PATTEXSUM : A pattern-based text summarizer

Elena Baralis, Luca Cagliero, **Alessandro Fiori**, and Saima Jabeen

{elena.baralis,luca.cagliero,alessandro.fiori,saima.jabeen}@polito.it

Dipartimento di Automatica e Informatica

Politecnico di Torino

Introduction

- Multi-document summarization
  - succint document collection representation
  - selection of a subset of not redundant and highly informative sentences
- State-of-art approaches
  - clustering-based
  - graph-based
  - linear programming

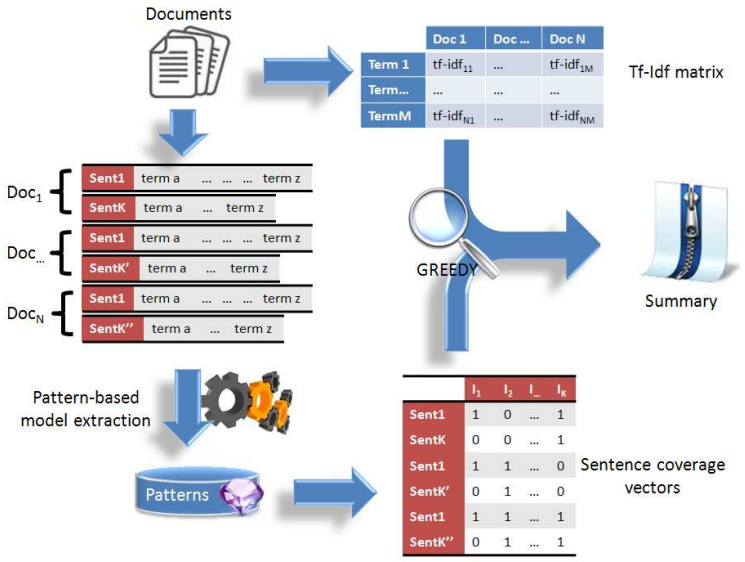
Innovative contribution

- The pattern-based model
  - transactional document representation to discover higher order correlations among document terms
  - extraction and selection of most valuable frequent itemsets
- Sentence evaluation
  - pattern-based model sentence coverage
  - tf-idf-based sentence relevance score
- Selection of a minimally not redundant subset of most relevant sentences

The method

Overview

- Two-way document collection representation
  - bag-of-word sentence representation
  - transactional data format
- Pattern-based model generation
- Sentence evaluation
  - sentence relevance score
  - sentence model coverage
- Sentence selection
  - set covering problem



Tf-idf matrix

- **Preprocessing**: Wordnet stemming algorithm
- Matrix composed of non-negative real value elements tc_{ik}
- Computed from the BOW representation of a sentence in the document collection D=$\{d_1, \ldots, d_n\}$

$$tc_{ik} = \frac{n_{ik}}{\sum_{r \in \{q\,:\,w_q \in d_k\}} n_{rk}} \cdot \log \frac{|D|}{|\{d_k \in D\ :\ w_i \in d_k\}|}$$

where:

- n_{ik}: number of occurrences of term w_i in document d_k
- $\sum_{r \in \{q\,:\,w_q \in d_k\}} n_{rk}$: sum of the number of occurrences of all terms in d_k
- $\log \frac{|D|}{|\{d_k \in D\,:\,w_i \in d_k\}|}$: inverse document frequency of term w_i

Transactional data representation

- Each sentence s_{jk} of the document collection D is represented as a transaction

$$tr_{jk} = \{w_1, \ldots, w_l\}$$

where:

- S_k=$\{s_{1k}, \ldots, s_{zk}\}$: set of sentences
- $tr_{jk} \subseteq s_{jk}$
- $w_q, w_r \in tr_{jk}$: sentence terms such that $w_q \neq w_r \; \forall \; q \neq r$

Document collection:

- $D \implies T$
- $T = \bigcup tr_{jk}$

Pattern-based model generation

Given:

- $\mathcal{D}(I)$: set of transactions covered by I
- $sup(I)=\frac{\mathcal{D}(\mathcal{I})}{|T|}$: support (observed frequency in D) of an itemset I

Model generation:

- Extraction of the most informative yet not redundant set of frequent itemsets
- Exploitation of the method presented in Mampaey et al.[a]
  - entropy-based heuristics for itemset evaluation
  - itemset mining and selection on-the-fly without postpruning
- Algorithm parameters:
  - min_sup: minimum support threshold
  - p: model size

[a] Mampaey, M., Tatti, N., & Vreeken, J. (2011). Tell me what I need to know: Succinctly summarizing data with itemsets. *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*

Sentence evaluation

Sentence relevance score Statistical sentence relevance evaluated on the document collection

$$SR(s_{jk}) = \frac{\sum_{i \mid w_i \in s_{jk}} tc_{ik}}{|t_{jk}|}$$

where:

- $|t_{jk}|$: number of distinct terms occurring in s_{jk}
- $\sum_{i \mid w_i \in s_{jk}} tc_{ik}$: sum of the tf-idf values associated with terms in s_{jk}

Sentence evaluation

Sentence relevance score Statistical sentence relevance evaluated on the document collection

$$SR(s_{jk}) = \frac{\sum_{i \mid w_i \in s_{jk}} tc_{ik}}{|t_{jk}|}$$

where:

- $|t_{jk}|$: number of distinct terms occurring in s_{jk}
- $\sum_{i \mid w_i \in s_{jk}} tc_{ik}$: sum of the tf-idf values associated with terms in s_{jk}

Sentence model coverage Pertinence of each sentence to the pattern-based model

- SC_{jk}=$\{sc_1, \ldots, sc_p\}$: sentence coverage vector
  - binary vector associated with each sentence $s_{jk} \in D$

$$sc_i = \mathbf{1}_{tr_{jk}}(I_i) \quad \text{where} \quad \mathbf{1}_{tr_{jk}}(I_i) = \begin{cases} 1 & \text{if } I_i \subseteq tr_{jk}, \\ 0 & \text{otherwise} \end{cases}$$

Sentence selection

- Set covering problem
  - selection of the minimum set of sentences
  - best coverage of the pattern-based model
  - maximize the sentence relevance score
- NP-hard problem
  - SC^*=$SC_1 \vee \ldots \vee SC_l$: summary coverage vector
  - greedy approach:
    - select the best non-overlapped sentence
    - best coverage of the model

Sentence selection

Greedy algorithm

Require: set of sentence relevance scores SR, set of sentence coverage vectors SC
Ensure: summary $\mathcal{S}$
{Initializations}
$\mathcal{S} = \emptyset$
$ESC = \emptyset$ {set of eligible sentence coverage vectors}
SC^* = all_zeros() {summary coverage vector with only 0s}

Sentence selection

Greedy algorithm

{Cycle until either SC^* contains only 1s or all the SC vectors contain only zeros}
while not (all_ones(SC^*) or only_zeros(SC^*)) **do**
    {Determine the sentences with the highest number of ones}
    ESC = max_ones_sentences()
    if ESC != $\emptyset$ **then**
        {Select the sentence with maximum relevance score}
        SC_{best} = $ESC[best]$ with $best$ = arg$_i$max SR_i
        {Update sets and summary_coverage_vector}
        $\mathcal{S} = \mathcal{S} \cup SC_{best}$
        SC^*= SC^* OR SC_{best}
        SC = $SC \setminus SC_{best}$
        {Update the sentence coverage vectors belonging to $\mathcal{V}$}
        for all SC_i in SC **do**
            $SC_i = SC_i$ AND $\overline{SC^*}$
        end for
    else
        break
    end if
end while
return $\mathcal{S}$

Experimental results

Experimental design

- Compared methods:
  - Open Text Summarizer (OTS)
  - TexLexAn
- ROUGE toolkit (version 1.5.5):
  - ROUGE-2, ROUGE-4
  - precision, recall, F-measure
- Experimental design: LOOCV

- Document collections:
  - real-life news articles
  - 10 news documents per category

Datasets	Description
Natural Disaster	Earthquake in Spain 2011
Royal Wedding	Prince William and Kate Middleton wedding
Technology	Microsoft purchased Skype
Education	Wealthy parents could buy their children places at elite universities
Sport	Australia defeat Pakistan in Azlan shah Hockey

