Finding Lexical Relations for the Reuse of Investigation Reports

Luc Lamontagne[1], Rim Bentebibel[2], Erwan Miry[1], Sylvie Despres[2]
[1] Department of Computer Science and Software Engineering,
Université Laval, Québec, Canada, G1K 7P4
[2] UFR Mathématiques Informatique – CRIP5 – Equipe IAD
Université René Descartes, 45 rue des Saints-Pères 75006 Paris France
{Luc.Lamontagne, Erwan.Miry}@ift.ulaval.ca
{rim.bentebibel, sd}@math-info.univ-paris5.fr

Abstract. In this paper, we propose a reuse approach for investigation reports. Air investigation reports are documents containing observations, findings and recommendations made by safety analysts about incidents involving an aircraft. Reusing these complex documents is a considerable task as substantial domain knowledge is required to exploit their content. We conducted experiments using statistical techniques to acquire salient lexical relations from a corpus of reports. Keyphrase extraction and cooccurrence analysis was performed to condensate the textual content of the reports. Text alignment was applied to find associations between the sections targeted as case problems and case solutions. We illustrate lexical relationships we obtained and we discuss how these can be exploited to reuse investigation reports.

Keywords: Keyphrase Extraction, Cooccurrences, Statistical Alignment.

1 Introduction

In this paper, we describe some work we conducted for reusing investigation reports as produced by the Canadian Transportation Safety Board (TSB). Air Investigation reports are long and complex documents containing observations, findings and action recommendations made by safety investigators assigned to the analysis of transportation incidents. Such reports are important as they can be used for legal purposes in case of prosecution by victim relatives. They also are a means to disseminate information within the air service community as they contain detailed descriptions of human, environmental and equipment factors that might have impacted air incidents.

Our goal is to devise a scheme for exploiting antecedent documents during the authoring of new reports. Case-based reasoning (CBR) techniques would contribute to this task by selecting previous documents relevant to a new incident (the retrieval phase) and by proposing sentences from specific sections of these documents to support the air investigators while editing a new report (the reuse phase). As our interest is to go beyond retrieval in our exploitation of CBR techniques, we will concentrate on the latter problem which constitutes a reduced form of case adaptation

[1]. Such a reuse task requires textual CBR techniques as the content of the documents is unstructured and as we want to preserve the syntactic formulation of the sentences to be reused from previous investigation reports.

Reuse of individual sentences requires knowledge models to assess the utility of the sentences. Such models could be acquired through a knowledge engineering effort performed manually using domain-specific documents and interview transcripts. But we deem this approach not feasible in the current context and we estimate that an approach relying on manual domain modeling would lack generality.

We propose to address this task from a statistical natural language processing (NLP) point of view. More specifically, our main effort is to determine if significant relationships among individual words and portions of text can be identified. If so, then we could exploit these relationships to reduce the complexity of the documents and to relate previous recommendations to new incident descriptions. We first describe the corpus that was used for our experiment and the task we want to perform. Then we explain the techniques that were applied to obtain our statistical models. We illustrate some of the models we acquired through our experimentation and discuss how these can be exploited for reuse purposes.

2 Characteristics of the Corpus

Our analysis is based on 162 air investigation reports published between the years 2002 and 2005[1]. Each document is divided into sections by delimiters such as *Summary*, *Analysis* and *Safety actions*. Not all of the delimiters are present in all of the reports. The documents are long ranging from 1500 to 5000 words. To exploit the reports, we make the assumption that the problem sections are those depicting the nature of the air incident, i.e. *Summary*, *Synopsis*, *Other factual information* and *Analysis*. Also we assume that solutions are sections reporting on conclusions and recommendations made by the investigators. This includes sections such as *Findings as to causes*, *Findings as to risk*, *Other findings* and *Safety actions*.

