



Natural Language Processing and Opinion Mining

Escuela de Verano de Inteligencia Artificial

L Alfonso Ureña

Universidad
de Jaén

Sociedad
Española para el
Procesamiento
del Lenguaje
Natural



Sociedad Española para el
Procesamiento del Lenguaje Natural

OVERVIEW

Introduction

Natural Language Processing

- Definition and Concepts
- Linguistic resources
- Applications

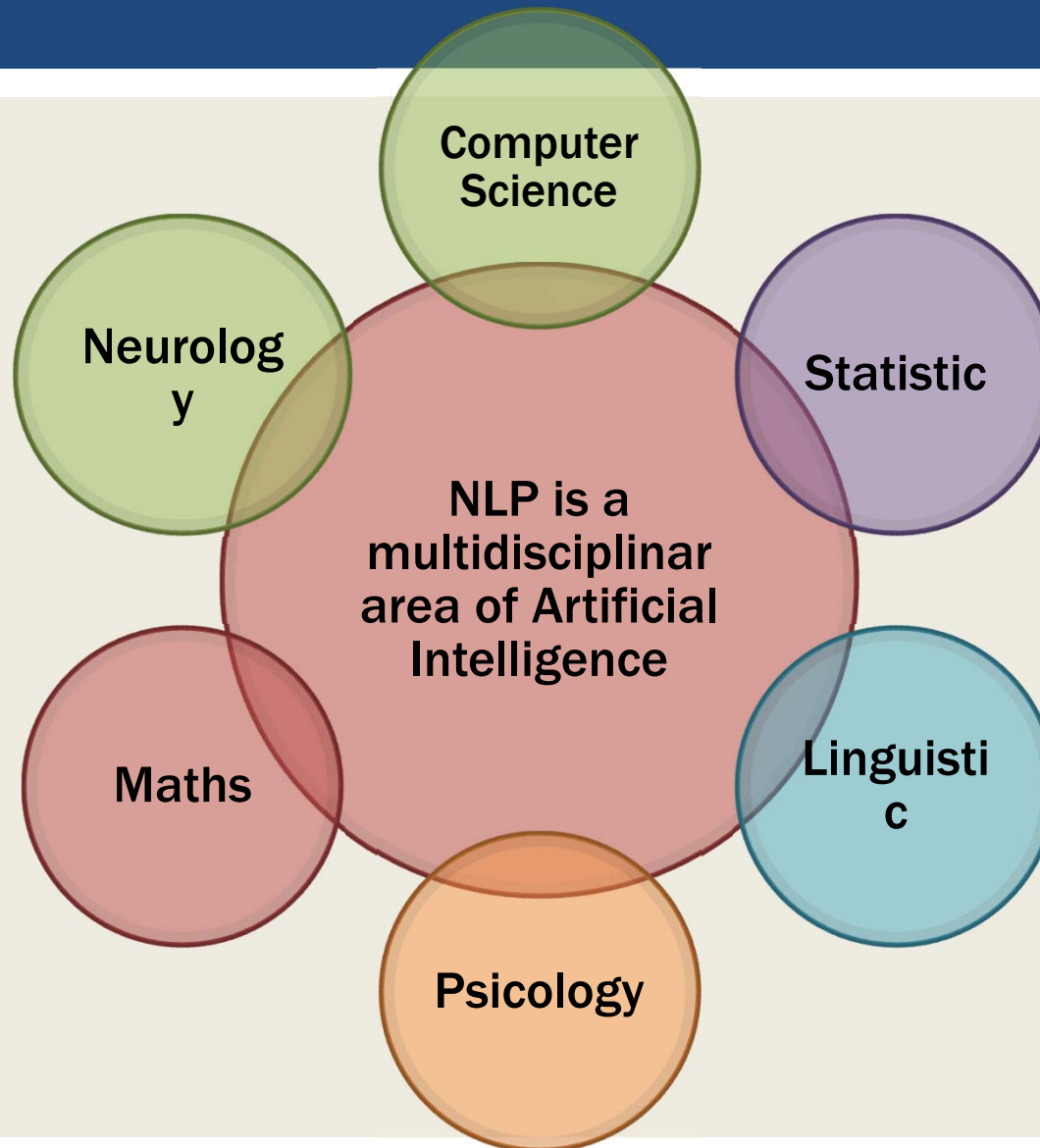
Opinion Mining

- Definition and Concepts
- New challenges and new tasks
- Applications
- Research trends



