STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

TEXT SUMMARIZATION BY FORMAL CONCEPT ANALYSIS APPROACH

DOINA TĂTAR, MIHAIELA LUPEA, AND ZSUZSANNA MARIAN

Abstract. This paper presents two original methods for text summarization (by extraction) of a single source document. The first method fulfills the two desiderata of summaries on the base of Formal Concept Analysis (FCA): 1.the relevance of a selected sentence is given by the introduction of the weight for a FCA concept, and 2. the minimal similarity to sentences previously selected is assured by a different coverage in the Concept Lattice. The second method realizes summarization by clustering the sentences. The new measure of similarity, *com*, has the roots also in FCA and provides the results close to the classical *cosine* measure.

At least to our knowledge, there are no previous attempts to solve the summarization of a text by FCA.

1. Introduction

Text summarization has become the subject of an intense research in the last years due to the explosion of the amount of textual information and it is still an emerging field [6], [8]. The extracts (which we are treating in this paper) are the summaries created by reusing portion of the input verbatim, while the abstracts are created by regenerating the extracted content [4]. However, research in the field has shown that most of the sentences (80%) used in an abstract are sentences which have been extracted from the text or which contain only minor modifications ([6]).

The paper is structured as follows: Some basic notions about FCA are given in Section 2. Our original method for summarization is described in Section 3. Section 4 presents the summarization by clustering with the *cosine* measure, on one hand, and the summarization by clustering with a new measure, inspired by FCA, on the other hand. We show here that these two

Received by the editors: March 14, 2011.

2000 *Mathematics Subject Classification.* 68T50, 03H65.

1998 *CR Categories and Descriptors.* I.2.7 [**Natural Language Processing**]: Discourse – *Coreference Resolution.*

Key words and phrases. text summarization, formal concept analysis.

measures are closely related. The last section contains conclusions and possible further work directions.

2. A short survey of Formal Concept Analysis (FCA)

FCA ([3]) has a big potential to be applied to a variety of linguistics domains as a method of knowledge representation, and provides a very suitable alternative to statistical methods. However, is somewhat surprising that FCA is not used more frequently in linguistic [7] (see [9] for another linguistic application). The reason could be that the notion of *"concept"* in FCA does not correspond exactly to the notion of *"concept"* as developed in Computational Linguistics.

Definition 1. *A* **formal context** $\mathbb{K} := (G, M, I)$ *consists of two sets* G *and* M *with* I *being a binary incidence relation between* G *and* M, $I \subseteq G \times M$. *The elements of* G *are called* **objects** *and the elements of* M *are called* **attributes** *of the context.*

The pair $(g, m) \in I$ is read as *"the object* g *has the attribute* m*"*.

The *derivative* of $A \subseteq G$ is $A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}$, the set of all attributes shared by the objects from A.

Dually, the *derivative* of $B \subseteq M$ is $B' = \{g \in G \mid \forall m \in B, (g, m) \in I\}$, the set of the objects which share all the attributes from B.

Definition 2. *A* **formal concept** *of the formal context* $\mathbb{K} = (G, M, I)$ *is a pair* (A, B), *with* $A \subseteq G$ *and* $B \subseteq M$ *such that:* $A' = B$ *and* $B' = A$. A *is called the* **extent** *and* B *is called the* **intent** *of the formal concept* (A, B).

Definition 3. *If* (A_1, B_1) *and* (A_2, B_2) *are concepts of a context* $\mathbb{K}$, (A_1, B_1) *is called* **subconcept** *of* (A_2, B_2) *provided that* $A_1 \subseteq A_2$ *(or equivalently,* $B_2 \subseteq B_1$*). In this case* (A_2, B_2) *is a* **superconcept** *of* (A_1, B_1) *and we write* $(A_1, B_1) \leq (A_2, B_2)$.

Definition 4. *For* $g \in G$, *the* **object concept** *is* $\gamma g := (g'', g')$ *and for* $m \in M$ *the* **attribute concept** *is* $\mu m := (m', m'')$. *The set of all formal concepts of a formal context together with the subconcept-superconcept order relation,* $\leq$, *forms o complete lattice called the* **concept lattice**.

