UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Automatic Data Recognition System in Natural Language

Presentado por Malte Jansen
en Universidad de Burgos — March 27, 2019
Tutor: Bruno Baruque Zanón

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. nombre tutor, profesor del departamento de nombre departamento, área de nombre área.

Expone:

Que el alumno D. Nombre del alumno, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, March 27, 2019

Vº. Bº. del Tutor:	Vº. Bº. del co-tutor:
D. nombre tutor	D. nombre co-tutor

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Named Entity Recognition (NER) is a subtask of Natural Language Processing (NLP). It focuses on extracting and classifying named entities from text. The kind of texts that will be focused on are informal texts such as tweets and reddit posts. Different existing NER tools will be evaluated on a number of datasets in order to select the best approach and configuration for informal texts.

Keywords

keywords separated by commas.

Contents

List of Figures

List of Tables

Introduction

Descripción del contenido del trabajo y del estrucutra de la memoria y del resto de materiales entregados.

Objetives

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

Theoretical concepts

This section will contain descriptions of fundamental concepts applied to the project.

3.1 Natural Language Processing(NLP)

Natural language processing is is a field of computer science and plays a big part in it's sub-field of artificial intelligence.

Some of today's use cases are information extraction, machine translation, summarization, search and human-computer interfaces.[1]

It describes techniques in which natural language is processed into a formal representation which computer can understand so they can further work with the data. In order to do that the spoken or written human language has to be analysed. It is not enough for the computer to know the significance of each word or even each sentence. Otherwise a simple dictionary could be easily used. In order to understand a text the capture of complete text correlations is necessary. This is a challenge for computer because of the complexity of human language and it's ambiguity. Computer unlike humans can't use their experience to understand a text but rather have to use artificial language algorithms and techniques. One of those techniques is called Named Entity Recognition(NER).

[1] Cf. Jason Collobert Ronan; Westan. *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning.* NEC Labs America.

3.2 Named Entity Recognition(NER)

Named entity recognition refers to the extraction of important information from a text. In named entity recognition the task is to identify which information is important and to then categorise this information. This results in named entities, that are particular terms in a text that are more informative than others or have a unique context. It can be used as a source of information for different NLP applications, such as answering questions, automatic translation or information retrieval.[2]

Figure 3.1 shows a sentence and it's possible entities.

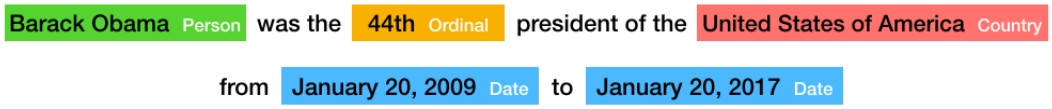

Figure 3.1: Example of entities in a phrase Source: https://github.com/floydhub/named-entity-recognition-template

3.3 Named Entity Recognition on informal Texts

The recognition of named entities in informal texts (eg.: social media posts) is a challenge. This is because they are written inherently differently than formal texts. Unlike formal texts, most social media posts do not follow strict rules, have different grammatical structures, contain misspelled words, informal abbreviations and multilingual vocabulary. They are also relatevely short. For example tweets have a 140 character limit, so not much context can be extracted. For these reasons named entity recognition is a much harder tas on informal texts.

[2]Cf. Behrang Mohit. *Natural Language Processing of Semitic Languages.* 2014, p. 221.

Techniques and tools

Esta parte de la memoria tiene como objetivo presentar las técnicas metodológicas y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Si se han estudiado diferentes alternativas de metodologías, herramientas, bibliotecas se puede hacer un resumen de los aspectos más destacados de cada alternativa, incluyendo comparativas entre las distintas opciones y una justificación de las elecciones realizadas. No se pretende que este apartado se convierta en un capítulo de un libro dedicado a cada una de las alternativas, sino comentar los aspectos más destacados de cada opción, con un repaso somero a los fundamentos esenciales y referencias bibliográficas para que el lector pueda ampliar su conocimiento sobre el tema.

Relevant aspects of the development of the project

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

Related works

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusions and future lines of work

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliography

[1] Jason Collobert Ronan; Westan. *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*. NEC Labs America. URL: https://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf.

[2] Behrang Mohit. *Natural Language Processing of Semitic Languages*. 2014, p. 221.