NATURAL LANGUAGE PROCESSING

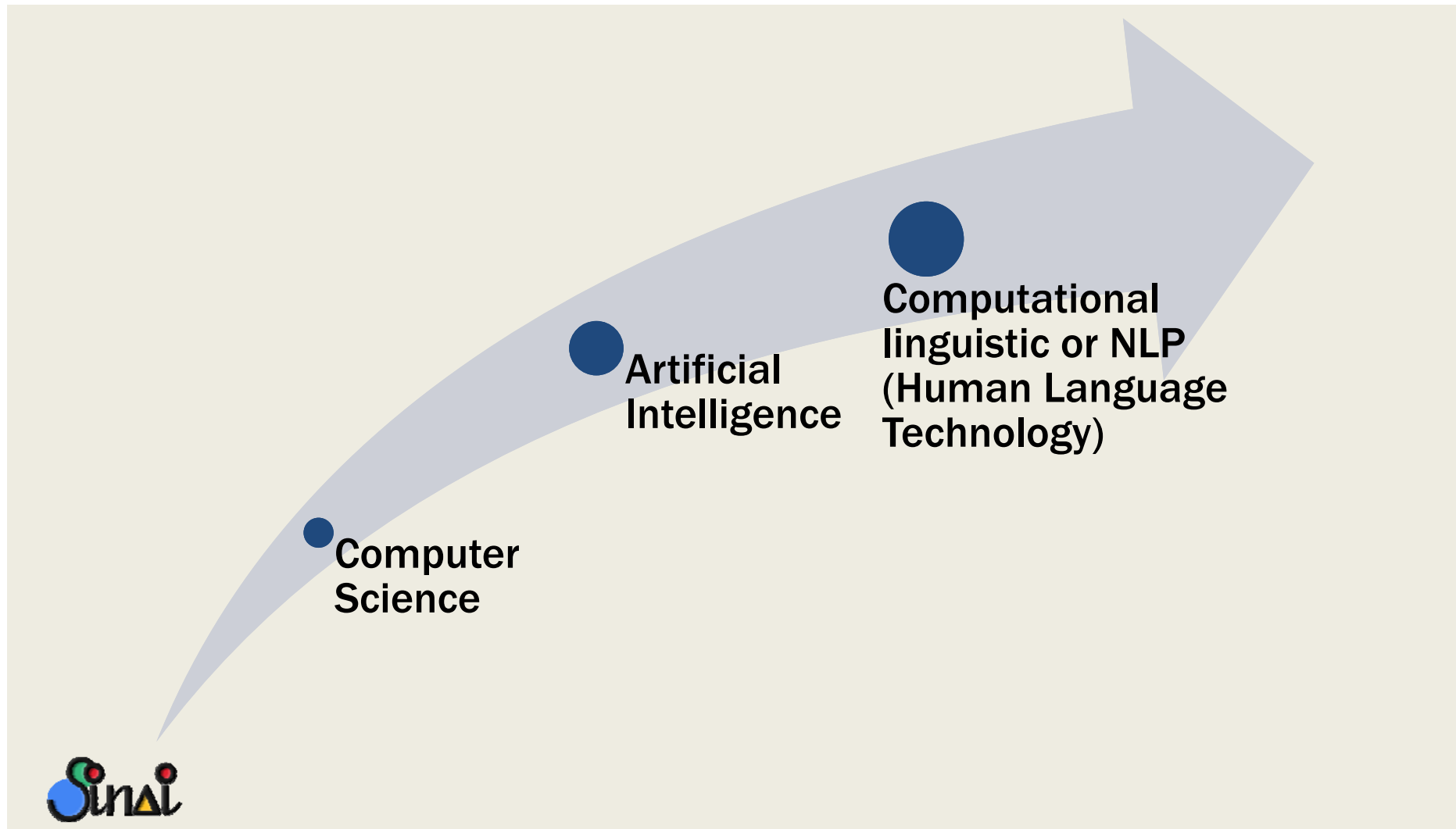
NATURAL LANGUAGE PROCESSING



**Final Goal:
Computers
can understand
humans**



WHERE DOES NLP COME FROM?



WHY IS NLP DIFFICULT?

Computers are not brains

- There is evidence that much of language understanding is built into the human brain

Computers do not socialize

- Much of language is about communicating with people

Key problems

- Representation of meaning
- Language presupposes knowledge about the world
- Language presupposes communication between people
- Language is ambiguous

WHY IS NLP DIFFICULT?

AMBIGUITY

AMBIGUITY

LANGUAGE LEVELS

Phonetics and Phonology

Ice cream vs I scream



Lexical

Polysemy: bank



Syntax

John saw the man on the mountain with a telescope



Semantic

Peter gives a cake to the children



Discourse

Referential: He ordered her to put it over that



Pragmatic

Irony, metaphor...

WHY IS NLP INTERESTING?

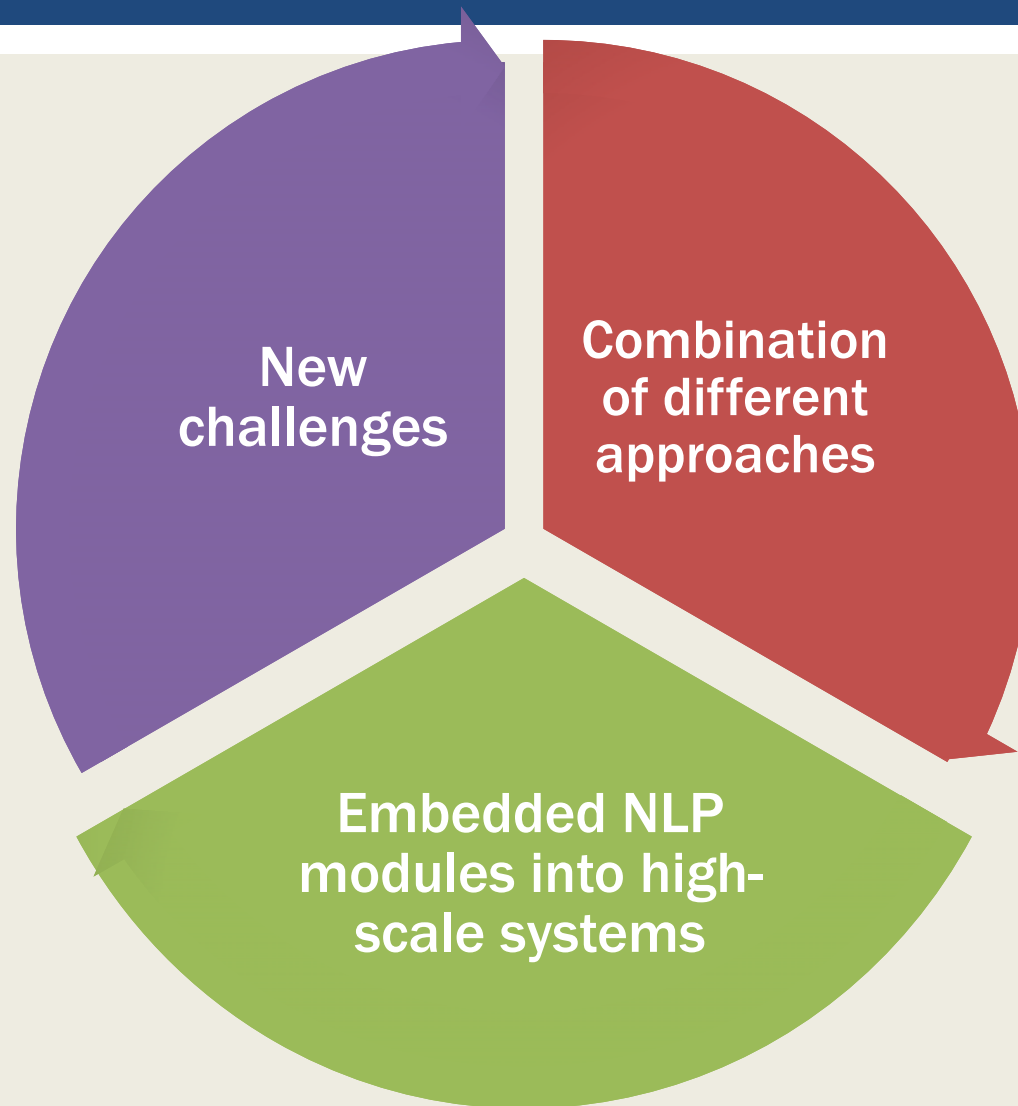
APPLICATIONS

- Spelling Suggestions/Corrections
- Grammar Checking
- Information Extraction
- Text Categorization
- Dialog Systems
- Speech Recognition
- Machine Translation
- Information Retrieval
- Question Answering
- Opinion Mining

And much more



CURRENT TRENDS



LINGUISTIC RESOURCES

LINGUISTIC RESOURCES

Most of resources are not a clear definition and sometimes the definitions are mixed one with each other

Definition

- Data set and their descriptions in electronic format to build, improve and evaluate natural language applications

Goal

- Include as much information as possible from linguistic resources or using NLP techniques in order to obtain systems more efficient

TYPES OF LINGUISTIC RESOURCES

Lexicons

Dictionaries

Ontologies

Corpora

GENERAL LEXICONS

Word repositories

- List of words, a language's inventory of lexemes...

