The University of Alicante at MultiLing 2015: approach, results and further insights

Marta Vicente, Óscar Alcón, Elena Lloret
MultiLing 2015

GPLSI

Universitat d'Alacant
Universidad de Alicante

Overview

- Motivation & Context
- Multilingual Single-document Summarization
  - UA-DLSI approach
- Experiments & Evaluation
  - Language choice & Datasets
  - Experimental Setup
  - Results & Analysis
- Conclusions & Next Steps

Motivation & Context

redpacientes.com

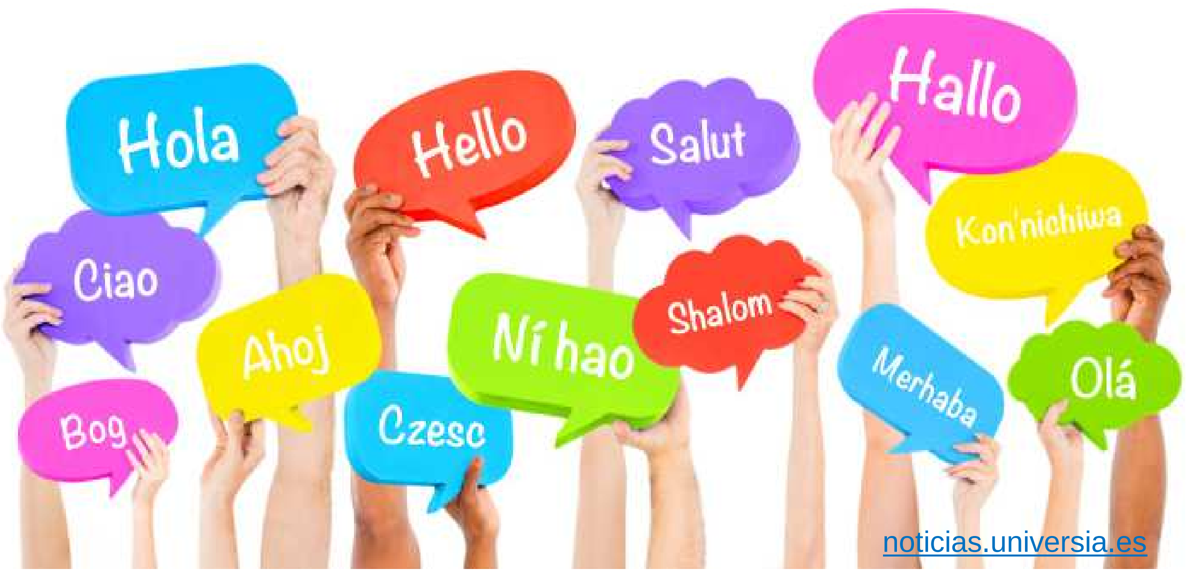
High volumes of information

- Difficult to manage
- What is the relevant information for us?

Motivation & Context

… in multiple languages

- Information lost if we cannot understand all the languages



Motivation & Context

Multilingual Summarization as a key technology

Determines the most relevant information

Deal with multiple languages

Multilingual Single-document Summarization (MSS)

To evaluate performance of systems generating a single document summary from Wikipedia articles in some of the languages provided (at least 3 languages from 38 available languages)

UA-DLSI Approach

- Technique employed
  - PCA: Principal Component Analysis
    - A statistical technique focused on the synthesis of information to compress and interpret the data.
    - Provides a way to determine the most relevant key terms of a document

UA-DLSI Approach

- Technique employed
  - PCA: Principal Component Analysis
    - A statistical technique focused on the synthesis of information to compress and interpret the data.
    - Provides a way to determine the most relevant key terms of a document
- Our **contribution**
  - Incorporation of lexical-semantic knowledge
    - Named Entity Recognition
    - WordNet + EuroWordNet

UA-DLSI Approach

- Summary Generation Process

Interpretation	• Creating the lexical-semantic matrix
Transformation	• PCA in action: determining key concepts
Summary Generation	• Sentence selection and ordering

UA-DLSI Approach

Interpretation

- Creating the lexical-semantic matrix

- Basic linguistic processing: sentence segmentation, tokenization, stopwords removal
- Identification of Named Entities and synonyms
- We group a set of synonyms under the same concept by the most frequent sense approach for each term.

UA-DLSI Approach

Interpretation

- Creating the lexical-semantic matrix

- Basic linguistic processing: sentence segmentation, tokenization, stopwords removal
- Identification of Named Entities and synonyms
- We group a set of synonyms under the same concept by the most frequent sense approach for each term.

Result: an initial lexical-semantic matrix, sentence as rows in the matrix, sense units (concepts, named entities, terms) as columns.

UA-DLSI Approach

Transformation

- PCA in action: determining key concepts

- Applying PCA technique we obtain the principal components (eigenvectors) and its corresponding weight (eigenvalue).
- The first eigenvectors collect the major part of the information extracted from the covariance matrix
  - Eigenvectors are derived in decreasing order of importance

UA-DLSI Approach

Transformation

- PCA in action: determining key concepts

- Applying PCA technique we obtain the principal components (eigenvectors) and its corresponding weight (eigenvalue).
- The first eigenvectors collect the major part of the information extracted from the covariance matrix
  - Eigenvectors are derived in decreasing order of importance

Result: relevant concepts are determined

UA-DLSI Approach

Summary Generation	• Sentence selection and ordering

- Two strategies are proposed for building different types of summaries
  - **Generic summaries** → for each relevant concept, select *<u>one sentence</u>* in which it appears
  - **Topic-focused summaries** → for each relevant concept, select *<u>all the sentences</u>* in which it appears

