Identifying Phrases in Instructional Text Documents Using Instructional Objective and Goals.

Robinson Vida Noronha[1,2] **Clovis Torres Fernandes**[1]

[1]Technological Institute of Aeronautics - ITA, SP, Brazil

[2]Federal Center of Technological Education of Parana - CEFET, Pr, Brazil

{rvida@cefetpr.br; clovis@ita.br}

Abstract. *There are no consensual methodologies to aid teachers how to identify main facts or piece of domain elements inside instructional text documents. Teachers identify which facts are relevant to their instructional activities while reading document. This paper describes a helpful model and process to identify, select and order relevant facts and phrases inside an instructional text document. This process is guided by some pedagogical information such as "what is the Instructional Objective and its Goals" and "which are the relevant concepts or keywords of this domain". The main contribution of this report is how to use pedagogical information to select main facts and pieces of knowledge. The findings of a preliminary experiment shows that these model and process could be used to aid teachers to identify relevant facts and phrases inside text document.*

Key words: identifying relevant domain elements, scaffolding teacher, instructional goal and objectives.

1. Introduction

Teachers, in many situations of their school practice, must identify and select main facts or domain elements from instructional text documents. Some of these activities could be, among several others, the creation of Structural Communication exercises [Egan, 1976] and the selection of main concepts to be used like resources in activities of elaboration of concept maps.

There are no consensual methodologies to aid teachers to identify main facts or domain's elements from instructional text documents. While they are reading the text document, they take some notes or underline some phrases. They ask themselves why and how a specific phrase or word is important and deserves to be selected. While teachers do this task, they are guided by their own learning expertise. This is a

subjective and time-consuming task and the final result is influenced by the duration and tiredness of the teacher.

Analysis in the literature has pointed out a gap on how to identify relevant facts and phrases from text documents to be used in instructional activities and how they could be connected with instructional goals. The purpose of this research report is not to describe how to produce summary like the ones presented by Luhn [1999], Jones [1999], and Marcu [2000]. This paper describes a process on how to identify relevant facts and domain elements from an instructional text document. The difference between this process and others described in literature is that it uses information of teacher's expertise to guide the process. We name here this process as Extractor of Keywords and Phrases (EKP). The EKP has been implemented in Java language. The EKP process works together with DGG [Noronha, 2005] to automatically produce some sections of Structural Communication exercises.

The paper is organized as follows. Section 2 shows the target of the work and also it shows what has been done by others. Section 3 shows how instructional objective, instructional goals and keyword knowledge are structured. Section 4 describes the process of identifying and selecting phrases of document source. Section 5 describes the process of sorting selected phrases. Section 6 describes the EKP process. Finally, the last section presents some conclusions.

2. The Target and the Literature

The Automatic Text Summarization research area [Luhn, 1999; Jone, 1999; Marcu, 2000; Rath, 1999] give successful solutions to automate some tasks involved in a teacher process by identifying main facts or domain elements from instructional text documents, such as the following:

- To analyze the input text.
- To transform the text into a summary representation.
- To synthesize an appropriate output form.

The first two solutions could be adapted to aid teachers to identify the relevant domain elements, although there are few research papers reinforcing this statement. Synthesizing an appropriate output form is function of the context where the selected elements will be used.

In order to give such a context, unlikely the Automatic Text Summarization algorithms from the literature, EKP is guided by some pedagogical information like the following:

- “What is the Instructional Objective?”
- “What are the Instructional Goals?”
- “Which are the relevant concepts or keywords of the domain”.

All this information is provided and guided by the teacher when interacting with EKP.

This process could use some of the information text retrieval techniques such as stemming algorithms [Frakes, 1992], stoplists [Fox, 1992], thesaurus dictionary [Srinivasdan 1992], and boolean algorithms [Wartik, 1992]. Rath [1999] showed that computer selected only 64% of phrases selected by humans in scientific papers. Cañas et al, [2003] reported that specific algorithms based on relevant concepts were used to data mining. Their preliminary findings showed that 47% to 69% of relevant concepts stored in WWW were retrieved. The EKP is intended to reach or even to surpass the Rath's results, but this is not the aim of the work reported here. In fact, some of these techniques were experimented during this research work and the results obtained were poor. As a consequence, they were discarded from the EKP software.