A lexicon is the knowledge that a native speaker has about a language. This includes information about

- the form and meanings of words and phrases
- lexical categorization
- the appropriate usage of words and phrases
- relationships between words and phrases, and
- categories of words and phrases



SPECIALIZED LEXICONS

Proper nouns

Gazetteers

Jargons about any profession or activity

- Business
- Health
- Sport
- Tourism
- ...

Lexical and terminological databases



DICTIONARIES

MRD – Machine Readable Dictionary

Types: general, normative, specialized...

Depending on the number of languages:
monolingual/bilingual/multilingual

Encyclopedia

Thesaurus

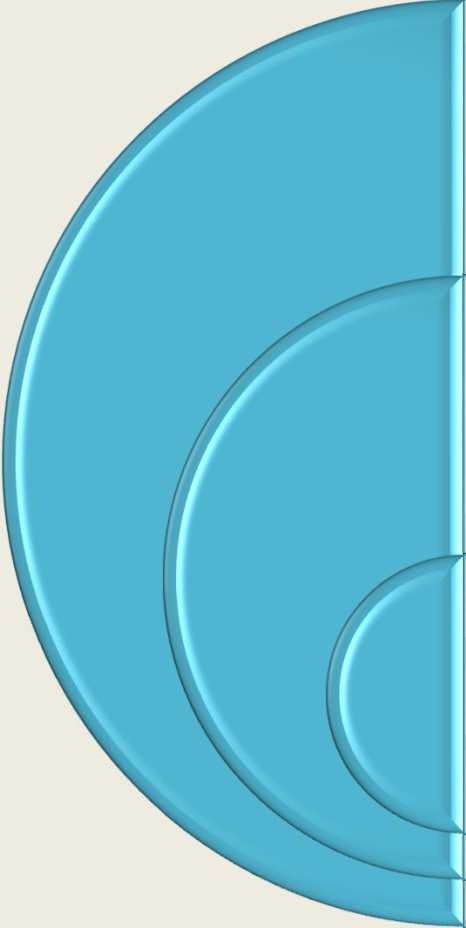
ONTOLOGIES

The term "ontology", also called Knowledge Base is not clearly defined, in fact it generates some controversy in the field of Artificial Intelligence (AI), since there is no clarity between ontology and lexicon

Definition (Gruber, 1993): An ontology is a formal, explicit specification of a shared conceptualization

An ontology formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts

CORPUS

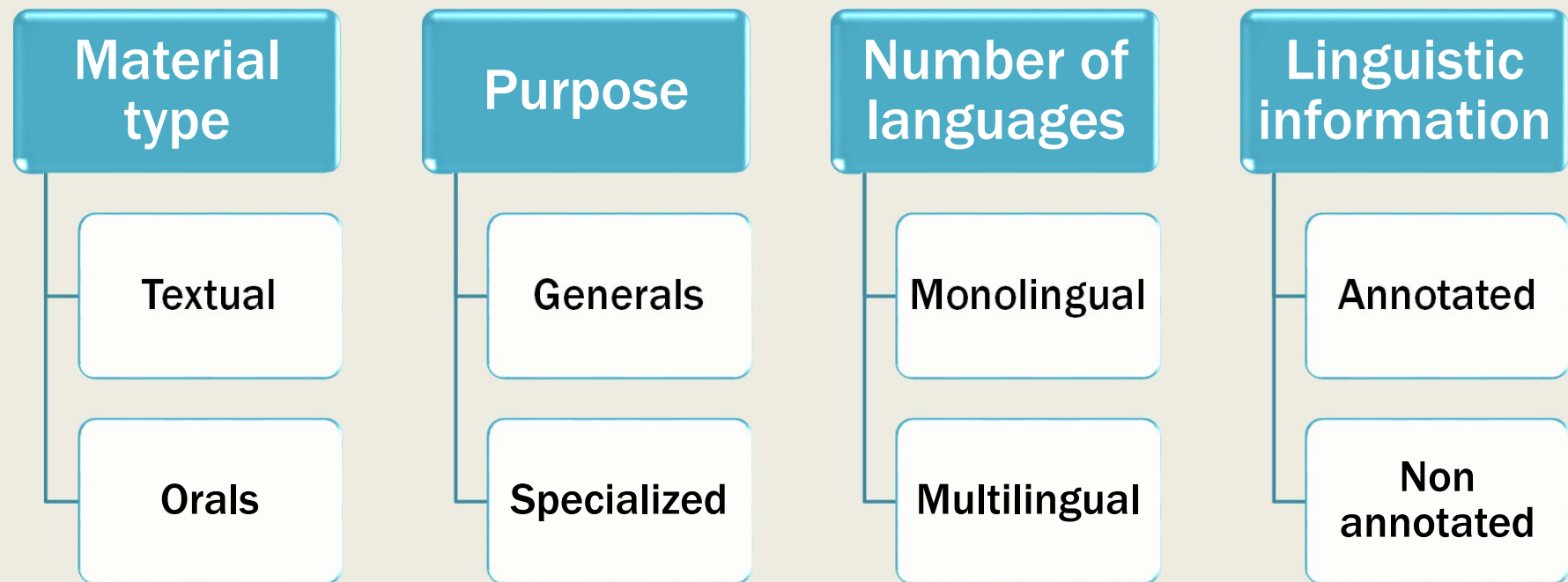


A corpus (corpora in plural) is a collection of representative documents of a language, a dialect or a subset of a language, which are used for linguistic analysis

Corpora with additional linguistic information are a valuable tool in order to integrate external information into NLP tasks

One of the most used linguistic resources

TYPES OF CORPORA



CORPUS

ACCORDING TO THE MATERIAL

Textual Corpora

- Brown Corpus
- Lancaster-Oslo/Bergen Corpus
LOB
- CREA (Corpus de Referencia del Español Actual)
- National Corpus of Polish NKJP
- Web as Corpus

Oral Corpora

- TI-DIGIT (digit recognition)
- Albayzin (different Spanish languages UPV-UPM-UGR-UAB-UPC)
- Pixi Corpora (dialogs in a book store in English and Italian)
- DCPSE (The Diachronic Corpus of Present-Day Spoken English)

CORPUS

ACCORDING TO THE PURPOSE

General Corpus

Brown Corpus

CREA

NKJP

Specialized Corpus

LEGA: Galician-Spanish
legal corpus (~6Mw)

Reuters-21578: oriented to
document categorization

Polish Corpus of Suicide
Notes (PCSN)

CORPUS

ACCORDING TO NUMBER OF LANGUAGES

Monolingual Corpus

Multilingual Corpus

- **Parallel Corpus**
 - Canadian Hansard (bilingual corpus)
 - Polyglot Bible (13 languages)
 - European parliament sessions
- **Comparable Corpus**
 - MultiText-East (articles of several journals from 6 different East-European countries)

