Introduction to Human Language Technologies and Relevant Applications for the Information Society

Elena Lloret Pastor
elloret@dlsi.ua.es
But before going into the topic ....

.... a little bit about what I come from
About Alicante …

- Where is **Alicante**?
About University of Alicante ...

- University of Alicante

Source of the pictures: www.ua.es
About University of Alicante …

- Polytechnic University College – EPSAlicante

- Computer Science Engineering
- Multimedia Engineering
- Robotics Engineering
- Health IT

- Fundamentals of architecture
- Technical architecture
- Civil Engineering

- Degrees

- Telecommunications (Sound and Image) Engineering
- Chemical Engineering

Source of the pictures: www.ua.es
About University of Alicante …

- Polytechnic University College - **EPSAlicante**

**ARA groups**

Reduced groups with at least 50% of the subjects in English

*Pre-requisites:* at least B1 for English and a good academic track record

Source of the pictures: www.ua.es
About me ...

- Full-time assistant lecturer at the Department of Software and Computing Systems

**Teaching**
- Databases (bachelor)
- Human Language Technologies (master)

**Research**
- Text Summarization
- Natural Language Generation
About GPLSI research group ...

Who are we?

- Research Group on Natural Language Processing and Information Systems (Dept. of Software and Computing Systems)
  - Since 1993
  - Team ~ 30 members
    - Senior researchers
    - Junior researchers
    - Ph.D. students
    - Technical staff
- More info
  - [http://gplsi.dlsi.ua.es/gplsi11/](http://gplsi.dlsi.ua.es/gplsi11/)
About GPLSI research group …

Research topics

- Information Retrieval
- Anaphora resolution
- Information Extraction
- Word Sense Disambiguation
- Sentiment Analysis
- Text Simplification
- Text Summarization
- …

Language Understanding

- Document planning
- Discourse planning
- Surface Realisation

Lang. Generation

Foro Internacional del Español
About GPLSI research group ...

Research topics

- Three pillars

Information Retrieval & extraction

Social Media Analysis

Text Summarization and generation

Czech Technical University in Prague
About GPLSI research group …

**GPLSI projects**

**Técnicas de deconstrucción aplicadas a las Tecnologías del Lenguaje Humano (MINECO, TIN2012-31224)**
- Develop a new paradigm for designing HLT processes
- Natural language deconstruction and reconstruction
- L-brick structure

**Socialising Around Media (CE, FP7-611312)**
- Create a social platform for distributing multimedia content
- Creation of dynamic social communities
- Standard multi-channel and multi-device

**DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0 (GVA, PROMETEOII/2014/001)**
- Create resources, tools and systems related to text mining and information generation
- Encourage knowledge transfer into business and society
About GPLSI research group …

GPLSI labs

GPLSI Social Rankings
• Trend analysis and monitoring in Social Media
  • http://gplsi.dlsi.ua.es/demos/socialrankings

GPLSI Election Maps
• Political analysis and monitoring in Social Media
  • http://gplsi.dlsi.ua.es/demos/electionmap/

GPLSI Alertas (www.ovtt.org)
• Technology monitoring and alert system
  • http://www.ovtt.org/informacion_alertas

GPLSI Observa (www.ovtt.org)
• Meta-searcher
  • http://www.ovtt.org/observa

GPLSI Compendium
• Automatic generation of text summaries
  • http://gplsi.dlsi.ua.es/demos/compendium/
Overview

- Human Language Technologies
- Current trends
- Open challenges
- Cutting edge research
- Conclusion and final remarks
Overview

- Human Language Technologies
  - Introduction
  - Research areas
- Current trends
- Open challenges
- Cutting edge research
- Conclusion and final remarks
**Human Language Technologies**  
*Introduction*

- **Language**
  - Key feature of human behaviour
  - The method of human communication, either spoken or written

- **Natural Language**
  - the method of human communication, either spoken or written, consisting of the use of words in a structured and conventional way
    - Written language \(\rightarrow\) process of knowledge transmission
    - Spoken language \(\rightarrow\) means of communication
Human Language Technologies

Introduction

- HLT \(\rightarrow\) covers a broad range of computational technologies with the goal of enabling people to communicate with machines using natural communication skills, keeping the language as the form of interaction, and information exchange (Cole, 1997)