Here are some characteristics we noted from the reports that had an impact on the techniques we selected:

- Some documents contain detailed descriptions of incidents that are difficult to understand for non-domain experts. For instance, the *Analysis* section often represent more than half of its corresponding report and provide a thorough analysis of chronological events. Capturing such an in-depth description is beyond the current state of the art of Textual CBR.
- The texts refer to individuals, locations, date, license numbers and other factual information. These are usually non recurrent (ex. a license number designates a single vehicle) and could be more useful if replaced by higher-level categories.
- The description of incidents often pertains to equipment problems, unfavorable environmental conditions (ex. weather, visibility, land barriers) and human errors occurring during flight. While fully automatic classification of these can require extensive training, capturing some of this information would be beneficial.

[1] http://www.tsb.gc.ca/en/reports/

3 Our Approach

The task to be accomplished by a CBR report authoring tool can be described as follows: Given a preliminary description of an new incident (sections such as *Summary*, *Analysis*…) provided by the investigator and an antecedent report to be reused, the CBR component should recommend to the investigator a choice of sentences potentially useful to the completion of the *Findings* sections.

Our approach is illustrated in Figure 1. First, we use *Summary* descriptions as case problems and sections about findings (*Findings as to causes, Findings as to risks, Other findings*) as case solutions. A basic preprocessing of these sections is performed as described in section 3.1. Then to cope with the complexity of a report, we reduce its content to a list of prominent keywords as described in section 3.2. To identify useful sentences, we build knowledge models consisting of associations between words of different sections. Techniques used to acquire these models are discussed in section 3.3. We finally discuss sentence reuse in section 3.4.

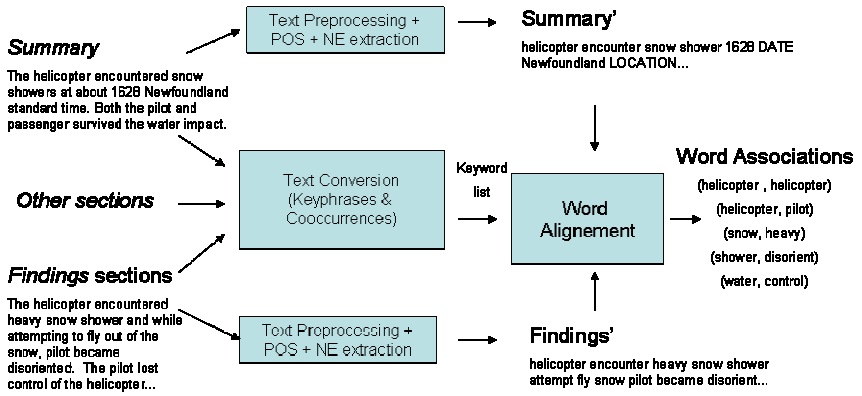

Fig. 1. Approach to acquire lexical knowledge required for sentence reuse

3.1 Preprocessing of the Investigation Reports

We initially processed the reports found on the Canadian TSB web site to remove non-textual content and to properly delimit each section. Then each section was segmented into sentences, part of speech tagging (POS) was applied and each individual word was converted into a normalized form using Porter Stemming algorithm. These manipulations were conducted using the GATE NLP platform [2]. Finally we removed from the internal representation of the problem (*Summary*) and solution (*Findings* sections) components the words belonging to functional grammatical categories carrying no meaning (prepositions, pronouns, conjunctions, adverbs…).

3.2 Conversion of the Texts

As mentioned previously, some sections of the investigation reports contain very detailed information about flight incidents and pertaining factors. To reduce the size and the complexity of these texts, we performed a statistical analysis of the documents to identify salient words that should be retained for recommendation purposes. The lengthy sections can then be replaced by a list of domain keywords.

3.2.1 Named Entity Extraction

The Air investigation reports contain numerous factual information and we decided to convert them into higher-level categories. We used the ANNIE information extraction component of GATE to locate *Date* and *Location* named entities and to insert entity roles in the sections. For example, *Quebec* would be supplemented with its role *LOCATION*. This gives an opportunity to find word associations that could not be extrapolated from specific locations or dates.