CORPUS

ACCORDING TO THE LINGUISTIC INFORMATION

**Annotated
Corpus: Includes
additional
information with
marks or labels**

- **3LB**: syntactically annotated corpus including Spanish, Catalan and Basque
- **SemCor** (Semantic Concordance): subset from Brown Corpus annotated syntactically and semantically using WordNet
- **CRATER**: corpus with technical texts (ITU - International Telecommunications Union), morphologically annotated for 3 languages (sp, fr, en)

Non annotated

- **Brown Corpus**
- **Gutenberg project**
- **Polyglot Bible**

APPLICATIONS

APPLICATIONS OF NLP

**According to
the purpose**

**According to
the goal**

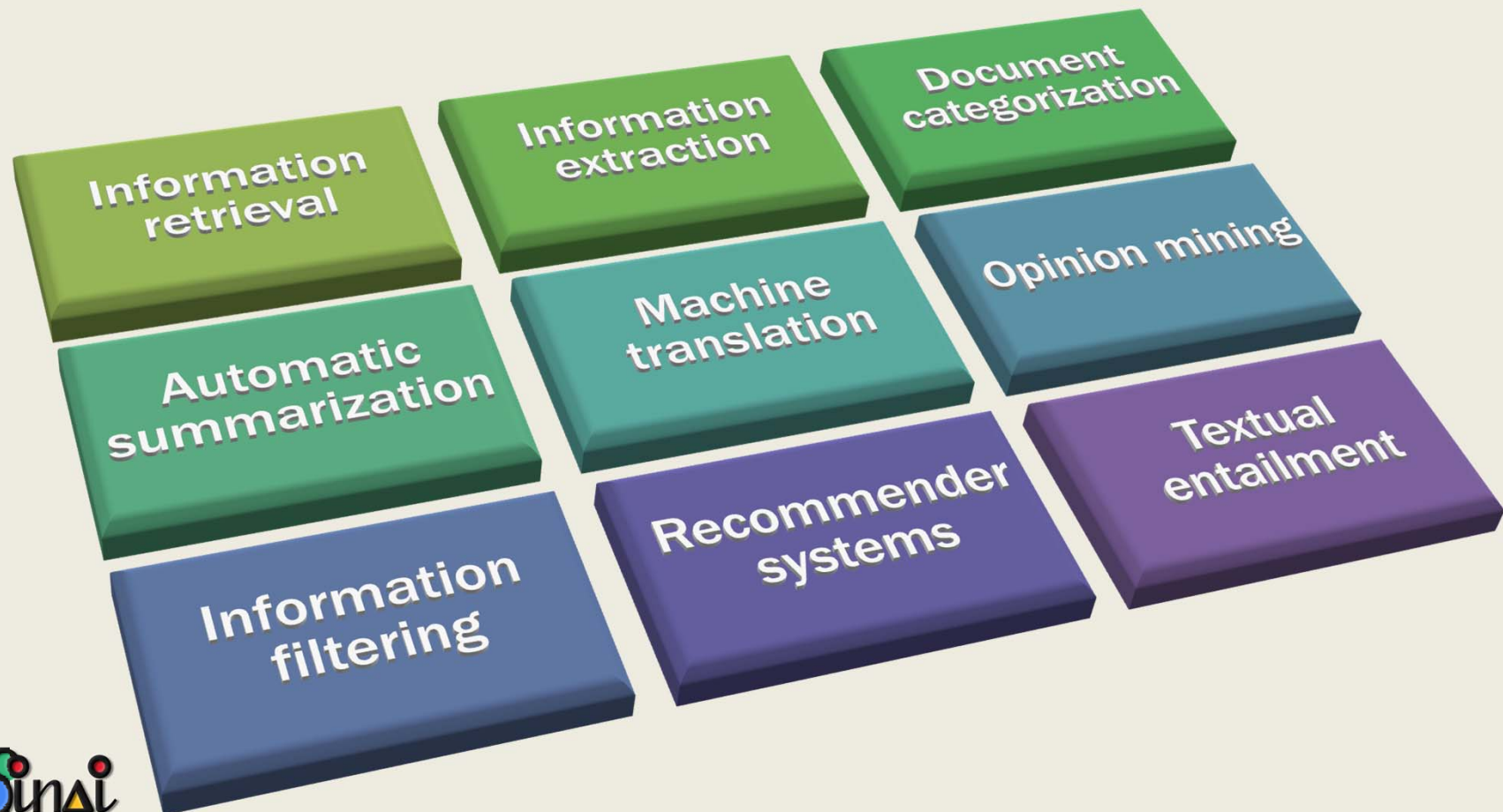
**Applications
based on
textual
treatment**

**Applications
based on
communication
human-
machine**

**No final
applications**

**Final
applications**

APPLICATIONS BASED ON TEXTUAL TREATMENT



APPLICATIONS BASED ON COMMUNICATION HUMAN-MACHINE

Applications for accessing other system

- Interfaces using natural language
- Applied to DBMS, expert systems, operating systems...

Dialog systems

- Study of human dialog
- Applied to games, tutorial systems, customer services...

APPLICATIONS ACCORDING THE GOAL

No final applications

- Word Sense disambiguation
- Entity recognition and classification

Final applications

- Information retrieval
 - Question answering systems
 - Multilingual and multimodal systems
- Document categorization
- Information filtering (spam systems)
- Information extraction
- Automatic summarization
- Text Mining
- Textual entailment
- ...

OPINION MINING & SENTIMENT ANALYSIS



MOTIVATION

What people think

Formal methods

Informal methods

Polls in electoral processes

Consumer surveys about products

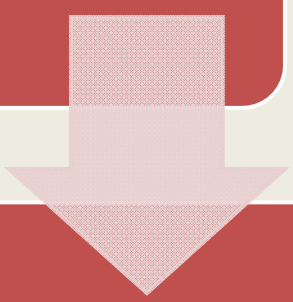
Client surveys about companies

Worker surveys on the company

Questions and chat with friends, relatives or acquaintances

MOTIVATION

Companies and businesses are realizing the importance of taking into account the opinions of Internet users about their products and services



It is necessary to build and improve systems based on the opinions from the web to help consumers choose products



**Collaborative
environment**



Subjectivity



Immediacy

OPINION MINING OR SENTIMENT ANALYSIS

Definition

Computational treatment of opinions, sentiments and subjectivity in textual documents