- Different terminology
  - Human Language Technologies - HLT
  - Natural Language Processing - NLP
  - Computational Linguistics - CL
  - Natural Language Engineering - NLE

HLT is a multidisciplinary field
- expertise in areas of linguistics, psychology, engineering, and computer science

HLT goals
- To find computational mechanisms that allow to recognise, understand, and generate natural language
Human Language Technologies

Research Areas

- HLT involve the development of:
  - Methods, techniques and tools
  - Resources
  - Applications
Methods, techniques and tools

- Lexical analysis
  - Part of speech tagger
  - Lemmatizer
  - Tokenizer

Human Language Technologies
Research Areas

- quickly (Adverb)
  - quick (Adjective)
    - -ly (Suffix)
Human Language Technologies

Research Areas

- Methods, techniques and tools
  - Syntactic analysis
    - Dependency parsing
    - Sentence segmentation

```
Peter sold his car
```

Diagram:
- Sentence
  - Subject (noun phrase)
    - Head
      - Peter
  - Predicate (verb phrase)
    - Direct object (noun phrase)
      - his
      - car
Human Language Technologies

Research Areas

- Methods, techniques and tools
  - Semantic analysis
    - Word Sense Disambiguation
    - Coreference resolution
    - Named Entity recogniser

**Examples of Named Entity Recognition:**

- **Peter**
  - **NAMED ENTITY:** PERSON
- **went**
- **to**
- **a**
- **bank**
  - **FINANCIAL ENTITY**
- **in**
- **Paris**
  - **NAMED ENTITY:** PLACE

**9 different meanings (Longman Dictionary):**

- Bank = financial entity
- Bank = side of a river
- Bank = place where human blood etc. is stored
Human Language Technologies
Research Areas

- Resources
  - Corpora
  - Annotation schemes
  - Ontologies
  - Knowledge bases
  - Dictionaries

Diagram:
- Means of transport
  - Vehicle
    - Wheeled vehicle
      - Motor vehicle
        - car
      - bike

Human Language Technologies

Research Areas

- Applications
  - Information Retrieval
  - Information Extraction
  - Question Answering
  - Text Classification
  - Sentiment Analysis
  - Text Summarisation
  - Text Simplification
Overview

- Human Language Technologies
- Current trends
  - Web 2.0 Language
  - Web 2.0 Content
- Open challenges
- Cutting edge research
- Conclusion and final remarks
Current Trends
Web 2.0 Language

- Web 1.0
  - Information ➔ passive user
- Web 2.0
  - More interaction, new ways of communication, collaborative information ➔ active and proactive user
  - New textual genres
    - Social Networks
    - Blogs
    - Fora
    - Microblogs (i.e., Twitter)
Current Trends
Web 2.0 Language

Web 2.0

- New language of communication
  - Abbreviations → LOL = laugh out loud, ur = you are, tht = that, plz = please
  - Slang → aarping, floorrobe, fungry,…
  - Spelling errors → sucess = success, playin = playing
  - Substitute characters/words for numbers → 4 = for, 2 = to (e42s = easy for you to say)
  - Emoticons → 😊, 😞, …

LANGUAGE DIFFICULT TO UNDERSTAND
Recent Research

Text informality
- Detect the informality degree of a text (very formal ... very informal)

Text Normalisation
- \textit{WAAAAAAAY} up great! $\rightarrow$ Way up great!
- \textit{its} my last wish to see \textit{u plz} $\rightarrow$ Its my last wish to see you please.
Current Trends
Web 2.0 Content

- Web 2.0 → increase of subjective information
  - about products, services, places, …
  - … coming from a wide range of information sources
Current Trends
Web 2.0 Content

- ... and different formats
- Opinions, comments, reviews, ratings, etc.

| Food         | ★★★★★
| Service      | ★★★★★
| Price/Value  | ★★★★★
| Atmosphere   | ★★★★★
| Overall      | ★★★★★
| Was it a pleasant experience? | Yes
| How many were seated at your table? | 2
| Do they accept reservations? | Yes
| Would you return? | Yes
| Credit cards accepted? | Yes

---

**Apple IPhone**

<table>
<thead>
<tr>
<th>Mobile Phones - Apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>排名第1 - Apple 1</td>
</tr>
</tbody>
</table>

