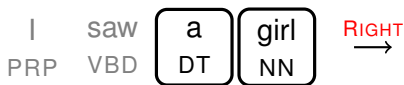
Resolución
del Trabajo

Algoritmo

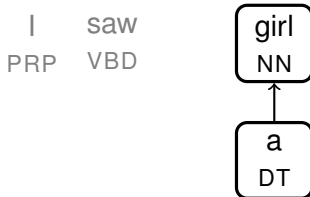
Resultados

Implementación

Conclusiones



(A)



(B)

FIGURA 2: DERECHA. (a) Antes. (b) Después.

ACCIÓN IZQUIERDA

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

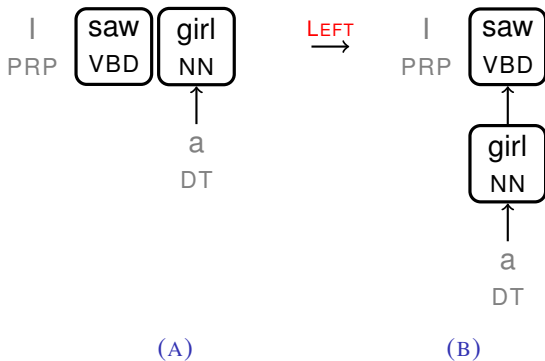


FIGURA 3: IZQUIERDA. (a) Antes. (b) Después

EJEMPLO – “*Sobre la oferta de IBM*”

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



EJEMPLO – “*Sobre la oferta de IBM*”

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

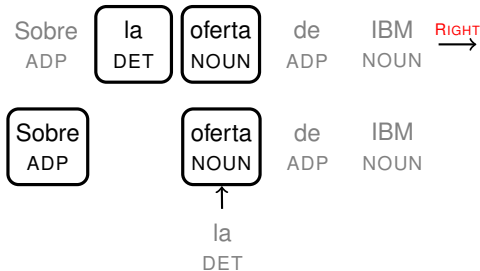
Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



EJEMPLO – “*Sobre la oferta de IBM*”

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

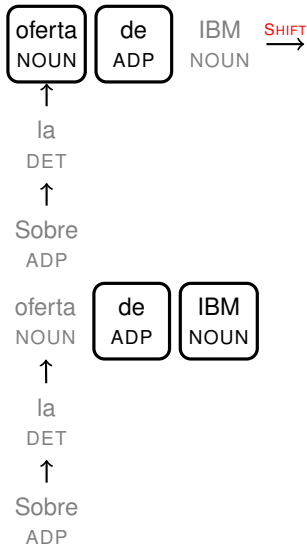
Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



EJEMPLO – “*Sobre la oferta de IBM*”

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

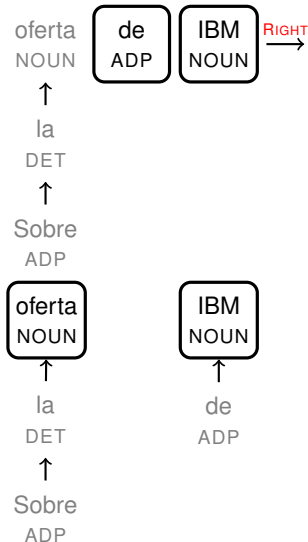
Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



SELECCIÓN DE CARACTERÍSTICAS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

Tipo	Valor
pos	POS <i>tag</i>
lex	La palabra
ch-L-pos	Nodo hijo modificando al padre por la izda.
ch-L-lex	Palabra del correspondiente ch-L-pos
ch-R-pos	Nodo hijo modificando al padre por la drcha
ch-R-lex	Palabra del correspondiente ch-R-pos

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

- Falta de Software Español

- Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

- Algoritmo

- **Resultados**

- Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

RESULTADOS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

Kernel: $(x' \cdot x'' + 1)^2$, Ctx: (2, 4)	TFG	ROHIT
Dep. Acc.	76 %	75 %
Root Acc.	67 %	70 %
Comp. Rate	15 %	11 %