Other names

Sentiment classification

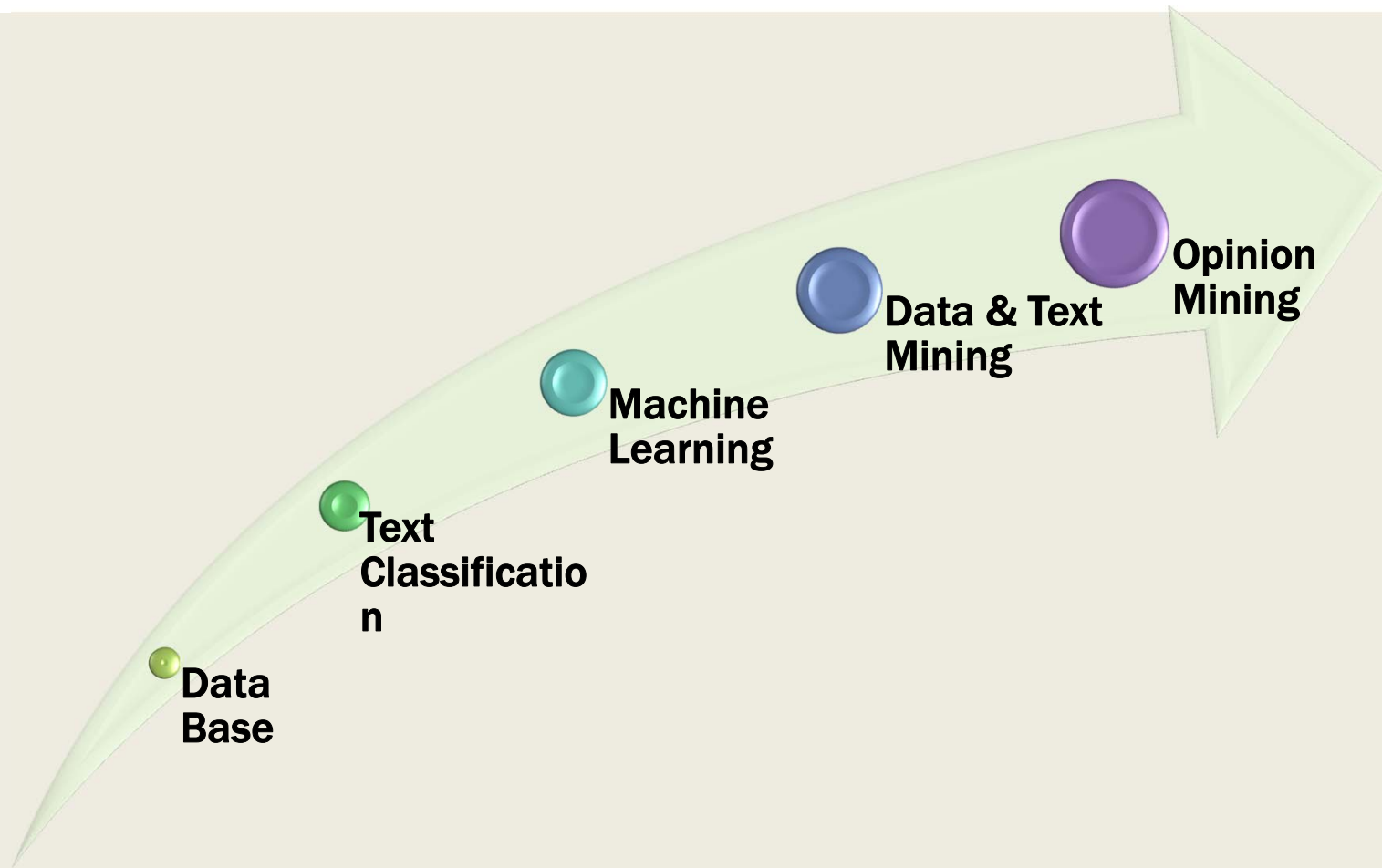
Subjectivity analysis

Review mining

Appraisal extraction

Affective computing

WHERE OM COMES FROM



COMPONENTS OF AN OPINION



Opinion holder: The person or organization that holds a specific opinion on a particular object



Object: on which an opinion is expressed



Opinion: a view, attitude, or appraisal on an object from an opinion holder



@alfonso_urena: i really like **Coruña:**)

WHY IS OPINION MINING DIFFICULT?

Inherent Ambiguity

- *“The film should be brilliant. The plot is fantastic, the characters are wonderful, the music is perfect, but the final result is not as expected”*

Dependency of the domain and context

- *“...better read the book ”*

World knowledge

- *“The director offers us another of his jewelry”*

Irony

- *“The battery is perfect to save money. You can only do one call”*

NEW CHALLENGES/NEW TASKS

Subjectivity detection

determine if user is looking for factual information or subjective

Opinionated document classification

Polarity classification (binary): identify if a subjective document includes a positive or negative opinion

Rating inference (multiclass): determine how much positive or negative a document is

Opinion extraction

Identify which documents or parts of documents containing opinions or feelings

Some sites are easy to identify the opinions but not other opinionated sites



NEW CHALLENGES/NEW TASKS

Emotion
classification

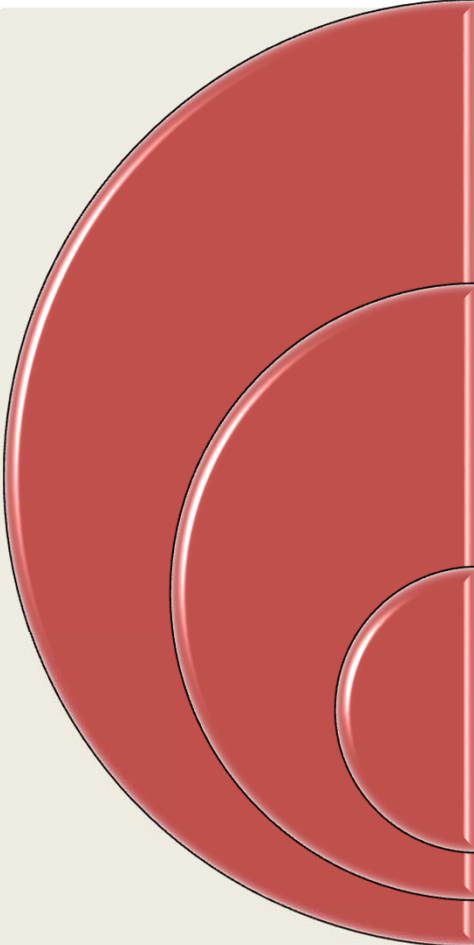
Point of view
classification

Opinion
Summarization

Humour or/and
irony detection

Detection and
tracking of
violent language

SUBJECTIVITY DETECTION



The first step should be to identify if a document contains subjective information or which parts of the document are subjectives

More difficult

J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin

“Learning subjective language”

Computational Linguistics, vol. 30, 2004

POLARITY CLASSIFICATION

Given a
opinionated
document,
determine if
the opinions
are positives
or negatives

Classification
of two
contrary
classes

*“sentiment
polarity”*
*“sentiment
polarity
classification”*
*“sentiment
classification”*

RATING INFERENCE OR DEGREE OF POSITIVITY

Multiclass Polarity classification using several levels of score



5 stars rating
(movies, music,
hotels...)

**Positive/Neutral/
Negative**
(Political strategies)

High/Medium/Low
(Product quality)

OTHER INTERESTING TASKS

Emotion classification

- Classification according to the emotion in a document: *anger, disgust, fear, happiness, sadness, surprise*
- Interesting for interfaces

Point of view classification

- Interesting in political scope
- Different point of view about Palestine problem, Irak war, USA election or Poland election

Opinion summarization

- Using a single document
- Multiple documents

OTHER INTERESTING TASKS

Textual Gender classification

- Literature, poetry, sport, news...