**Swotti Index**

- Sound Quality: ★★★★★
- Design: ★★★★★
- Display: ★★★★★
- Usability: ★★★★★
- Weight: ★★★★★
- Speed: ★★★★★
- Camera: ★★★★★
- Speakerphone: ★★★★★
- Size: ★★★★★
- Keypad: ★★★★★
- Reception Quality: ★★★★★

**RATINGS KEY**

- x5 - Excellent
- x4 - Above Avg.
- x3 - Average
- x2 - Below Avg.
- x1 - Poor

**Wishes Motel**

- #123 of 131 hotels in Miami
- ★★★★★
- 35 reviews

- “Average, terrible customer service” 03/29/2014
- “TERRIBLE Experience” 01/08/2014

---

Czech Technical University in Prague
Current Trends
Web 2.0 Content

- ... and different formats
  - Opinions, comments, reviews, ratings, etc.

“This hotel in Miami turned out to be one of the best hotels during our trip to the United States.” (Hampton Inn Hotel)

“MOST AWFUL EXPERIENCE EVER !!! The towels were dirty and bathroom was horrible. The pictures posted looked good but turned nightmare onsite. The room looked haunted.” (Wishes Motel)

“Wonderful meal, the food was exquisite! The atmosphere was amazing and timeless and we just loved everything about our dining experience” (The Hollywood Brown Derby Restaurant)

“I really don't like the touch screen on the iPhone.” (iPhone)
Subjective information
- Great value for individuals and companies
  - Help with the decision-making process
- Necessary to distinguish between objective and subjective information
  - Subjective information classification
    - Positive/Negative/Neutral

Recent Research
- Sentiment Analysis/Opinion Mining
  - GPLSI Social Rankings
Current Trends
Web 2.0 Content

GPLSI SOCIAL RANKINGS

CATEGORIES
POLÍTICA CANALES TV BANCA MÚSICALES TELEFONÍA FERIAS CURSOS DE INGLÉS COCHES

VISTAS
10 MINUTOS 1 HORA 1 DÍA 1 SEMANA 1 MES 1 AÑO

EVOLUCIÓN EN TIEMPO REAL - CANALES TV

RANKING EN TIEMPO REAL - CANALES TV - ORDENADOS DE MEJOR VALORADO A PEOR VALORADO

<table>
<thead>
<tr>
<th>ENTIDAD</th>
<th>EMISORES</th>
<th>RECEPTORES</th>
<th>REPUTACIÓN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 DISCOVERY</td>
<td>1626 lb.</td>
<td>998158 lb.</td>
<td>25.3</td>
</tr>
<tr>
<td>2 TELECINCO</td>
<td>8620 lb.</td>
<td>859559 lb.</td>
<td>23.6</td>
</tr>
<tr>
<td>3 ANTENA 3</td>
<td>6793 lb.</td>
<td>17057207 lb.</td>
<td>18.4</td>
</tr>
</tbody>
</table>

http://gplsi.dlsi.ua.es/de mos/socialrankings/
Current Trends
Web 2.0 Content

- Huge amount of available information
  - Opinionated or not
  - Impossible to process all this information in an effective and efficient manner
    - Determine the most relevant
  - Not understandable by everybody or too complex to be processed

- Recent Research
  - Text Summarisation
  - Text Simplification
Big Ben is the name of the bell in the clock tower of Westminster Palace. The clock was installed in 1854, but was not fully operational until 1859. Big Ben itself, the bell, was cast in 1856. The chimes are well known for its melody called "Westminster Quarters", but this is not only played by this clock tower but also by many others in London and Britain. The tower is 316 feet high. The name Big Ben applied originally only to the bell but was probably after Sir Benjamin Hall the first Commissioner of Works. Each year a lot of money is spent to keep the clock running. The clock tower is not open for public visitors. It is located at the Thames and can be seen from the Westminster Bridge.
Text Simplification

Wikipedia

Prague (ˈpraːɡə; Czech: Praha, [ˈpraɦa] (listen)) is the capital and largest city of the Czech Republic. It is the fifteenth-largest city in the European Union. It is also the historical capital of Bohemia. Situated in the north-west of the country on the Vltava River, the city is home to about 1.24 million people, while its larger urban zone is estimated to have a population of nearly 2 million. The city has a temperate climate, with warm summers and chilly winters. The origin of the name Praha is rather associated with the word prah (that means a 'threshold'), which is a rapid on the river. […]

Simple Wikipedia

Prague (Praha in Czech) is the capital and the biggest city of the Czech Republic. It has got 1,24 mil. people. Prague is one of the most beautiful European cities since the Middle Ages. Often called the "City of 100 towers", the "Rooftop of Europe" or the "Heart of Europe", Prague was a place where many merchants, artists and inventors met. Prague is full of historical monuments and shows all major artistic styles. The historical centre of Prague is situated on both banks of the Vltava river. […]

Czech Technical University in Prague
Overview

- Human Language Technologies
- Current trends
- Open challenges
  - Personalisation
  - Non-literal language detection
- Cutting edge research
- Conclusion and final remarks
Open Challenges

Personalisation

- Dynamically adapt the output of HLT systems according to user profile and preferences
  - Several criteria (need, purpose, language, topic, etc.)

Language (English)
Domain and topic (sports, ice hockey)
Purpose (concise information, breaking news)

First period over. Latvia v Czech Republic in #Prague 1-0, Norway vs. USA in #Ostrava 1-1 after 20 minutes. #IIHFWorlds
Open Challenges
Non-literal Language Detection

- In written or spoken language, figurative language is frequently used. This includes idioms or metaphors, the meaning of which is not literal

The 1930s was a difficult decade for Americans. Under the cloud of the Great Depression, the average American learned to do without comforts and luxuries. Many had to do without necessities, such as groceries and gasoline. People were lucky to have even low-paying jobs, when many of their neighbors were unemployed, impoverished, and frightened. […]

Rafael Nadal: Is the US Open champion better than ever?
"I am going to try to arrive in good shape to Wimbledon, and if not I am going to look at the rest of the season," he said, and so it proved as the backlog of matches caught up with him and a second successive early exit followed in London. […]

Czech Technical University in Prague
Open Challenges

Non-literal Language Detection

- Metaphor detection poses one of the great challenges in HLT
  - It requires deep language understanding
  - Distinction when an expression has literal or non-literal meaning

Literal meaning

Avoid *rolling up your sleeves* and pant-cuffs, because sparks or hot metal could deposit in the folds; also, wear your trousers outside your work boots.

Non-literal meaning

"We've got *to roll up our sleeves* to deal with this deficit and debt." (to get ready to start work)
Overview

- Human Language Technologies
- Current trends
- Open challenges

**Cutting edge research**
- Text Simplification
- Text Summarisation
- Natural Language Generation

- Conclusion and final remarks
A text can include diverse linguistic phenomena that can difficult its comprehension

- Complex sentences
- Coreference
- Non-literal language
- Rare words
- Specialised terminology

**Pulp Fiction**

*Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary.* The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. [...] The film's title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.

Cutting Edge Research

Text Simplification

Goal
- To facilitate reading comprehension
  - Adapt documents to make them easier to people with cognitive problems, poor literacy, non-native, ...

Original

Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary.

Simplified

Pulp Fiction is a 1994 American crime film. The film was directed by Quentin Tarantino. Quentin Tarantino co-wrote the screenplay of Pulp Fiction with Roger Avary.
Flexible Interactive Reading Support Tool (FIRST)

To use HLT to simplify documents in Bulgarian, English, and Spanish to people with Autism Spectrum Disorders

European project ➔ FP7 (http://www.first-asd.eu/)

Duration: 01/10/2011 – 30/09/2014

9 partners from 4 countries

3 Universities (UK, ES, ES)
3 clinical partners (UK, ES, BG)
2 companies (UK, BG)
1 NGO (BE)
Types of complexity within FIRST project

- **Structural complexity**
  - Simplification of coordinate and subordinate sentences

- **Semantic complexity**
  - Detection and resolution of coreferential entities
  - Detection and resolution of difficult words
  - Detection and resolution of non-literal language (i.e. idioms)

- **Document complexity**
  - Text Summarisation
  - Generation of tables of contents
Types of simplification

- Lexical simplificación
  - Identification and explanation of difficult terminology (e.g., specialised terminology, rare words)

**Pulp Fiction**

Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. […] The film's title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.
Types of simplification

- Syntactic simplification
  - Substitute a complex sentence by a set of simpler ones

Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary.

- Pulp Fiction is a 1994 American crime film.
- The film was directed by Quentin Tarantino.
- Quentin Tarantino co-wrote the screenplay of Pulp Fiction with Roger Avary.
Types of simplification
  Semantic simplification
  Identification and explanation of non-literal language

**Pulp Fiction**
Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. [...] The film's title refers to the **pulp magazines** and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.

**Pulp magazines** are inexpensive fiction magazines published from 1896 through the 1950s. The typical pulp magazine was 7 inches (18 cm) wide by 10 inches (25 cm) high, 0.5 inches (1.3 cm) thick, and 128 pages long. Pulps were printed on cheap paper with ragged, untrimmed edges.
Cutting Edge Research

Text Simplification

- Types of simplification
  - Semantic simplification
    - Identification and explanation of non-literal language

**Pulp Fiction**
Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. [...] The film's title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.

**Hardboiled** fiction is a literary genre that shares to some degree its characters and settings with crime fiction (especially detective stories).
Types of simplification

- Document simplification
  - Facilitate the pre-reading and understanding of the document
    - Table of contents
    - Summary

**Pulp Fiction**
Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. [...] The film’s title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.
Cutting Edge Research
Text Simplification

- Types of simplification
  - Document simplification
    - Facilitate the pre-reading and understanding of the document
      - Table of contents
      - Summary

**Pulp Fiction**
Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. [...] The film’s title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.
Cutting Edge Research
Text Simplification

- Types of simplification
  - Document simplification
    - Facilitate the pre-reading and understanding of the document

- Table of contents
- Summary

TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
<th>Concept N</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEYWORDS</td>
<td>KEYWORDS</td>
<td>KEYWORDS</td>
<td>KEYWORDS</td>
</tr>
</tbody>
</table>

Segment 1 ----> KEYWORDS
Segment 2 ----> KEYWORDS
Segment 3 ----> KEYWORDS
......
Segment N ----> KEYWORDS
Pulp Fiction is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, ironic mix of humour and violence, nonlinear storyline, and host of cinematic allusions and pop culture references. […]

The film's title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.
Types of simplification

- Document simplification
  - Facilitate the pre-reading and understanding of the document
    - Table of contents
    - **Summary (20% compression ratio)**

---

**Pulp Fiction**

*Pulp Fiction* is a 1994 American crime film directed by Quentin Tarantino, who co-wrote its screenplay with Roger Avary. The film was nominated for seven Oscars, including Best Picture; Tarantino and Avary won for Best Original Screenplay. It was also awarded the Palme d'Or at the 1994 Cannes Film Festival. It is neither neo-noir nor a parody of noir."
We’ve already learnt the “what”, but … how are the different types of simplification performed?

- A wide range of HLT techniques and approaches
  - Knowledge-based vs. Corpus-based
  - e.g. Text Summarisation
Summary? ➔ A condensed version of one or more texts containing the most important information (Mani, 2001)

Aim of Text Summarisation

- To generate a summary automatically
  - Avoid redundancy
  - Detect relevant information
  - Present it in a condensed form

A quantum computer is any device for computation that makes direct use of distinctively quantum mechanical phenomena, under the basic principle that the quantum properties of particles can be used to represent and structure data, and quantum mechanisms can be used to perform operations with this data.

Quantum computer
A quantum computer is any device for computation that makes direct use of distinctively quantum mechanical phenomena, such as superposition and entanglement, to perform operations on data. In a classical (or conventional) computer, information is stored as bits; in a quantum computer, it is stored as qubits (quantum bits).

The basic principle of quantum computation is that the quantum properties can be used to represent and structure data, and that quantum mechanisms can be devised and built to perform operations with this data. Although quantum computing is still in its infancy, experiments have been carried out in which quantum computational operations were executed on a very small number of qubits.

Research in both theoretical and practical areas continues at a frantic pace, and many national government and military funding agencies support quantum computing research to develop quantum computers for both civilian and national security purposes, such as cryptanalysis. If large-scale quantum computers can be built, they will be able to solve certain problems exponentially faster than any of our current classical computers (for example Shor's algorithm).

Quantum computers are different from other computers such as DNA computers and traditional computers based on transistors.

Some computing architectures such as optical computers may use classical superposition of electromagnetic waves, but without some specifically quantum mechanical resources such as entanglement, they have less potential for computational speed-up than quantum computers. The power of quantum computers, integer factorization is believed to be computationally infeasible with an ordinary computer for large integers that are the product of only a few prime numbers (e.g., products of two 300-digit primes).