EKP was devised to aid teachers on how to identify and sort facts and concepts from text documents based on instructional objective and goals. EKP is also intended to aid teachers to produce Structural Communication [Egan, 1976] exercises and will be part of an authoring framework to produce Structural Communication exercises based on ill-structured problems [Noronha et al., 2004]. The set of selected phrases is used to compose a Structural Communication's Response Matrix. This matrix stores relevant domain elements. Learners make up their solutions to ill-structured problems using some of these elements stored in a Response Matrix like a musician compose his music by selecting some of tones from a tonal scale.

To help explain the structure of the EKP process, a symbolic example is presented throughout the paper. This symbolic example was derived after a deep analysis of a real case, considering an instructional document, instructional objective and instructional goals is written in Portuguese and deals with drug problems, traffic and shantytown. The instructional document was obtained from Toledo (2004).

The EKP process has the following steps which will be detailed next:

- The teacher defines instructional objectives, instructional goals and knowledge keywords for a specific domain.
- EKP identifies relevant phrases from instructional texts.
- EKP sorts selected phrases.

3. Defining Instructional Objectives, Instructional Goals and Knowledge Keywords

The representation model defined during this research was a typical graph structure illustrated in Figure 1. In this graph, the root represents an Instructional Objective. The first level of graph represents the Instructional Goals. Each Instructional Goal (IG) has a set of knowledge keywords associated. The keywords represent pre-defined concepts or core aspects chosen by the teacher and that should be known by the learner. Here a knowledge keyword is depicted by KWK followed by a positive integer number.

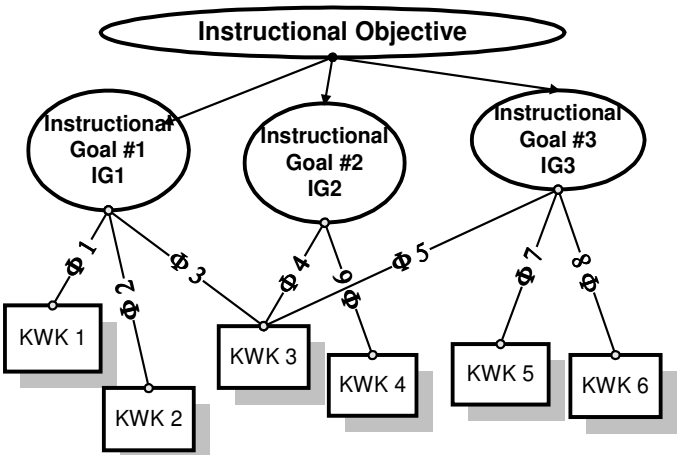

Figure 1 - Representation model for the instructional information of the domain.

The knowledge keywords can have different relevantly levels when they are connected to different instructional goals. The keyword "drug's legalization", for example, could be more relevant to an instructional goal like "What are the author's solutions to violence in shantytown?" than to an instructional goal like "Which are the aspects of problem of violence in shantytown?". The relevancy level is represented by factor **Φ** and is function of the relationship between an instructional goal (IG) and a keyword (IG-KWK).

Based on this representation model, at this step, the teacher has to define for a specific domain the instructional objective, the instructional goals, the knowledge keywords and the relevance level associated to all connections IG-KWK.

4. Identifying Phrases from Instructional Texts

This step of the process is automated. EKP identifies relevant phrases or facts from instructional text documents. These phrases aim to be relevant to instructional goals. This process uses a sequential search to locate, identify and select phrases that contain inside of them some knowledge keywords (KWKs).

Table 1 shows a possible example of the selection's result. In this table, the first line indicates that the first knowledge keyword (KWK1) is from some of document's phrases. F1, F3 and F7 numbered these phrases. One specific KWK can appear in one or more phrases and one selected phrase can contain one or more keywords. The two first lines of Table 1 show that sentence F1 contains the keywords numbered by KWK1 and KWK2.

The number of KWKs and its respective connection parameter Φ define the relevance level of a phrase with respect to a specific IG, as illustrated by Figure 1 and Table 1. The document phrases are classified as "Not Relevant" when they don't have any knowledge keyword associated. These phrases are discarded at the end of the process.

Figure 2 shows, for the current example, the relationship among some selected phrases and instructional goals. For instance, phrases F5, F8 and F9 and their respective

instructional goal, IG2. So, for example, the phrases F5, F8 e F9 are relevant to IG2 or have each one specific relevant level with respect to IG2.

Table 1 - Document Text's phrases that contain keywords knowledge.

Knowledge keyword	Selected Phrases
KWK 1	F1, F3, F7
KWK 2	F1, F2, F4
KWK 3	F5
KWK 4	F9, F8
KWK 5	F6, F7
KWK 6	F10, F3, F7

The list of sorted phrases is determined by the computation of the weight of the phrases. Table 2 shows the calculus of the weight for each phrase. Each phrase has a numeric value that corresponds to the sum of the relationship value IG-KWK (Φ_N) for each KWK found in it. For instance, KWK2(Φ2**)** stands for the relevance level of KWK2 with respect to IG1, namely Φ2. So, the weight of F5, for instance, is KWK3(Φ3) + KWK3(Φ4)+KWK3(Φ5). Considering Φ3=Φ4=Φ=1, the weight of F5 with respect to IG1, IG2 and IG3 is 3, which can be represented as "3*F5".

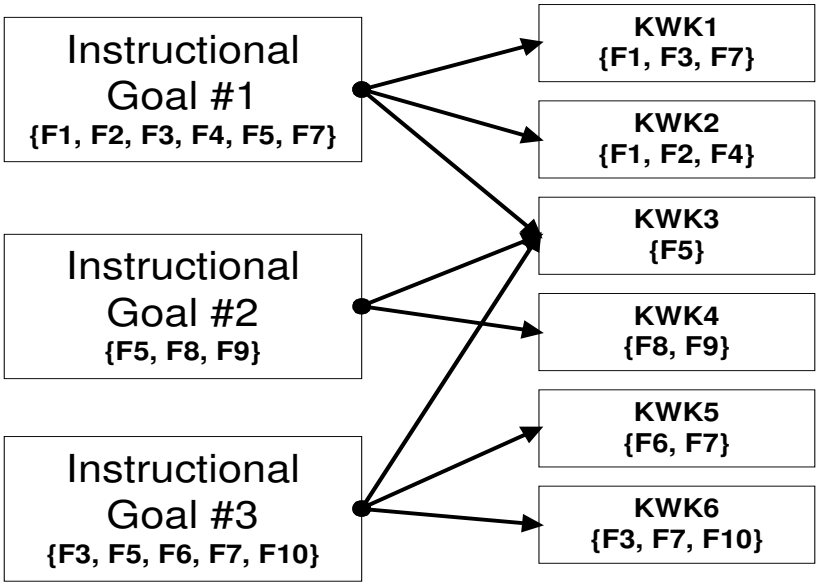

Figure 2- Relationship between selected phrases knowledge keywords and Instructional Goal.

Table 2 - Calculus of phrases' weight through keyword quantity and relationship IG-KWK.

Phrase	Weight
F1	KWK2(Φ2) + KWK1(Φ1)
F2	KWK2(Φ2)
F3	KWK1(Φ1) + KWK6(Φ8)
F4	KWK2(Φ2)
F5	KWK3(Φ3)+ KWK3(Φ4) + KWK3(Φ5)
F6	KWK5(Φ7)
F7	KWK1(Φ1) + KWK5(Φ7)+ KWK6(Φ8)
F8	KWK4(Φ6)
F9	KWK4(Φ6)
F10	KWK6(Φ8)

5. Sorting Selected Phrases

It was defined three rules for sorting out the selected phrases. The main ideas of this process are the following: i) which phrases are common to IGs's intersections; ii) which phrases are connected with each IG; iii) all IGs should have the same importance to reach the instructional objective, so that the amount of selected phrases associated with each IG should be almost always the same; as a result, the difference of number of phrases among IGs must be minimized as much as possible.

The relevant sequence of phrases is defined by the following sorting rules:

1) Instructional Goal Intersection. One phrase is more important to reach an instructional objective when this phrase is connected to a great number of instructional goals. Table 3 shows one example of the application of this rule. F5 is common to all defined IG. F5 connects with IG1 through Φ3, with IG2 through Φ4 and with IG3 through Φ5, as illustrated by Figure 1 and Figure 2. The first line of Table 3 shows the phrase F5 and its value of weight considering $\Phi_3 = \Phi_4 = \Phi_5 = 1$.

Table 3. Relevant phrases and IGs connections.

IG's Rule	Phrase's weight
$IG_1 \cap IG_2 \cap IG_3$	**3*F5**
$IG_1 \cap IG_2$	**2*F5**
$IG_1 \cap IG_3$	**2*F3+2*F5+3*F7**
$IG_2 \cap IG_3$	**2*F5**

Continuing with the application of this rule, one identifies which phrases are connected with a pair of IG. F3 and F7 are common to IG1 and IG3. The selected

phrases will be F5, F3 e F7. Following this way, the phrases connected with more than one IG will selected. The relevant sequence, until now, will be {F5, F7, and F3}.

2) Union of Instructional Goal. The set of non-sequenced phrases will be sorted in reason of their weight associate. These values are showed in Table 2. Considering $\Phi_N = 1$, for all connections, the result of employing this rule will be (2)* F1, (1)* F2, (1)* F4, (1)* F6, (1)* F8, (1)* F9, (1)* F10. They indicate that F1 is more relevant phrase of this subset because its weight is 2. The phrases with the same value of weight are considered similar. The relevant sequence, until now, will be {F5, F7, F3, F1}.

3) Instructional Goal Balance. Table 4 synthesizes the amount of selected and sorted phrases so far and their IGs. This balancing rule tries to reduce divergences introduced by the sorting process. The second column of Table 4, (Qf) indicates the sum of weight of sorted selected phrases (F5, F7, F3, F1) so far. F1 and F7 were selected because they have two KWK-IG links, so they are represented in Table 4 by (2*). IG2 has only one relevant sentence, the sentence F5. The IG1 has four relevant elements and its weight is equal to 5. IG3 has three relevant elements with weight equal to 4. This difference between IG1, IG2 and IG3 must be reduced. So, IG2 needs more selected phrases to sway the balance. Table 4 shows that IG1 has associated 5 phrases, IG3 4 phrases and IG2 has just one phrase. This difference should be minimized. Figure 2 shows the other phrases connected with IG2, namely F8 e F9. These phrases are inserted in the set of sorted and selected phrases. The relevant sequence, so far, will be {F5, F7, F3, F1, F8, and F9}.

   Following the step, it will be selected phrases F6 or F10 to reduce the difference of amount of selected phrases between IG3 and IG1.

Table 4. Selected phrases and IGs.

IG	Qf	Selected Phrases
IG1	5	F5, F3, F7, 2*F1
IG2	1	F5
IG3	4	F3, F5, 2*F7

The result of *Extractor of Keywords and Phrases* for the current example are the following:

- Relevant Selected Phrases Sorted by EKP: {F5, F7, F3, F1, F8, F9, (F6 or F10)}.
- Low Relevant Phrases: {F2, F4, (F6 or F10)}. This set corresponds to phrases selected but not sorted by EKP.

The final result is showed in Table 5. The Qf value should be the same for all IGs, but IG2 does not achieve it. To reduce this difference among Qf values, the pedagogical information could be modified to achieve this purpose, i.e. add new KWKs connected with IG2, or change the text document.

Table 5. Final selected phrases and IGs.

IG	Qf	Selected Phrases
IG1	5	F5, F3, F7, 2*F1
IG2	3	F5, F8, F9
IG3	5	F3, F5, 2*F7, {F6 or F10}

6. EKP Process

The main idea of this process is very simple and it is an adaptation of Luhn's algorithms [Luhn, 1999]. This process is illustrated in Figure 3. The process continuously breaks the text document until reaching single words. Then it follows by identifying which phrases have the pre-defined keywords and store this phrases. The phrases must be sorted by the rules. Instructional objective, IGs, IG-KWK relationship and the source of instructional text document are input data to this process.

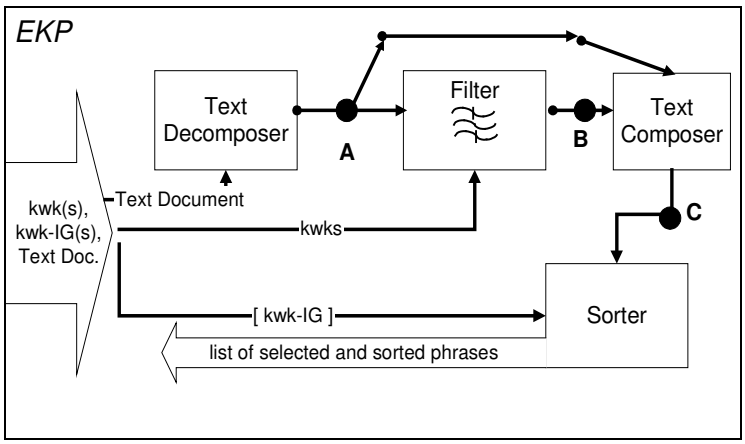

Figure 3. Block diagram for the EKP process.

The text document is decomposed and EKP produces a list of words and text punctuation, represented by point "A" in Figure 3. This list also records the quantity of occurrence of the word and its location. The number of paragraph, number of line and position of the word in the phrase represent the word's localization. The structure used was frames according to the following example:

```
Structure :
   [[label] [ [paragraph:sentence:position]
   ... [paragraph:sentence:position] ]]
Example:
[ [tráfico] [ [2:7:13] [4:10:40] ] ]
```

The Filter selects which phrases have occurrence of keywords, represented by point "B" in Figure 3. The Text Composer mounts the selected phrases, represented by

point “C” in Figure 3. The mounted phrases are then sorted by the three rules described in Section 4. The list with the sorted and selected phrases is the result of this process.

This process was briefly tested with a real teacher involved with instructional activities. The set of phrases selected and sorted out by the process was considered relevant by him.

This test also showed that information text retrieval techniques did not give a significant contribution to increment the number of selected phrases; neither it was able to identify more phrases relevant to the instructional objective.

The stemming algorithms proposed by Frakes [1992] aims to identify other somehow related words that could be also used as knowledge keywords. These words were named “auxiliary keywords”. In this case, these algorithms do not increase the result of the process. The Filter when using the three stemming algorithms proposed by Frakes produces result lists with few differences compared to the lists obtained with the normal Filter depict above. In fact, some other phrases were also selected. However, some of them were not related with the subject and instructional goals. For example, these algorithms selected the keyword “trafico” (traffic in English) the word “ficou” (past tense of verb “to stay”). The word “ficou” is very common in Portuguese and has a little discrimination features.

Table 5 exemplifies the auxiliary keywords obtained from keyword “tráfico”. Another feature observed is about Dice’s Coefficient [Frakes, 1992]. This coefficient is used to discriminate similar words. How to define its value does not have methodology to aid and, following the findings observed, it could not be used to all keywords in a practical example. When Dice’s coefficient assumed 0.3, the example listed in Table 5, the list of words contained some usefull words like “*tráfico*” (traffic) and “*traficantes*” (traffic dealers) and unusefull words like “*químicos*” (chemists) and “*tráfico*” (traffic).

Table 5 – Auxiliary keywords.

Stemmer Dice's Coefficient	Auxiliary keywords of keyword ***Tráfico*** **(traffic).**
0.3	***Químicos; públicos; encontra; contra; ficando; médicos; ficou; político; tráfico; ficar; traficantes***
0.4	***Ficou; tráfico; ficar***
0.5	***Ficou ; tráfico***
0.7	***Tráfico***

A thesaurus dictionary also was tested in an implementation of the process. The thesaurus dictionary was accessed by the Filter in an effort to find other words that could be used like keywords. This technique also proved to be useless. It is not so common that authors use synonymous words to represent main concepts or principles. Using thesaurus dictionary it produced the same list that was produced without dictionary. The process became just a time-consuming one.

The stemming algorithms and thesaurus dictionary were premature removed from the Filter module. These two techniques increased just the processing time and

their results were useless. More experiments must be carried on with real teachers and diverse knowledge domains, in order to validate the effectivity of the Filter in conjunction with the thesaurus dictionary when identifying relevant phrases.

8. Conclusions, Limitations and Future Works

How to identify relevant facts from instructional text documents is still an open question. This work provides a simple process to identify and select relevant phrases from text documents. This process is intended to aid instructional authors to create Structural Communication exercises.

This paper showed a process to aid teachers to carry out this difficult task. The process and model defined and reported in this paper could be extended to encompass some open research questions, such as how to identify "good" instructional materials to be used during instructional activities.

However, the EKP process and model have yet some limitations: the knowledge keywords just represent concepts, but some teachers might want to represent semantic ideas; the sorting process has as an outcome a list with some phrases and each phrase has a numeric value and all the process can not be defined or modified by the user; the teacher just can inform which are the instructional objective, instructional goals, knowledge keywords and instructional text document; finally, the criteria of the sorting process are fixed and could only be modified by changing the computer program.

We can envision the following future work for this research:

- To define some adaptability features for this process.
- To conduct more experiments with EKP trying to identify how many of the selected phrases are really used and which of them are not used and which of them must be edited by the author.
- To experiment the EKP with many different kinds of text document that could be used in instructional activities and considering many distinct instructional objective and instructional goals.

Our final target is to develop Structural Communication exercises, but during this research one question emerged: Could this process be used to select relevant instructional material?

Nowadays, the Internet have many instructional material scattered around all over the world. Could this process and model be used within an e-learning environment to identify and select the most relevant instructional material to the student?

Feigenbaum [2003] outline two grand challenges in achieving Computational Intelligence:

- To build a large knowledge base by reading text, reducing knowledge engineering effort by one order of magnitude.
- To distill from the WWW a huge knowledge base and build a system of "semantic scrappers".

We believe that identifying relevant instructional piece of information inside text documents could be considered a tiny step toward solving Feigenbaum challenges.

References

Egan, K., Structural Communication. Fearon Publishers, (1976).

Cañas, A., J., Carvalho, M., Arquedas, M., “Mining the web to suggest concepts during concept mapping: preliminary results”. In proceeding of XIII Simpósio Brasileiro de Informática em Educação, November, 2002, Unisinos, Brazil.

Feigenbaum, E., A., “Some challenges and grand challenges for computational intelligence”. Journal of the ACM, 50, (1), 2003, pp. 32-40.

Fox, C., "Lexical Analysis and Stoplists", Information Retrieval: Data Structures and Algorithms, William B. Frakes, Ricardo Baeza-Yates (eds), Prentice Hall PTR, 1992, pp. 102-130.

Frakes, W., B., "Stemming Algorithms", Information Retrieval: Data Structures and Algorithms, William B. Frakes, Ricardo Baeza-Yates (eds), Prentice Hall PTR, 1992, pp. 131-160.

Jones, K. S., "Automatic Summarizing: Factors and Directions", Advances in Automatic Text Summarization, Inderjeet Mani and Mark T. Maybury (eds), The MIT Press, 1999, pp. 1-14.

Luhn, H., P., "The Automatic Creation of Literature Abstracts", Advances in Automatic Text Summarization, Inderjeet Mani and Mark T. Maybury (eds), The MIT Press, 1999, pp. 15-22.

Marcu, D., The Theory and Practice of Discourse Parsing and Summarization, A Bradford Book, The MIT Press, 2000.

Rath, G., J., Resnick, A., Savage, T. R., "The Formation of Abstracts by Selection of Phrases", Advances in Automatic Text Summarization, Inderjeet Mani and Mark T. Maybury (eds), The MIT Press, 1999, pp.287-292.

Srinivasdan, P., "Thesaurus Algorithms", Information Retrieval: Data Structures and Algorithms, William B. Frakes, Ricardo Baeza-Yates (eds), Prentice Hall PTR, 1992, pp. 161-218.

Noronha, R., V., Villalba, C., Romiszowski, A., Fernandes, C., T., “An Authoring Tool for the Design of Structural Communication Exercises in WEB-based Environments”, 11th International Congress of Distance Learning, September 07-10, 2004, Salvador –BA, Brazil.

Noronha, R., V., “Authoring Ideas for Developing Structural Communication Exercises”. YRT Paper. Proceedings of 12th International Conference on Artificial Intelligence in Education - AIED 2005. July 18 to 22, 2005. Amsterdam, Netherlands.

Toledo, R., P., “A cheirada, a tragada e o sangue”, Veja Magazine, edition 1851 publishing date April, 28, 2004. Available on-line at http://veja.abril.com.br/280404/pompeu.html, 2004.

Wartik, S., "Boolean Operations", Information Retrieval: Data Structures and Algorithms, William B. Frakes, Ricardo Baeza-Yates (eds), Prentice Hall PTR, 1992, pp. 264-292.