Humor detection

- Determine if a document includes humor content

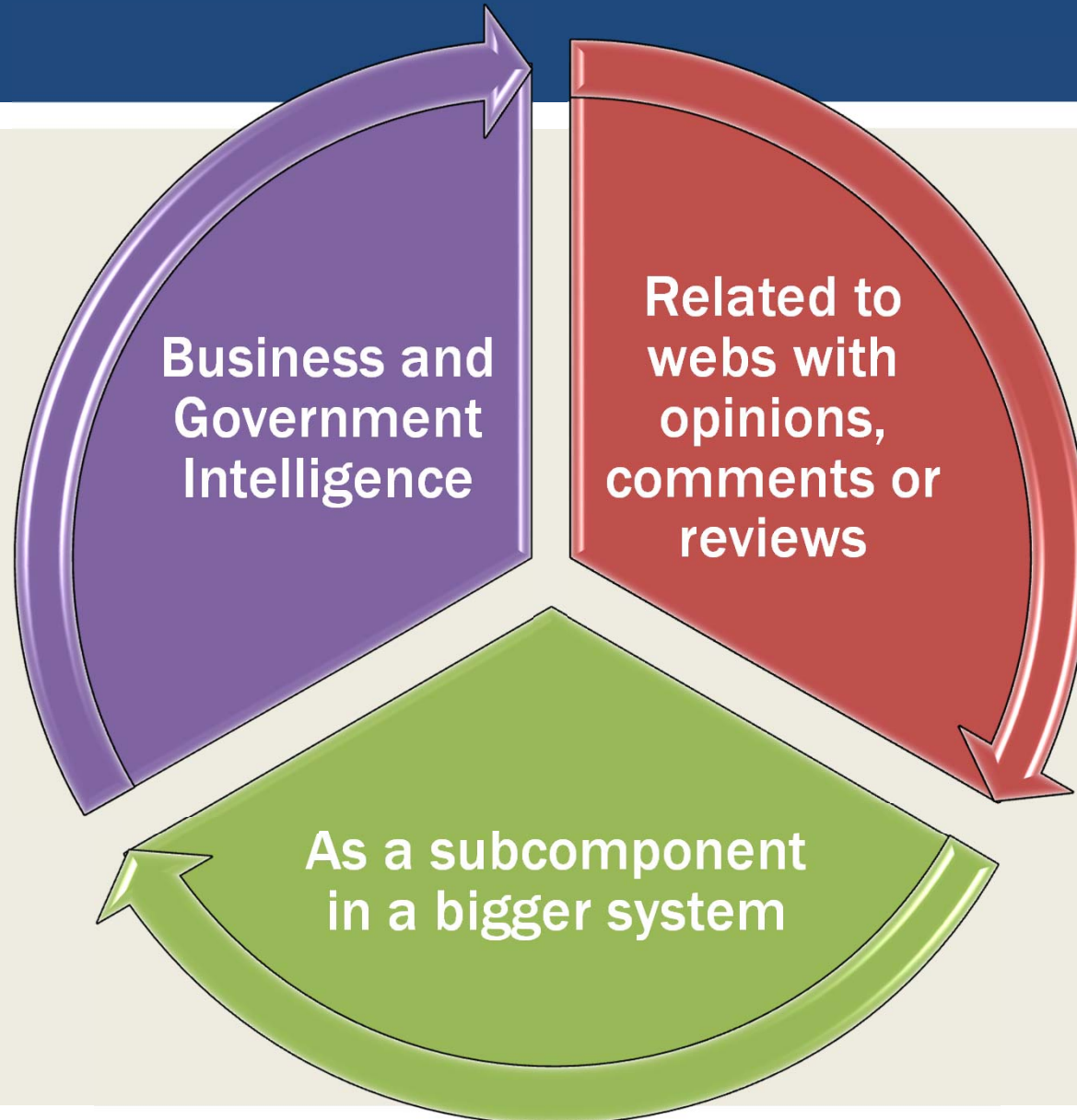
Detection and tracking of violent language

- Terrorist attacks, bullying...

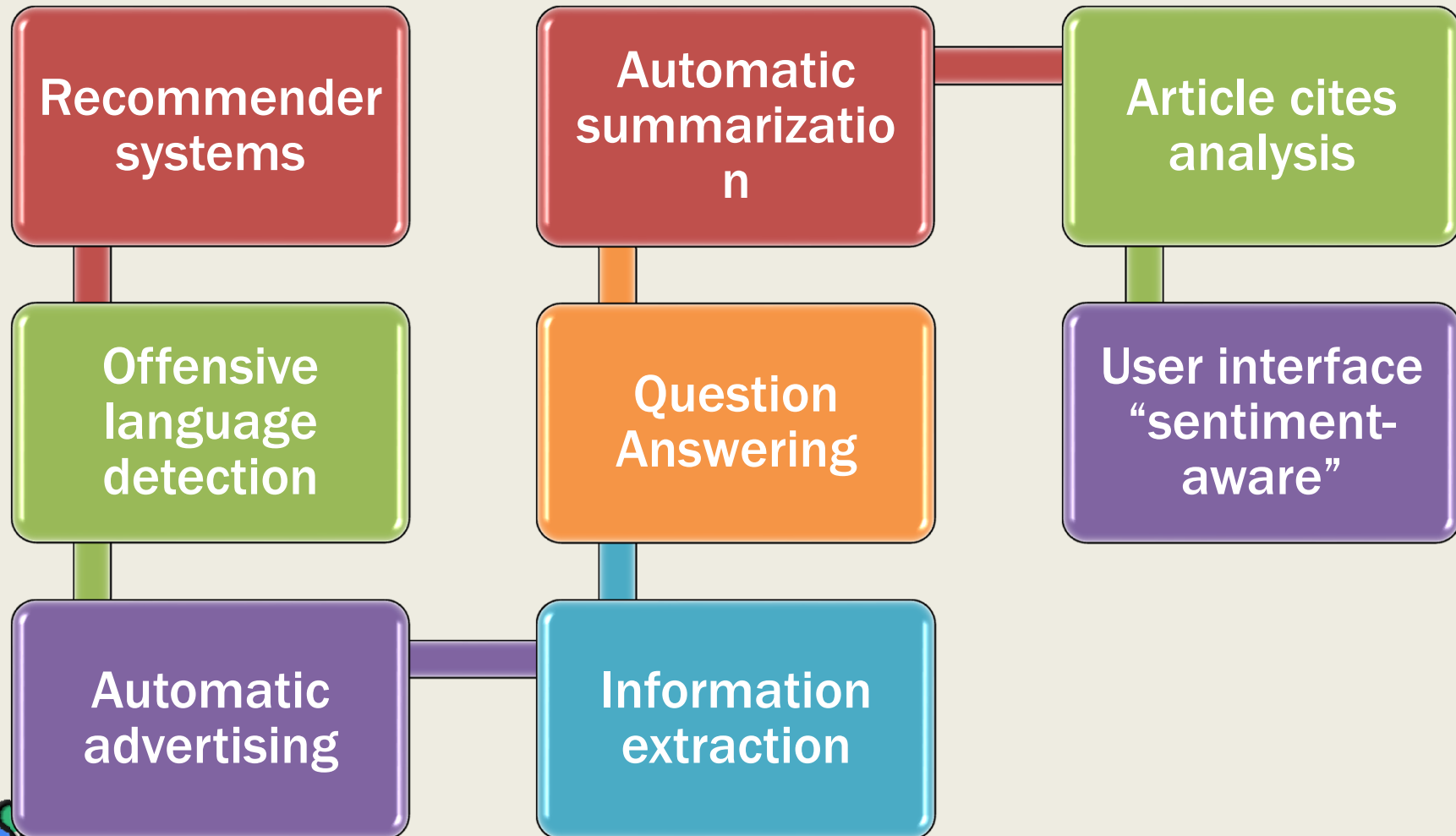
Generation of product comparatives

- Useful for consumers who hesitate between similar products

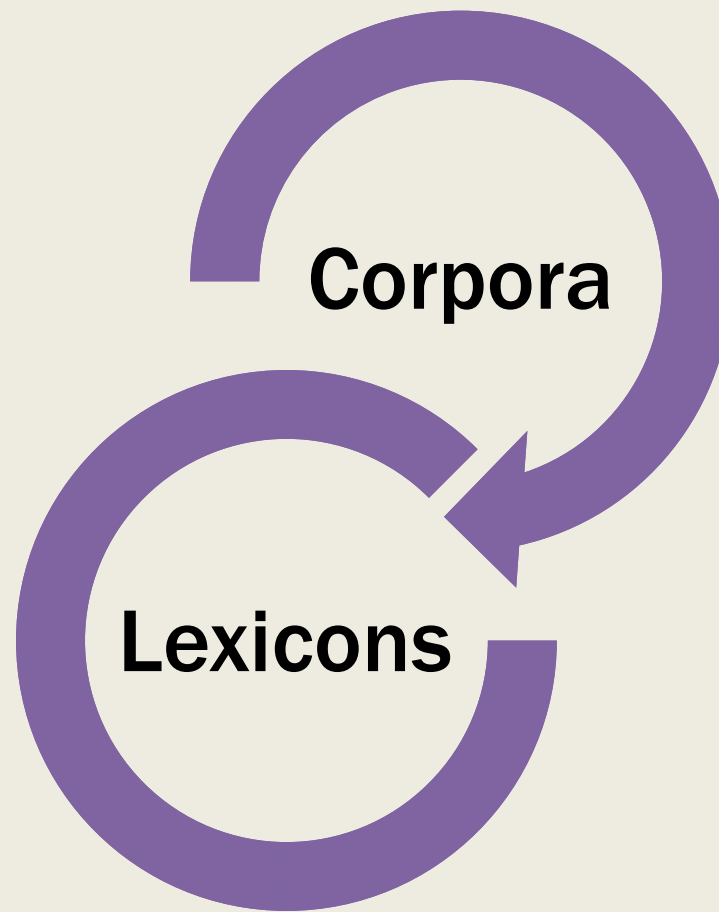
APPLICATIONS



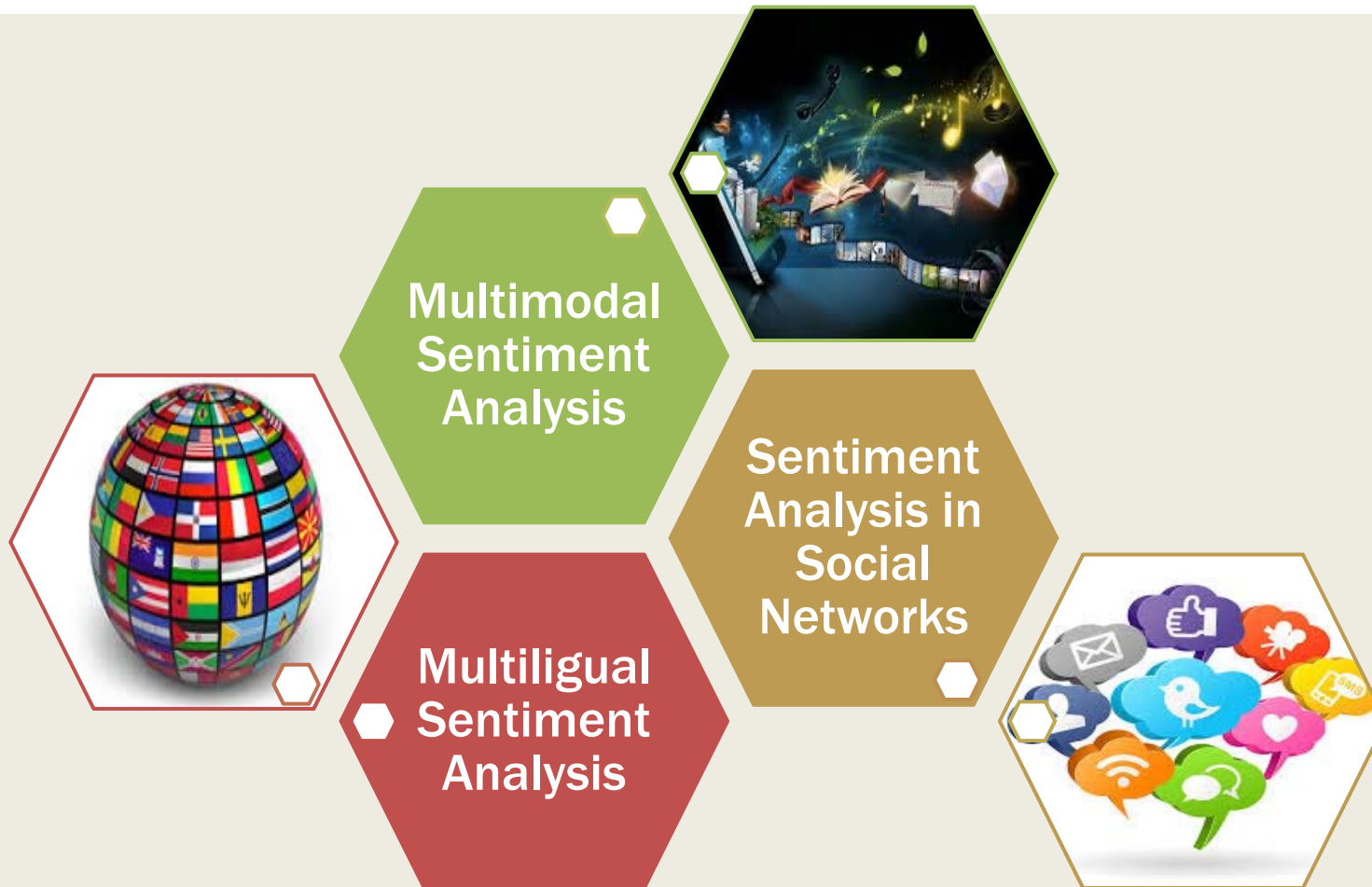
APPLICATIONS AS A SUBCOMPONENT IN A BIGGER SYSTEM



RESOURCES FOR OM



RESEARCH TRENDS



MORE INFORMATION

LINGUISTIC RESOURCES ON THE INTERNET

- ACL – The Association for Computational Linguistics
 - <http://www.aclweb.org/>
- ELDA – Evaluations and Language resources Distribution Agency
 - <http://www.elda.org/>
- ELRA – European Language Resources Association
 - <http://www.elra.info/>
- LDC – Linguistic Data Consortium
 - <http://www ldc.upenn.edu/>
- ELSNET (European Network in Language and Speech)
 - <http://www.elsnet.org/>
- Multext (Multilingual Text Tools and Corpora)
 - <http://www.lpl.univ-aix.fr/projects/multext/>
- CLR (Consortium for Lexical Research)
 - <http://crl.nmsu.edu/Tools/CLR/>
- List of linguistic resources from the Stanford university
 - <http://nlp.stanford.edu/links/linguistics.html>
- Instituto Cervantes OESI: Spanish site dedicated to linguistic technologies
 - <http://oesi.cervantes.es>
- Computational linguistic in Poland
 - <http://clip.ipipan.waw.pl/>



INTERNATIONAL CONFERENCES

- ACL (EACL, NAACL) - Annual Meeting of Association for Computational Linguistics
- COLING - International Conference on Computational Linguistics
- SEPLN - Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
- TSD - Text, Speech and Dialogue
- RANLP - Recent Advances in Natural Language Processing
- NLDB - Natural Language and Information Systems
- CLEF - Cross-Lingual Evaluation Forum
- PACAL RTE Challenge - Recognising Textual Entailment



JOURNALS

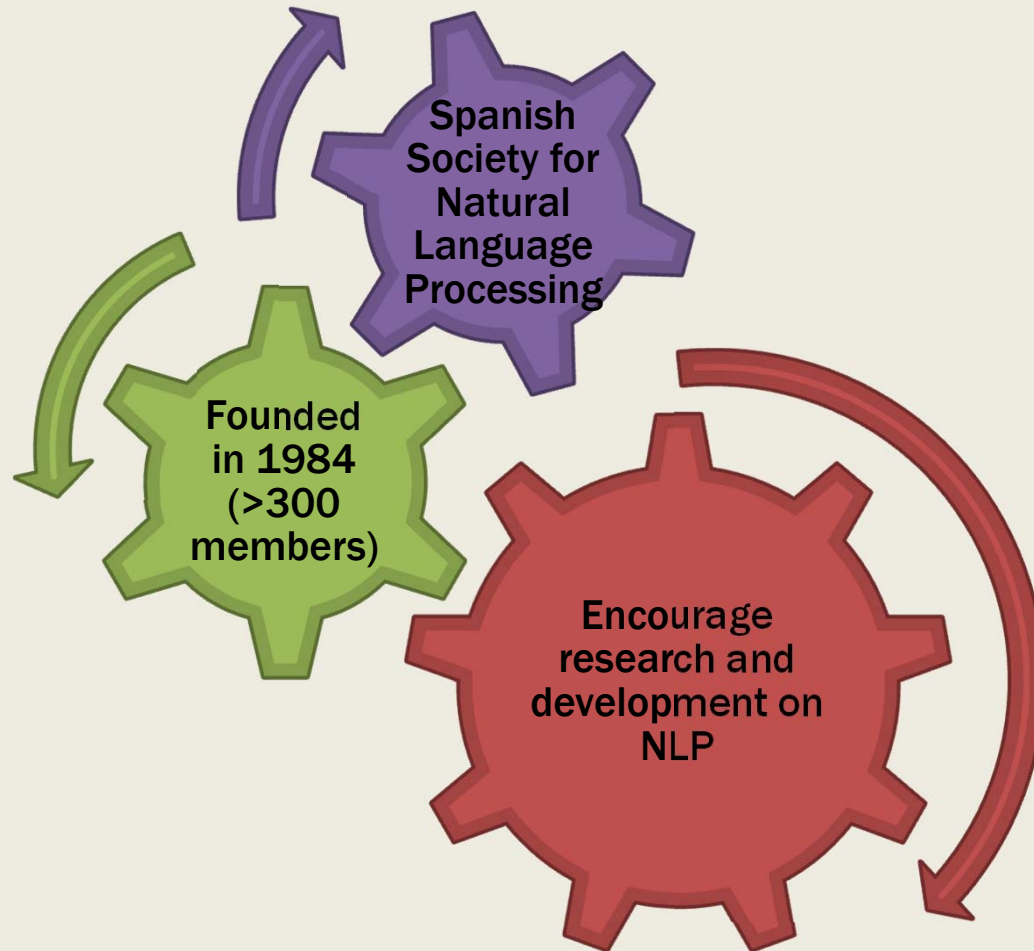


Journals

- Computational Linguistics
- Journal of Artificial Intelligence Research
- Artificial Intelligence
- Computing and Humanities
- ACM of Communications
- Journal of Intelligence Systems
- Information Retrieval Journal
- Machine Translation
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Procesamiento del Lenguaje Natural
- Revista Iberoamericana de Inteligencia Artificial
- Novática (Tecnologías del Lenguaje)

SEPLN: SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

[HTTP://WWW.SEPLN.ORG](http://www.sepln.org)



RED TEMÁTICA EN TRATAMIENTO DE INFORMACIÓN MULTILINGÜE Y MULTIMODAL HTTP://SINAI.UJAEN.ES/TIMM/



National
Thematic
Research
Network for
Multilingual
and Multimodal
Information
Management

Created
by SINAI
in 2006

>150
multidisciplinary
researchers &
institutions



MOOCTLH

UA L'EDUCACIÓ DIGITAL DEL FUTUR
LA EDUCACIÓN DIGITAL DEL FUTURO



moocTLH

g+1 22

Iniciar sesión

[Avisos](#) [Curso](#) [Foro](#) [Registro](#)

Nuevos retos en las Tecnologías del Lenguaje Humano

Este es un curso en línea abierto y masivo (massive open online course) centrado en las tecnologías del lenguaje humano (TLH). Te vamos a mostrar el estado actual de las tecnologías en la esperanza de que te aporte ideas para tu negocio. O igual te atrae la investigación y descubres que esto te interesa y querrías trabajar en ello.

El 7 de enero de 2014, empezamos...

[moocTLHinfo](#)
[moocTLHtube](#)
[moocTLHtwitter](#)
[moocTLHcomunidad](#)



Registrar



TIME FOR QUESTIONS



L. Alfonso Ureña López
laurena@ujaen.es