By comparison, a quantum computer could solve this problem more efficiently than a classical computer using Shor's algorithm to find its factors.

This ability would allow a quantum computer to “break” many of the cryptographic systems in use today, in the sense that there would be a polynomial time (in the number of bits of the integer) algorithm for solving the problem. In particular, most of the popular public key ciphers are based on the difficulty of factoring integers, including forms of RSA. These are used to protect secure Internet pages, encrypted email, and many other types of data.

Breaking these would have significant ramifications for electronic privacy and security.

The only way to increase the security of an algorithm like RSA would be to increase the key size and hope that an adversary does not have the resources to build and use a powerful enough quantum computer. It seems plausible that it will always be possible to build classical computers that have more bits than the number of qubits in the largest quantum computer.
Big Ben is the name of the bell in the clock tower of Westminster Palace. The clock was installed in 1854, but was not fully operational until 1859. Big Ben itself, the bell, was cast in 1856. The chimes are well known for its melody called "Westminster Quarters", but this is not only played by this clock tower but also by many others in London and Britain. The tower is 316 feet high. The name Big Ben applied originally only to the bell but was probably after Sir Benjamin Hall the first Commissioner of Works. Each year a lot of money is spent to keep the clock running. The clock tower is not open for public visitors. It is located at the Thames and can be seen from the Westminster Bridge.
Summary types
Automatic summarisation process (Hovy, 2005)

- **Topic Identification**: What the document is about
- **Interpretation**: Understand the document
- **Summary Generation**: Produce the final summary using NLG techniques

Automatic summarisation process (*Hovy, 2005*)

- **Extractive summarisation**
  - **Topic Identification**
    - What the document is about
  - **Interpretation**
    - Understand the document
  - **Summary Generation**
    - Produce the final summary using NLG techniques

---

Cutting Edge Research
Text Summarisation

GPLSI COMPENDIUM

Alicante is a city in Spain, the capital of the province of Alicante and of the comarca of Alacantí, in the south of the Valencian Community. It is also a historic Mediterranean port. The population of the city of Alicante proper was 334,329, estimated as of 2011, ranking as the second-largest Valencian city. Including nearby municipalities, the Alicante conurbation had 462,281 residents. The population of the metropolitan area (including Elche and satellite towns) was 771,061 as of 2011 estimates, ranking as the eighth-largest metropolitan area of Spain.

Amongst the most notable features of the city are the Castle of Santa Bárbara, which sits high above the city, and the port of Alicante. The latter was the subject of bitter controversy in 2006–2007 as residents battled to keep it from being changed into an industrial estate.

The Santa Bárbara castle is situated on Mount Benacantil, overlooking the city. The tower (La Torre) at the top, is the oldest part of the castle, while part of the lowest zone and the walls were constructed later, on the 18th century.

The promenade Explanada de España, lined by palm trees, is paved with 6.5 million marble floor tiles creating a wavy form and is one of the most lovely promenades in Spain.

http://gplsi.dlsi.ua.es/demos/compendium/
Cutting Edge Research
Text Summarisation

- GPLSI COMPENDIUM CORE

[Diagram showing processes of surface linguistic analysis, redundancy detection, topic identification, relevance detection, summary generation, and extractive summary generation]
Cutting Edge Research
Text Summarisation

GPLSI COMPENDIUM (Lloret, 2011)

Legend:
- Core Stages
- Additional Stages
- Input
- Input for the additional stages
- Types of Summaries (output)

- Surface Linguistic Analysis
- Subjective Information Detection
- Redundancy Detection
- Relevance Detection
- Summary Generation
- Extractive Summary
- Information Compression and Fusion
- Abstractive Summary
- Topic Identification
- Query-Focused Summary
- Sentiment-Based Summary

Cutting Edge Research

Text Summarisation

- Surface Linguistic Analysis
  - Pre-processes the input text by employing HLT state-of-the-art tools
  - Sentence segmentation
  - Tokenisation
  - Part-of-Speech tagging
  - Stemming
  - Stop word identification
The main idea behind the use of TE for detecting redundancy is that those sentences whose meaning is already contained in other sentences can be discarded, since the information has been previously mentioned.