3.2.2 Keyphrase Extraction

Keyphrases are word collocations that can play the role of descriptors for long and complex documents relying on a specific domain terminology. Machine learning has recently been applied successfully to automatically extract keyphrases from texts and we used the KEA algorithm [3] to condensate information contained in long sections such as A*nalysis*. The result is a classifier learned to extract information about the type of vessel involved in the incident, the environmental factors and references to equipment devices. To build an extraction model, we initially tagged eight (8) reports and then iteratively applied the algorithm to our test corpus and manually corrected the results. After a few iterations, we ended up with a classifier that selects up to 30 keyphrases for each of the investigation reports of our corpus. To illustrate the results we obtained, here is a keyphrase list for the A03O0341 report (2003).

airstrip turbine	*aircraft got airborne*	*turbine-equipp*	*ELT*
airborne	*morning*	*take-off run*	*loaded*
de Havilland	*passenger*	*local air operator*	*skis*
de Havilland DHC-3	*horizontal stabilizer trim*	*base Armstrong*	*snow*
snowmobile	*surviv passenger*	*metal tower*	*STC*
turbine engine	*M601E turbine engine*	*cover approximat*	*take-off*
improving	*main landing gear*	*inch of snow*	*tail*

3.2.3 Cooccurrence Analysis

As keyphrase extraction is limited to word collocations (i.e. consecutive words), we also applied an algorithm to extract cooccurrences from our reports. Cooccurrences are statistically dependent words that are not required to be adjacent [4]. Sometimes this approach captures relationships that can be interpreted at a semantic level (for instance, relationships between nouns of different phrases). To extract the cooccurrences, we cumulated for all the reports the counts of all the pairs of words present in the same sentence and we applied a Chi² test to select the valid pairs. Here are some examples of cooccurrences extracted from the reports:

civil fault	*forecast weather*	*valv shutoff*	*human injuri*
taken action	*liabl civil*	*chamber re-open*	*precipit drizzle*

3.3 Alignments of Words from Different Sections

At this point, cases consist of sequences of words representing problems (*Summary* and keyword lists) and solutions (*Findings* sections). Alignment of these components consists of finding influences between words belonging to different sequences. To obtain word alignments, we built a translation model $t(prob_i|sol_j)$ which estimates the probability that a problem word can be associated with (or generated from) a solution word. As part of our experimentation, an IBM1 translation model was learnt using an EM algorithm [4]. Here are some examples of probabilistic word relationships:

aerodrom proxim (0.703)	*airborn take-off (0.066)*	*fault ignit (0.278)*
take-off departur (0.22)	*injuri prevent (0.10)*	*injuri minimum (0.61)*

3.4 Sentence Reuse

Once a translation model is built, the reuse phase consists of recommending sentences to the investigator by decreasing value of probability. The probability of a solution sentence is estimated by the Bayes rule $P(Sol|Prob) = P(Prob|Sol)P(Sol)$. And $P(Prob|Sol)$ is partly estimated by cumulating $t(prob_i|sol_j)$ for all the possible word alignments between a solution sentence and an antecedent problem description. Evaluating this reuse scheme in practice is difficult as domain expertise would be required to assess the relevance of each sentence. But we are conducting a leave-one-in evaluation and we obtained a precision of 57% on preliminary trials. However further investigation is required to fully assess the potential of this scheme.

4 Conclusion

We outlined in this paper an approach for the reuse of investigation documents to support the authoring of new incident reports by investigators. The approach relies on statistical techniques to construct models that can be exploited for recommending sentences to the user. Some experiments have been conducted to acquire statistical models from a corpus of 162 reports and we are currently completing a leave-one-in evaluation of this scheme. We would like, for future work, to explore how to replace the translation model with other statistical approaches like cooccurrence analysis.

References

1. Lamontagne, L. ; Lapalme, G. ; (2004) "Textual Reuse for Email Response", *Advances in Case-Based Reasoning*, LNCS, vol. 3155 , Springer-Verlag, pp.234-246.
2. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Environment for Robust NLP Tools and Applications. *Proceedings ACL'02*.
3. Witten I. H., Paynter G. W., Frank E., Gutwin C. and Nevill-Manning C. G. (1999) "KEA: Practical automatic keyphrase extraction." Proc. DL '99, pp. 254–256.
4. Manning, C., Schütze, H.; (1999) *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA.