A Pragmatic Approach to Summary Extraction in Clinical Trials

Graciela Rosemblat
National Library of Medicine
NIH, Bethesda, Maryland
rosem@nlm.nih.gov

Laurel Graham
National Library of Medicine
NIH, Bethesda, Maryland
lagraham@mail.nih.gov

Background and Introduction

ClinicalTrials.gov, the National Library of Medicine clinical trials registry, is a monolingual clinical research website with over 29,000 records at present. The information is presented in static and free-text fields. Static fields contain high-level informational text, descriptors, and controlled vocabularies that remain constant across all clinical studies (headings, general information). Free-text data are detailed and trial-specific, such as the Purpose section, which presents each trial's goal, with large inter-trial variability in length as well as in technical difficulty. The crux of the trial purpose is generally found in 1-3 sentences, often introduced by clearly identified natural language markers.

In the Spanish cross-language information retrieval (CLIR) ClinicalTrials.gov prototype, individual studies are displayed as abridged Spanish-language records, with Spanish static field descriptors, and a manual Spanish translation for the free-text study title. The Purpose section of these abbreviated documents only contains a link (in Spanish) to the full-text English record. The premise was that the gist could be obtained from the Spanish title, the link to the English document, and the Spanish descriptors. However, in a recently conducted user study on the Spanish CLIR prototype, Spanish-speaking consumers did not use the Purpose section link, as doing so entailed leaving a Spanish webpage to go to an English one. Further, feedback from an earlier study indicated a need for some Spanish text in the Purpose section to provide the gist of the trial while avoiding the information overload in the full-text English record. Thus, in an alternative display format, extractive summarization plus translation was used to enhance the abbreviated Spanish document and supplement the link to the English record. The trial purpose--up to three sentences--was algorithmically extracted from the English document Purpose section, and translated into Spanish via post-edited machine translation for display in the Spanish record Purpose section (Rosemblat et al., 2005).

Our extraction technique, which combines sentence boundary detection, regular expressions, and decision-based rules, was validated by the user study for facilitating user relevance judgment. All participants endorsed this alternative display format over the initial schematic design, especially when the Purpose extract makes up the entire Purpose section in the English document, as is the case in 48% of all trials. For Purpose sections that span many paragraphs and exceed 1,000 words, human translation is not viable. Machine translation is used to reduce the burden, and using excerpts of the original text as opposed to entire documents further reduces the resource cost. Human post-editing ensures the accuracy of translations. Automated extraction of key goal-describing text may provide relevant excerpts of the original text via topic recognition techniques (Hovy, 2003).

1 RegExp Detection and Pattern Matching

Linguistic analysis of the natural language expressions in the clinical trial records' Purpose section was performed manually on a large sample of documents. Common language patterns across studies introducing the purpose/goal of each trial served as cue phrases. These cue phrases contained both quality features and the rhetorical role of GOAL (Teufel and Moens, 1999). The crux of the purpose was generally condensed in 1-3 sentences within the Purpose section, showing definite patterns and a limited set of productive, straightforward linguistic markers. From these common patterns, the ClinicalTrials.gov Purpose Extractor Algorithm (PEA) was devised, and developed in Java (1.5) using the native regexp package.

Natural language expressions in the purpose sentences include three basic elements, making them well suited to regular expressions:

a) A small, closed set of verbs (*determine, test*)
b) Specific purpose triggers or cues (*goal, aim*)
c) Particular types of sentence constructs, as in:
This study will evaluate two medications…

PEA incorporates sentence boundary detection (A), purpose statement matching (B), and a series of decision steps (C) to ensure the extracted text is semantically and syntactically correct:

A) To improve regexp performance and ensure that extraction occurred in complete sentences, sentence boundary detection was implemented. Grok (OpenNLP), open source Java NLP software, was used for this task, corpus-trained and validated, and supplemented with rules-based post-processing.

B) Regular expressions were rank ordered from most specific to the more general with a default expression should all others fail to match. The regexp patterns allowed for possible tense and optional modal variations, and included a range of all possible patterns that resulted from combining verbs and triggers, controlled for case-sensitivity. The default for cases that differed from the standard patterns relied solely on the verb set provided.

C) Checks were made for (a) length normalization (a maximum of 450 characters), with purpose-specific text in enumerated or bulleted lists overriding this restriction; and (b) discourse markers pointing to extra-sentential information for the semantic processing of the text. In this case, PEA determines the anchor sentence (main crux of the purpose), and then whether to include a leading and trailing sentence, or two leading sentences or two trailing ones, to reach the 3-sentence limit.

RegExp Patterns	Description	Case
PURPOSE	Sentence label (purpose)	Yes
To VERB_SET	Study action starts section	No
In THIS STUDY	General actions in study	No

Table 1. Some purpose patterns used by PEA

2 Evaluation

Manual PEA validation was done on a random sample of 300 trials. For a stricter test, the 13,110 studies with Purpose sections short enough to include in full without any type of processing or decision were not part of the random sample. Judgments were provided by the authors, one of whom was not involved in the development of PEA code. The 300 English extracts (before translation) were compared against the full-text Purpose sections in the clinical trials, with compression rate averaging 30%. Evaluation was done on a 3-point scale: perfect extraction, appropriate, wrong text. Inter-annotator agreement using Cohen's kappa was considered to be good (Kappa = 0.756987). Table 2 shows evaluation results after inter-rater differences were reconciled:

CRITERIA	TRIALS	RATIO
Perfect extraction	275	92%
Appropriate extraction	18	6%
Extraction of wrong text	7	2%

Table 2: Results: 300 Clinical trials random sample

3 Conclusion

This pragmatic approach to task-specific (purposive) summary extraction in a limited domain (ClinicalTrials.gov) using regular expressions has shown a 92% precision. Further research will determine if this method is appropriate for CLIR and query language display via machine translation and subsequent post-editing in clinical trials information systems for other registries and sponsors.

Acknowledgements

The authors thank Tony Tse and the anonymous reviewers for valuable feedback. Work supported by the NIH, NLM Intramural Research Program.

References

Eduard Hovy. 2003. Text Summarization. In Ruslan Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 583-598). Oxford University Press.

Graciela Rosemblat, Tony Tse, Darren Gemoets, John E. Gillen, and Nicholas C. Ide. 2005. *Supporting Access to Consumer Health Information Across Languages.* Proceedings of the 8th International ISKO Conference. London, England. pp. 315-321

Grok part of the OpenNLP project. [Accessed at http://grok.sourceforge.net]

Simone Teufel and Marc Moens. 1999. Argumentative classification of extracted sentences as a step towards flexible abstracting. In *Advances in Automatic Text Summarization*, I. Mani and M.T. Maybury (eds.), pp. 155-171. MIT Press.