Performance comparison

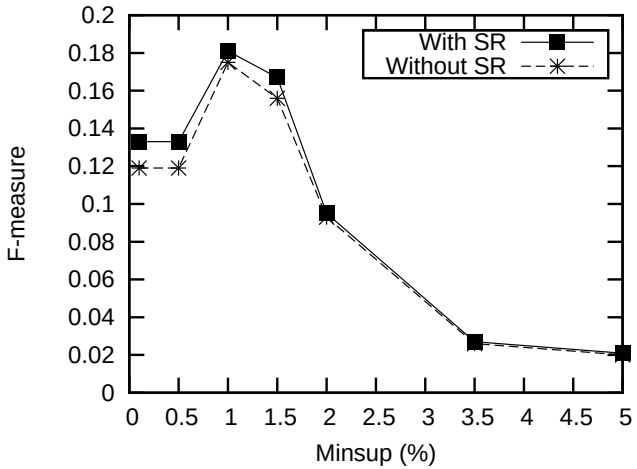
Performance comparison in terms of ROUGE-2 score

dataset	PATTEXSUM				OTS			TexLexAn		
	p	R	Pr	F	R	Pr	F	R	Pr	F
Natural Disaster	16	**0.116**	**0.288**	**0.141**	0.040	0.120	0.053	0.038	0.114	0.045
Royal Wedding	12	**0.036**	**0.215**	**0.058**	0.034	0.174	0.054	0.030	0.150	0.047
Technology	5	**0.141**	**0.465**	**0.210**	0.042	0.208	0.067	0.042	0.172	0.065
Sports	10	**0.145**	**0.297**	**0.189**	0.055	0.133	0.075	0.071	0.149	0.093
Education	8	**0.039**	**0.241**	**0.064**	0.036	0.170	0.054	0.034	0.150	0.051

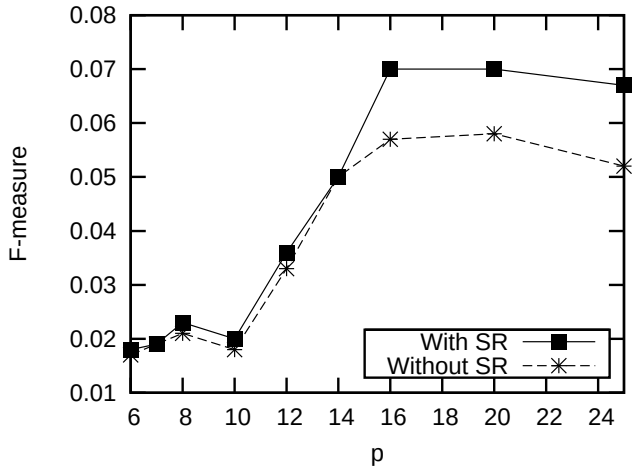
Performance comparison in terms of ROUGE-4 score

dataset	PATTEXSUM				OTS			TexLexAn		
	p	R	Pr	F	R	Pr	F	R	Pr	F
Natural-Disaster	16	**0.060**	**0.125**	**0.068**	0.005	0.012	0.006	0.005	0.011	0.006
Royal-wedding	12	**0.009**	**0.082**	**0.015**	0.003	0.018	0.005	0.003	0.018	0.005
Technology	5	**0.113**	**0.356**	**0.167**	0.009	0.065	0.016	0.003	0.011	0.005
Sports	10	**0.059**	**0.112**	**0.077**	0.004	0.010	0.006	0.022	0.036	0.027
Education	8	**0.017**	**0.141**	**0.030**	0.003	0.012	0.005	0.003	0.009	0.004

Parameter analysis



Technology. p=5. Impact of the support threshold.



Natural Disaster. min_sup=1.5%. Impact of the pattern-based model size.

Conclusions and future works

- Discovery of higher order correlations among terms
- Generation of pattern-based model with highly informative itemsets
- Combination of model coverage and statistical sentence relevance
- Selection of the minimal set of most relevant sentences
- Good performance achieved on real-life news articles
- Future works:
  - incremental summary updating
  - novel set covering approaches

Thanks for the attention!