UA-DLSI Approach

- Two strategies are proposed for building different types of summaries
  - **Generic summaries** → for each relevant concept, select *one sentence* in which it appears
  - **Topic-focused summaries** → for each relevant concept, select *all the sentences* in which it appears

Result: the summary is obtained

Experiments & Evaluation

- Language choice

English	Adding lexical-semantic knowledge requires some resources available for these languages
German	Named Entity Recognition → Standford NER
Spanish	Semantic knowledge → WordNet + EuroWordnet

Experiments & Evaluation

- Datasets provided

Training

- 30 articles/language
- Human-generated summaries
- Character length/target summary

Test

- 30 articles/language
- **NO** human-generated summaries
- Character length/target summary

Experiments & Evaluation

- Datasets provided

Training

- 30 articles/language
- Human-generated summaries
- Character length/target summary

Test

- 30 articles/language
- **NO** human-generated summaries
- Character length/target summary

Concision & precision are required

Lang	Compr. ratio
DE	2.75%
EN	7.19%
ES	5.21%

Experiments & Evaluation

- Experimental Setup

Criteria → **System configurations**

Types of summaries

- **T1** Generic Summary
- **T3** Topic-focused Summary

Types of Knowledge included

- **LI** Language-independent approach
- **LEX** Using Lexical Knowledge (Name Entity Recognition)
- **SEM** Semantic Knowledge (WordNet,…)

Words in the PCA matrix construction

- **OWFH** Only words in Wikipedia headings
- **AW** All words in the document

UA-DLSI 1	T1	LI	AW	
UA-DLSI 2	T3	LEX	SEM	OWFH
UA-DLSI 3	T1	LI	OWFH	
UA-DLSI 4	T1	LEX	SEM	AW
UA-DLSI 5	T3	LI	OWFH	
UA-DLSI 6	T3	LEX	SEM	AW

Experiments & Evaluation

- Experimental Setup
  - Baselines
    - LEAD
      - System that selects the leading substring of the article's body having the same length as the human summary
    - ORACLES
      - Select sentences from the body text that cover the tokens in the human sentences using as few sentences as possible
  - MSS 2015 Participants (5 systems)
    - *"BGU-SCE" – "CCS" – "EXB" – "LCS-IESI" – "UA-DLSI"*

Experiments & Evaluation

- Results & Analysis

ROUGE 1, F-measure

	UA - DLSI 1	UA - DLSI 2	UA - DLSI 3	UA - DLSI 4	UA - DLSI 5	UA - DLSI 6	Lead	Oracles	Best performance
en	0.45605	0.42703	0.40551	**0.45627** (15/22)	0.42419	0.42727	0.42907	0.60983	BGU-SCE 5 0.49361
es	**0.48977** (8/13)	0.47141	0.46979	0.48454	0.47691	0.47193	0.46599	0.61691	CCS 4 0.52126
de	0.34110	0.33725	**0.36236** (7/13)	0.34317	0.33437	0.34553	0.32230	0.52759	CCS 4 0.38803

UA-DLSI 1	T1	LI	AW	
UA-DLSI 2	T3	LEX	SEM	OWFH
UA-DLSI 3	T1	LI	OWFH	
UA-DLSI 4	T1	LEX	SEM	AW
UA-DLSI 5	T3	LI	OWFH	
UA-DLSI 6	T3	LEX	SEM	AW

Experiments & Evaluation

Results & Analysis

	UA - DLSI 1	UA - DLSI 2	UA - DLSI 3	UA - DLSI 4	UA - DLSI 5	UA - DLSI 6	Lead	Oracles	Best performance
en	0.45605	0.42703	0.40551	**0.45627** (15/22)	0.42419	0.42727	0.42907	0.60983	BGU-SCE 5 0.49361
es	**0.48977** (8/13)	0.47141	0.46979	0.48454	0.47691	0.47193	0.46599	0.61691	CCS 4 0.52126
de	0.34110	0.33725	**0.36236** (7/13)	0.34317	0.33437	0.34553	0.32230	0.52759	CCS 4 0.38803

- When summarizing Wikipedia articles, generic summarization has been shown to be more appropriate.
- Compress ratio values for German are the highest (2.75% against 7.19% for English), which is reflected in the scores obtained for this language.
  - These high values are quite challenging for a summarization task. In this sense, we keep looking into new options to improve our implementation

Conclusions & Next Steps

- Potentials
  - 1st time we participate in a summarization competition
  - Promising results were obtained
  - PCA is a very good technique for language-independent summarization
  - Wikipedia title headings were meaningful enough to build the PCA matrix in our summarization process
- Limitations
  - Lexical and semantic knowledge is dependent on the performance of the existing tools and resources
  - Need for going beyond extractive summarization
  - Compression ratio of Wikipedia articles too high compared with other type of documents

Conclusions & Next Steps

- Potentials
  - 1st time we participate in a summarization competition
  - Promising results were obtained
  - PCA is a very good technique for language-independent summarization
  - Wikipedia title headings were meaningful enough to build the PCA matrix in our summarization process
- Limitations
  - Lexical and semantic knwoledge is dependent on the performance of the existing tools and resources
  - Need for going beyond extractive summarization
  - Compression ratio of Wikipedia articles too high compared with other type of documents

Future Work → analyze PCA with other types of knowledge in order to advance the generation of abstractive summarization

Thank you for your attention!

Universitat d'Alacant
Universidad de Alicante

GPLSI

Elena Lloret Pastor

Tel. +34 96 590 2448
Campus de Sant Vicent del Raspeig
03690 – Alacant
elloret@dlsi.ua.es