3. Our approach: The basic idea of summarization by FCA

In the paper [5], the authors show that the exploiting of the diversity of topics in text has not received much attention in the summarization literature. However, they propose as *different topics* the different clusters (exactly as in older clustering methods), and for a reduced *redundancy*, a weighting scheme (for each sentence) which finds out the best scored sentences of each cluster. The authors of [2] assert that the main step in text summarization is the identification of the most important "concepts" which should be described in the

summary. By "concepts", they mean the named entities and the relationships between these named entities (a different vision from our FCA concepts).

In our method we use the FCA concepts and the idea that the quality of a summary is given by how many FCA concepts in the original text can be preserved in it with a minimal redundancy. The process of summarization is defined as extracting the minimal amount of text which covers a maximal number of "important" FCA concepts. The "importance" of a FCA concept is given by its generality in the concept lattice and by the number of the concepts "covered" by it. The most important sentences are selected to be introduced in the summary, keeping a trace of the concepts already "covered".

The basic idea is to associate with a text $T = \{S_1, \cdots, S_n\}$ a formal context (G, M, I) and a concept lattice CL ($\leq$ relation from Definition 4):

- the objects are the sentences of the text: $G = \{S_1, \cdots, S_n\}$;
- the attributes are represented by the set M of the *most frequent* terms (nouns and verbs) in T;
- the incidence relation I is given by the rule: $(S_i, t) \in I$ if the term t occurs in the sentence S_i.

Definition 5. *The weight $w(c)$ of a concept $c \in Conc$ is $w(c) = |\{m | c \leq \mu m\}|$, where $Conc$ is the set of all concepts (nodes) of the concept lattice CL.*

Definition 6. *An object concept S_i covers the concept c if $\gamma S_i \leq c$.*

The object concepts cover a bigger number of concepts if they are located in the lower part of the concept lattice CL. In other words, we are firstly interested in the sentences S_i such that γS_i are direct superconcepts of the bottom of the concept lattice CL. Let us denote this kind of sentences by $Sentence_{bottom}$. The algorithm introduces sequentially in the summary Sum the sentences from $Sentence_{bottom}$ which cover a maximal number of attribute concepts at the introduction time.

Summarization by FCA (SFCA algorithm):

Input: A text $T = \{S_1, \cdots, S_n\}$, the concept lattice CL, the set of concepts $Conc$, the set of concepts $Sentence_{bottom}$, the length L of the summary.

Output: A summary Sum of the text T with the length L.

Step 1. The set of covered concepts is empty, $CC = \emptyset$ and $Sum = \emptyset$.

Step 2. $\forall S_i \in Sentence_{bottom}$, $S_i \notin Sum$ calculate the weight: $w(S_i) = |\{m | \gamma S_i \leq \mu m, \mu m \in Conc \backslash CC\}|$.

Step 3. Choose in the summary the sentence with the maximum weight: $Sum = Sum \cup \{S_{i^*}\}$, $i^* = argmax_i\{\ w(S_i)\}$.

Step 4. Modify the set of covered concepts: $CC = CC \cup \{c | \gamma S_{i^*} \leq c\}$.

Step 5. Repeat from the **Step 2**, until the length of Sum becomes L.

Experiment

We tested our method on ten texts from DUC2002 documents. For the first text (Text1) having 15 sentences, the concept lattice is given in Figure 1. We chose as attributes the verbs and nouns with a frequency ≥ 3.

All the concepts from $Sentence_{bottom}$ are $\gamma S_1, \gamma S_2, \gamma S_3, \gamma S_6, \gamma S_7, \gamma S_8, \gamma S_9, \gamma S_{10}, \gamma S_{11}, \gamma S_{12}, \gamma S_{13}$, simply denoted by $1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13$ in the concept lattice. $L = 30\%$ length(Text1) $= 5 \leq |Sentence_{bottom}|$=11.

The first sentence introduced in the summary is S_8, since γS_8 covers the attribute concepts: $\mu puerto, \mu rico, \mu weather, \mu gilbert, \mu storm, \mu mph, \mu say$, the maximal number of covered concepts is $w(S_8)$=7. The algorithm provides the summary $Sum = \{S_1, S_3, S_7, S_8, S_{10}\}$ with a precision of 60% (Table 1) .