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

1 MOTIVACIÓN E INTRODUCCIÓN

- Falta de Software Español

- Introducción al NLP

2 OBJETIVOS

3 RESOLUCIÓN DEL TRABAJO

- Algoritmo

- Resultados

- Implementación

4 CONCLUSIONES Y VÍAS FUTURAS

IMPLEMENTACIÓN

PLANIFICACIÓN

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

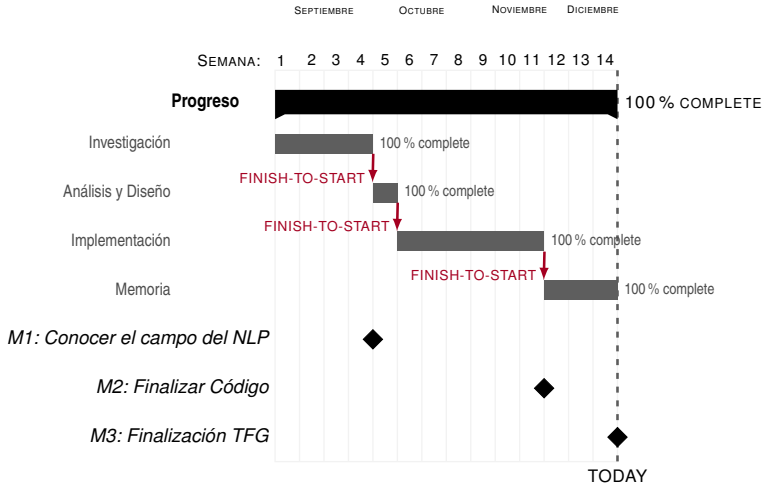
Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones



IMPLEMENTACIÓN

POR QUÉ EN SCALA

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

- Programación OO.
- Programación **Funcional**.
- Sintaxis breve.
- Escalable.
- Implementa algunos patrones.
- TRAITS.
- Amplio abanico reglas de visibilidad.

IMPLEMENTACIÓN

VENTAJAS DE SCALA

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo
Resultados
Implementación

Conclusiones

```

class WordCountMapper extends MapReduceBase
implements Mapper<IntWritable, Text, Text,
    ↪ IntWritable> {

    static final IntWritable one = new
    ↪ IntWritable(1);
    // Value will be set in a non-thread-safe
    ↪ way!
    static final Text word = new Text();

    @Override
    public void map(IntWritable key, Text
    ↪ valueDocContents,
    OutputCollector<Text, IntWritable> output,
    ↪ Reporter reporter) {
        String[] tokens = valueDocContents
        ↪ .toString().split("\\s+");
        for (String wordString: tokens) {
            if (wordString.length > 0) {
                word.set(wordString.toLowerCase());
                output.collect(word, one);
            }
        }
    }
}

class WordCountReduce extends MapReduceBase
implements Reducer<Text, IntWritable, Text,
    ↪ IntWritable> {

    public void reduce(Text keyWord,
    ↪ java.util.Iterator<IntWritable>
    ↪ counts,
    OutputCollector<Text, IntWritable> output,
    ↪ Reporter reporter) {
        int totalCount = 0;
        while (counts.hasNext()) {
            while (counts.hasNext()) {
                totalCount += counts.next().get();
            }
            output.collect(keyWord, new
            ↪ IntWritable(totalCount));
        }
    }
}

```

```

class ScaldingWordCount(args : Args) extends
    ↪ Job(args) {
    TextLine(args("input"))
    ↪ .read
    ↪ .flatMap('line -> 'word) {
        line: String =>
        ↪ line.trim().toLowerCase().split("\\s+")
    }
    ↪ .groupBy('word){ group => group.size('count)
    ↪ .write(Tsv(args("output")))
}

```

METODOLOGÍA: TDD Y BDD

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

Test-Driven Development

- Rojo.
- Verde.
- Refactorizar.

Behavior-Driven Development

- *Given.*
- *When.*
- *Then.*

FILOSOFÍA

Aplicar **TDD** a problemas de Aprendizaje

- R2 *Value*.
- ROC y AUC.
- Matriz confusión.
- Establecer un **baseline**.
- Intentar mejorarlo en cada iteración.

TDD PARA AA

EJEMPLO

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

```
class DependencyParserCheckBaselineSpec extends Specification
  with GWT
  with StandardRegexStepParsers {def is = s2""
    When training the model, set the following baselines ${featuresBaseline.start}
      Given Train data set: es_ancora-converted-train1
      Given Test data set: es_ancora-converted-test1
      When Genenaring Vocabulary
      Then Dep. Acc should be at least: 70%
      and Root Acc should be at least: 50%
      and Comp. Acc should be at least: 3%                                ${featuresBaseline.end}
  ""
  // ...
```

CONCLUSIONES Y VÍAS FUTURAS

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

Conclusiones

- Implementado parseo para **Castellano**.
- Uso de **SVMs**.
- SCALA.
- **TDD** para AA.

Vías futuras

- Más algoritmos.
- Código 100 % funcional.
- Más fases del **Pipeline**.
- SPARK.

Dep Parsing
Castellano

A. Alcalde

Motivación e
Introducción

Falta de Software
Español

Introducción al NLP

Objetivos

Resolución
del Trabajo

Algoritmo

Resultados

Implementación

Conclusiones

Gracias por su atención



[algui91/NLP_Dependency_Parsing](https://github.com/algui91/NLP_Dependency_Parsing)