- The man was **killed** last week
- The man is **dead**
- The man was **shot** in his shoulder
- The man is **dead**

---

**Redundancy Detection**
- Identifies and removes redundant information
- Textual Entailment (Ferrández, 2007)

---

Óscar Ferrández, Daniel Micol, Rafael Muñoz, Manuel Palomar: DLSITE-1: Lexical Analysis for Solving Textual Entailment Recognition. NLDB 2007: 284-294

---

Czech Technical University in Prague
Cutting Edge Research

Text Summarisation

- **Topic Identification**
  - Identifies the most relevant topics
  - Term frequency (*Luhn, 1958*)

- Most frequent words (without considering stop words) can be considered representative of the main topics of a document

---

Relevance Detection

- Computes a weight for each sentence, depending on its importance
  - The Code Quantity Principle (Givón, 1990)
    - Coding element \( \rightarrow \) noun phrase

- Sentences containing a noun phrase including high frequent words will be considered more important

- Score for each sentence \( \rightarrow \) \( \text{relevance}_{\text{sent}} = \frac{1}{|NP|} \sum_{w \in NP} \text{tf}_w \)
Cutting Edge Research

Text Summarisation

- Summary Generation
  - Summary size
    - number of words
    - compression rate
  - The highest scored sentences up to a desired length are selected and extracted
  - Sentences are ordered as they appear in the document

- Type of summaries (output)
  - Generic extracts $\Rightarrow$ COMPENDIUM$_E$
Cutting Edge Research

Natural Language Generation

- Initial language comprehension stage
  - Information retrieval and extracción
  - Relevant fragment detection (only extraction)
  - Simplification, when necessary
- … but
  - Once the relevant fragments have been detected, how they can be fused/combined/joined to produce a new coherent text?
    - Natural Language Generation
      - Sentence and information ordering
      - Sentence realisation
Cutting Edge Research
Natural Language Generation

- Macroplanning
- Microplanning
- Realisation

Information representation

Text/voice in natural language
Cutting Edge Research
Natural Language Generation

- Macroplanning
- Microplanning
- Realisation

Information representation

NLG process

Text/voice in natural language

CLASE
Laptop, tecra A8

PROPIEDADES
manufacturedBy: toshiba
hasProcessor: intel Core 2
hasMemoryInGB: 2
hasHardDiskInGB: 110
hasSpeedInGHz: 2
hasPriceInEuro: 850

SALIDA
"Tecra A8 is a laptop, manufactured by Toshiba. It has an Intel Core 2 processor, 2 gb ram and a 110 gb hard disk. Its speed is 2 ghz and it costs 850 Euro."
Cutting Edge Research
Natural Language Generation

- Macroplanning
- Microplanning
- Realisation

(Reiter & Dale, 2000)

Cutting Edge Research
Natural Language Generation

WHAT TO SAY?

HOW TO SAY?

INPUT
COMMUNICATION GOAL
DISCOURSE HISTORY
USER MODEL
KNOWLEDGE BASES

OUTPUT

PREPROCESSING
MACROPLANNING
MICROPLANNING
REALISATION

WHAT TO SAY?

Document Plan
Content Selection
Content Structuring
Aggregation
Lexicalisation
Reference Expression Generation

HOW TO SAY?

Discourse Plan
Linguistic Realisation
Structure Realisation
Overview

- Human Language Technologies
- Current trends
- Open challenges
- Cutting edge research
- Conclusion and final remarks
Conclusion and final remarks

- Final goal of Human Language Technologies
  - To allow human-computer interaction using natural language

- HLT
  - Research Area on the rise
    - Horizon 2020 EU
  - Huge relevance for the society
    - A wide range of real applications
Conclusion and final remarks

- Existing challenges
  - Improve language comprehension
    - New textual genres, Web 2.0
    - Knowledge inference and deduction
  - Natural language generation
  - Information reliability & credibility
  - Complex linguistic phenomena detection
    - Non-literal language
    - Irony, sarcasm
Conclusion and final remarks

- Hot research HLT topics
  - Text Simplification
    - Facilitate the comprehension of texts
  - Text Summarisation
    - Provide key ideas while reducing the length of the document/s
  - Natural Language Generation
    - Produce human utterances automatically
Děkuji!!!!

Elena Lloret Pastor

Tel. +34 96 590 2448
Campus de Sant Vicent del Raspeig
03690 - Alacant
elloret@dlsi.ua.es
Introduction to Human Language Technologies and Relevant Applications for the Information Society

Elena Lloret Pastor
elloret@dlsi.ua.es