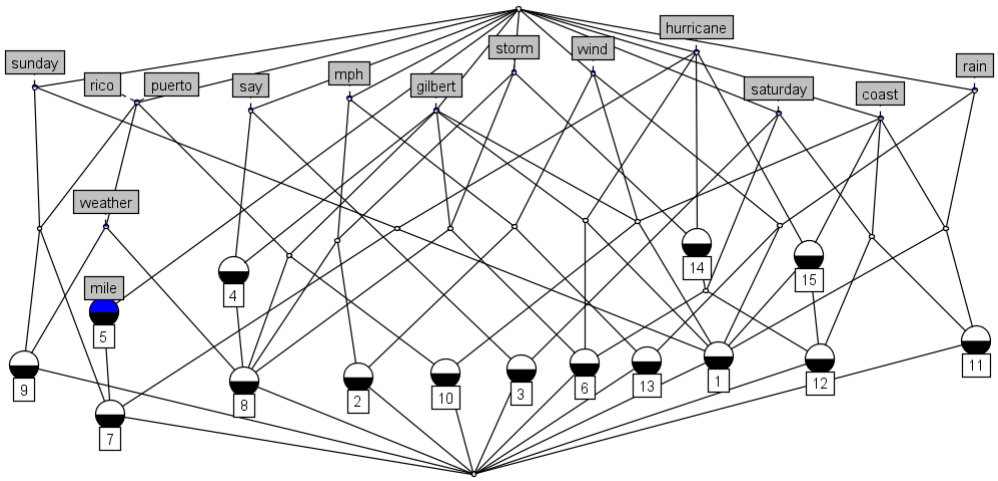

FIGURE 1. Concept Lattice for Text1

4. Summarization by Clustering

One of the first attempts to cluster sentences of a text was the paper [10]. In this section we show the results obtained by applying the *cosine* measure to cluster the sentences (as vectors) and to obtain further the summary.

Summarization by Sentences Clustering - SSC Algorithm

Input: A text $T = \{S_1, \cdots, S_n\}$, the length L of the summary.

Output: A summary Sum of the text T with the length L.

Steps: 1. calculate the frequency of the terms (verbs and nouns) in each sentence, 2. calculate the total frequency of the terms for all sentences, 3. choose the most frequent terms, 4. represent each sentence as vector using the frequent terms, 5. apply the hierarchical clustering algorithm for

	Text1 n=15	Text2 n=27	Text3 n=21	Text4 n=9	Text5 n=28	Text6 n=35	Text7 n=26	Text8 n=11	Text9 n=44	Text10 n=13
SFCA	60%	44%	43%	66%	55%	45%	66%	75%	73%	75%
SSC	40%	44%	43%	66%	22%	54%	33%	50%	46%	50%

TABLE 1. SFCA algorithm *versus* SSC algorithm - precisions with respect to manual summaries

$T = \{S_1, \cdots, S_n\}$ based on the similarity measure $sim(S_i, S_j)$, 6. build the summary (select from each cluster the sentence with the minimal index and re-traverse the clusters applying the same selection rule until L is reached).

Details regarding the implementation of SSC algorithm:

- The length of the summary is $L = 30\%n$, n is the length of the text.
- The number of clusters is equal with the length of the summary.
- The frequency for the m most frequent terms (nouns and verbs) used to represent the sentences as vectors is ≥ 2 or ≥ 3 such that $m \approx n$.
- In the bottom-up hierarchical clustering algorithm we begin with a separate cluster for each sentence and we continue by grouping the most similar clusters until we obtain a specific number of clusters (L). We have used:

  - as similarity measure between two sentences S_i and S_j:
    1) $sim(S_i, S_j) = cosine(V(i), V(j)) = \frac{\sum_{k=1}^{m} f(i,t_k) * f(j,t_k)}{\sqrt{\sum_{k=1}^{m} f^2(i,t_k) * \sum_{k=1}^{m} f^2(j,t_k)}}$ or
    2) a new measure denoted as *com* and defined as follows: $sim(S_i, S_j) = com(V(i), V(j)) = \frac{\sum_{k=1}^{m} min(f(i,t_k), f(j,t_k))}{\sum_{k=1}^{m} max(f(i,t_k), f(j,t_k))}$
  - as similarity between two clusters C1 and C2, for merging them:
    1) *single-link clustering*: the similarity of two most similar members $sim(C1, C2) = max\{sim(S_i, S_j) | S_i \in C1 \text{ and } S_j \in C2\}$;
    2) *complete-link clustering*: the similarity of two least similar members $sim(C1, C2) = min\{sim(S_i, S_j) | S_i \in C1 \text{ and } S_j \in C2\}$.

In the language of FCA, the new proposed similarity measure $sim(S_i, S_j) = com(V_i, V_j)$ represents the ratio of the number of the common attributes of objects S_i, S_j and the total number of attributes of these.

The SSC algorithm was implemented to work with all combinations for similarity of two sentences (*cosine* or *com*) and similarity between two clusters (*min* for complete-link clustering or *max* for single-link clustering).

According to the results obtained on the same input ten texts, the new measure *com* behaves like *cosine* measure with a precision greater than 80%.

Examples of summaries for **Text2**, 27 sentences, 25 frequent terms, $L = 9$:

- *cosine+min*: $\{S1, S2, S4, S5, S7, S11, S13, S15, S17\}$
- *com+min*: $\{S1, S2, S4, S7, S11, S13, S15, S17, S24\}$
- *cosine+max*: $\{S1, S4, S5, S7, S8, S11, S17, S23, S24\}$
- *com+max*: $\{S1, S2, S4, S8, S11, S12, S17, S23, S24\}$

From Table 1 we remark that both algorithms (SFCA and SSC) work with a good precision, but the precision is better when we apply SFCA-algorithm.

5. Conclusions and further work

The algorithms described in this paper are fully implemented and the evaluation indicates acceptable performance when compared against human judgment of summarization. However, we are currently looking at ways of expressing both properties of a good summary (the coverage and the distinctiveness) and the introduction of the number of occurrences of a term in a sentence using the multi-valued formal contexts ([3]).

References

[1] R. Barzilay, M. Elhadad, Using lexical chains for Text summarization, in J. Mani and M. Maybury editors, Advances in Automated Text Summarization, MIT Press, 1999, pp. 111-122.
[2] E. Filatova, V. Hatzivassiloglou, Event-based extractive summarization, Proceedings of the ACL-04, Barcelona, 21-26 July, 2004, pp. 104-111.
[3] B. Ganter, R. Wille, Formal Concept Analysis. Mathematical Foundations, Ed. Springer, 1999.
[4] E. Hovy, Text summarization, in R. Mitkov editor, The Oxford Handbook of Computational Linguistics, Oxford University Press, 2003, pp. 583-598.
[5] T. Nomoto, Y. Matsumoto, A new approach to unsupervised Text summarization, Proceedings of SIGIR 2001, September 9-12, 2001, New Orleans, pp. 26-34.
[6] C. Orasan, Comparative evaluation of modular automatic summarization systems using CAST, PhD Thesis, University of Wolverhampton, 2006.
[7] U. Priss, Linguistic application of Formal Concept Analysis, in Ganter, Stumme, Wille editors, Formal Concept Analysis, Foundations and Applications, Lecture Notes in Artificial Intelligence 3626, 2005, pp. 149-160.
[8] D. Radev, E. Hovy, K. McKeown, Introduction to the Special Issues on Summarization, Computational Linguistics 28, 2002, pp. 399-408.
[9] D. Tatar, E. Kapetanios, C. Sacarea, D.Tanase, Text Segments as Constrainted Formal Concepts, Proceedings of Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, 23-26 September 2010, IEEE Computer Society, pp. 223-228.
[10] Y. Yaari, Segmenting of expository text by hierarchical agglomerative clustering, Proceedings of Recent Advances in NLP, Tzigov Chark, Bulgarie, 1997, pp. 59-65.

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: {dtatar,lupea}@cs.ubbcluj.ro, marianzsu@yahoo.com