Extractive Summarization in Clinical Trials Protocol Summaries: A Case Study

Graciela Rosemblat, Laurel Graham, and Tony Tse

U.S. National Library of Medicine, National Institutes of Health,
Bethesda, Maryland, USA

Abstract. This paper presents a method for extracting purpose statements from clinical trial protocol summaries. Simple summarization technology based on regular expressions and natural language processing techniques were applied in a controlled environment with structured sectioning to address an expressed user need: providing access to information about the purpose of specific clinical trials, originally in English, to Spanish speakers. Following an analysis of manually annotated data, based on a cascade of criteria, the "Purpose Extractor Algorithm" was developed to select tightly-focused candidate excerpts out of lengthy descriptions, for translation into Spanish. The extracts reduce the translation task and provide purposive content in biomedical text. The results were validated in a focused user study. It is anticipated that this extractive summarization approach may be generalized to documents from other databases as the algorithm can be tailored to different applications or needs.

Keywords: Extractive summarization, biomedical text, regular expressions.

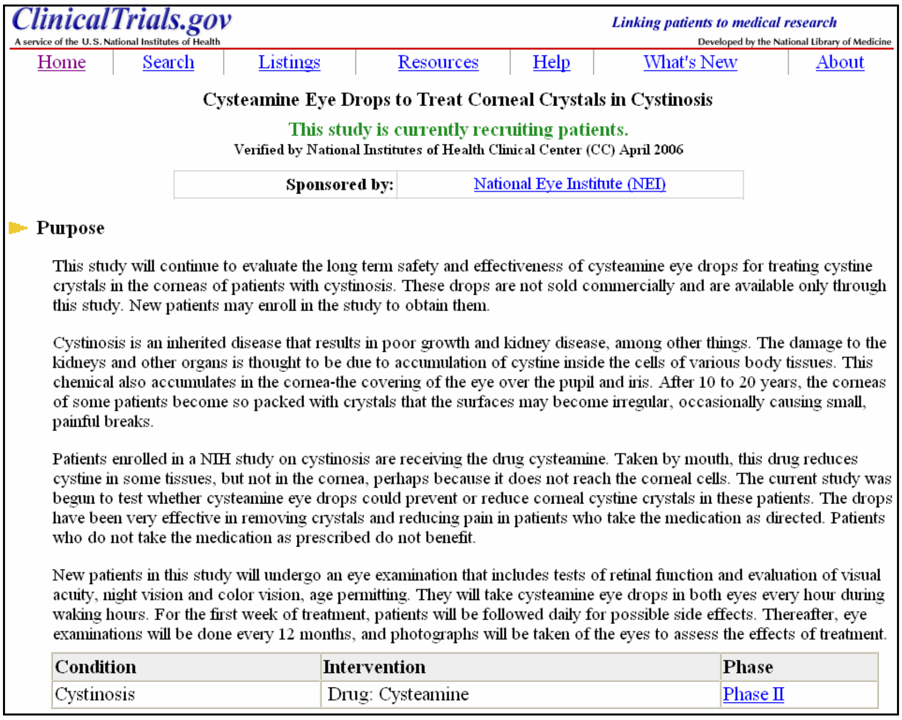
1 Background and Introduction

This paper describes a text extraction approach for summarizing purposive information and uses a case study to illustrate its application to address an actual user need. Unlike other current extractive approaches which focus on single-document summarization [1], this approach targets specific sections of documents (e.g., Purpose section) and, hence, takes advantage of layout design and structured sectioning. The process reported in this paper “condenses” purposive information content, using topic recognition techniques [2] to reduce English-language free text of varying length and detail to three-sentence extracts that convey the gist of a narrative. The free-text purpose descriptions are generally authored by different data providers, such as phramaceutical companies, federal organizations, and other institutions that conduct clinical trials. Thus, the position of the specific purpose sentences varies greatly within the section. This rules out location heuristics as used in the Edmundsonian paradigm for ranking sentences for extraction [3]. Our approach relies on the definite linguistic and discourse (rhetorical) patterns used by the authors of the purposive phrases. It flags sentences that contain such phrases for extraction, not as one of several features [4], but when used in specific sentence constructs. This extraction method also differs from similar approaches on unsupervised detection of semi-fixed cue phrases [5], or those which impose intrasentential distance restrictions [6], since it does not depend on whether a specific syntactic or grammatical relation holds between the constituents of the phrase.

The context for this work was an earlier project to develop a Spanish-English cross language information retrieval (CLIR) prototype of the US National Library of Medicine (NLM) clinical trials registry, ClinicalTrials.gov (http://clinicaltrials.gov). This English-language website provides information on about 42,000 clinical research studies; each record includes “static” and free-text data fields [7]. Static fields contain phrases and controlled vocabulary terms that rarely change (e.g., standard headings, such as “Purpose,” and information about the enrollment status, condition(s) being studied, and intervention name(s)). In contrast, free-text fields consist of detailed, trial-specific data subject to variable text length and frequent updates, such as the Purpose Description (Table 1). This important field, which provides information about the purpose of the study, appears in the top half of clinical trial records (Figure 1).

Table 1. Examples illustrating variability in the Purpose Description of clinical trial records.

Clinical Trial ID	Text in Purpose Description Field
NCT00022360	RATIONALE: Drugs used in chemotherapy use different ways to stop tumor cells from dividing so they stop growing or die. PURPOSE: Phase I trial to study the effectiveness of taurolidine in treating patients who have recurrent or progressive glioma.
NCT00188370	A group of researchers at the Ontario Cancer Institute/ Princess Margaret Hospital have discovered that a very specific form of cell death 'apoptosis' can be detected using high-frequency ultrasound imaging. This type of cell death is recognized to occur in tumours in response to various different chemotherapeutic drugs and in response to radiation therapy. This group of researchers has confirmed that high-frequency ultrasound can detect apoptosis in response to tumour treatments experimentally using cell culture and experimental animal systems. The ultrasound approach is now being evaluated clinically in a 3-year clinical trial enrolling a target of 200 patients including Hodgkin's disease and non-Hodgkin's disease lymphoma patients, melanoma patients and patients with basal cell carcinoma. Our hope is to be able to use this type of imaging system in the future to clinically monitor the effects of therapy on tumours and rapidly detect tumours which are not responding so that changes in therapy can be made much quicker than presently possible.
NCT00004697	OBJECTIVES: I. Determine whether intravenous choline supplementation will reverse the hepatic steatosis and improve liver function in patients who receive long term total parenteral nutrition.


ClinicalTrials.gov
A service of the U.S. National Institutes of Health
Linking patients to medical research
Developed by the National Library of Medicine
Home | Search | Listings | Resources | Help | What's New | About

Cysteamine Eye Drops to Treat Corneal Crystals in Cystinosis

This study is currently recruiting patients.
Verified by National Institutes of Health Clinical Center (CC) April 2006

Sponsored by:	National Eye Institute (NEI)

Purpose

This study will continue to evaluate the long term safety and effectiveness of cysteamine eye drops for treating cystine crystals in the corneas of patients with cystinosis. These drops are not sold commercially and are available only through this study. New patients may enroll in the study to obtain them.

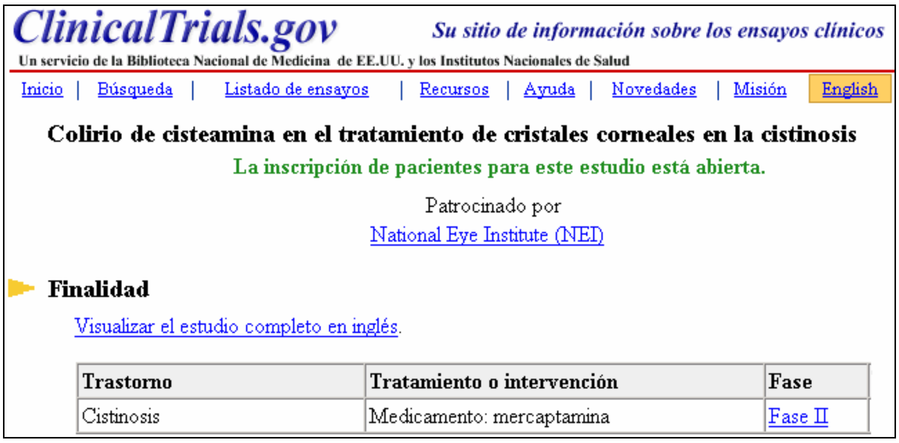
Cystinosis is an inherited disease that results in poor growth and kidney disease, among other things. The damage to the kidneys and other organs is thought to be due to accumulation of cystine inside the cells of various body tissues. This chemical also accumulates in the cornea-the covering of the eye over the pupil and iris. After 10 to 20 years, the corneas of some patients become so packed with crystals that the surfaces may become irregular, occasionally causing small, painful breaks.

Patients enrolled in a NIH study on cystinosis are receiving the drug cysteamine. Taken by mouth, this drug reduces cystine in some tissues, but not in the cornea, perhaps because it does not reach the corneal cells. The current study was begun to test whether cysteamine eye drops could prevent or reduce corneal cystine crystals in these patients. The drops have been very effective in removing crystals and reducing pain in patients who take the medication as directed. Patients who do not take the medication as prescribed do not benefit.

New patients in this study will undergo an eye examination that includes tests of retinal function and evaluation of visual acuity, night vision and color vision, age permitting. They will take cysteamine eye drops in both eyes every hour during waking hours. For the first week of treatment, patients will be followed daily for possible side effects. Thereafter, eye examinations will be done every 12 months, and photographs will be taken of the eyes to assess the effects of treatment.

Condition	Intervention	Phase
Cystinosis	Drug: Cysteamine	Phase II

Fig. 1. Top half of an individual clinical trial record. Note lengthy Purpose Description text.


ClinicalTrials.gov
Su sitio de información sobre los ensayos clínicos
Un servicio de la Biblioteca Nacional de Medicina de EE.UU. y los Institutos Nacionales de Salud
Inicio | Búsqueda | Listado de ensayos | Recursos | Ayuda | Novedades | Misión | English

Colirio de cisteamina en el tratamiento de cristales corneales en la cistinosis

La inscripción de pacientes para este estudio está abierta.

Patrocinado por
National Eye Institute (NEI)

Finalidad

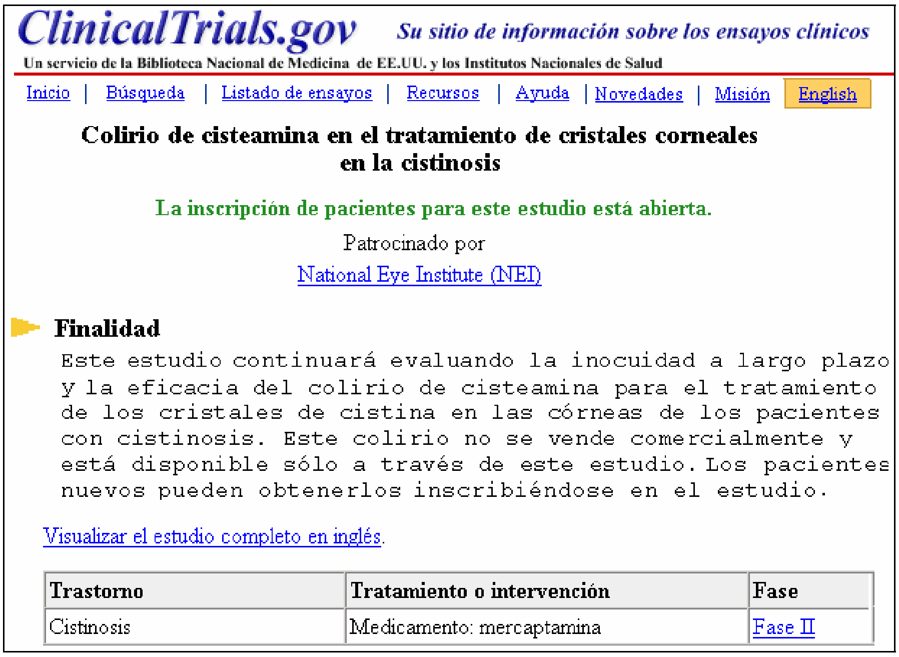
Visualizar el estudio completo en inglés.

Trastorno	Tratamiento o intervención	Fase
Cistinosis	Medicamento: mercaptamina	Fase II

Fig. 2. Top half of the initial Spanish language version of the clinical trial record shown in Figure 1. There is no descriptive text in the Purpose section.

The translation of free-text fields into Spanish proved to be a significant challenge: static fields allowed for one-time translation of descriptors for display in multiple records, an option not viable for free-text fields due to their size and variability. In the initial phase of the project, nearly all free-text fields (e.g., Purpose Description) in the Spanish-language record (Fig. 2) were displayed with links labeled in Spanish to the corresponding English record. The title, essential for Spanish-speaking users, was the only free-text field to be translated into Spanish via manual post-editing of machine translation output. However, applying the same technique to other free-text fields was not feasible due to the human cost involved in post-editing much longer text passages. As an alternative, links were provided to the English-language full text for use by Spanish-English bilingual users. Further, supplying a link to the original trial in English resolved the issue of updating and synchronizing potentially changing data.

During a 2004 pre-pilot study using these translated clinical trial records, participants indicated that, along with the link, some purpose information was needed in each record to allow users to obtain the gist of the study. Based on this feedback, an alternate format [8] was developed and used in a subsequent user study (Fig. 3). This new format included an algorithmically extracted excerpt from the English-language full text Purpose Description. The excerpt was then machine-translated into Spanish, manually post-edited, and included in the Spanish records [9]. When shown both formats, Spanish-speaking user study participants unanimously preferred this new format:


ClinicalTrials.gov
Su sitio de información sobre los ensayos clínicos
Un servicio de la Biblioteca Nacional de Medicina de EE.UU. y los Institutos Nacionales de Salud
Inicio | Búsqueda | Listado de ensayos | Recursos | Ayuda | Novedades | Misión | English

Colirio de cisteamina en el tratamiento de cristales corneales en la cistinosis

La inscripción de pacientes para este estudio está abierta.

Patrocinado por
National Eye Institute (NEI)

Finalidad

Este estudio continuará evaluando la inocuidad a largo plazo y la eficacia del colirio de cisteamina para el tratamiento de los cristales de cistina en las córneas de los pacientes con cistinosis. Este colirio no se vende comercialmente y está disponible sólo a través de este estudio. Los pacientes nuevos pueden obtenerlos inscribiéndose en el estudio.

Visualizar el estudio completo en inglés.

Trastorno	Tratamiento o intervención	Fase
Cistinosis	Medicamento: mercaptamina	Fase II

Fig.3. Top half of the Spanish record shown in Figure 2, with purpose statement included.

The strength of this work comes from the application of summarization to:

- facilitate bilingual access to purpose information in clinical trial records and
- reduce the workload by translating excerpts instead of full text.

Translating the entire Purpose Description text was not viable, as it may span many paragraphs. While machine translation can reduce the burden, human post-editing is needed to ensure accuracy. Hence, a Purpose Extractor Algorithm was developed to extract candidate purpose statements for translation.

2 Text Analysis

Manual linguistic analysis of the natural language expressions used in the Purpose Description of the clinical trial records was performed (in English) by the first author of this manuscript, on a representative sample of documents. The analysis revealed common patterns across documents in the language used to introduce a "key purpose sentence" in each Purpose Description. These textual markers combined quality / rhetoric features [10] in a single sentence. Despite its variable position within the text in each document, the key purpose sentence followed a distinct pattern, rendering a limited set of straightforward and productive linguistic markers, thereby eliminating the need for automatic acquisition of indicators. However, as the markers are not domain-specific but style-dependent, some knowledge modeling may be needed to apply this extractive process to text from other databases or systems in other domains.

2.1 Rationale for Regular Expressions

Initially, the first two authors of this paper had considered that the task for text extraction could be equated with knowledge representation, and a template system comprising a handful of alternate sentences had been discussed. Each template sentence contained empty slots for key components, namely, the labels for diseases and/or interventions studied in each trial. The fillers for the empty slots would be extracted from the English-language clinical trial purpose text by an NLM-developed, knowledge-based semantic interpreter (SemRep). SemRep uses underspecified syntactic analysis and structured domain knowledge from the NLM Unified Medical Language System$^{®}$ (UMLS$^{®}$) to identify semantic predications in biomedical text [11].

However, the rigidity of the template system (one-size-fits-all) was a critical concern, as it would not provide enough flexibility for representing long and coordinated sentences with rich descriptive information, or which simply did not fit the templates. Data providers are not required to follow specific or uniform formatting guidelines for free-text fields. This limitation soon underscored the need for a different approach.

The focus was then turned to regular expressions. The consistent language in the purposive text rendered it well suited for extractive summarization techniques based on regular expressions (regexp), long used in natural language processing [12]. Three basic elements fit this approach:

- A small, closed set of 32 verbs: *ascertain, assess, attempt, characterize, collect, compare, conduct, determine, develop, ensure, estimate, evaluate, try, examine, explore, extend, follow, hypothesize, identify, intend, investigate, look at, measure, monitor, observe, plan, propose, provide, research, seek, study, test*;
- Purpose triggers or cues, such as *purpose, objective, aim* and *goa*l; and
- Particular types of sentence constructs, as in: *The objective of this study is to assess the efficacy of… This pilot study will evaluate…* Or: *To determine…*

2.2 Identification of Regular Expressions and Sentence Delimiters

Text extraction using regular expressions for identifying and matching purposive sentences was done in Java 1.5, native regexp package. The expressions were ranked by specificity, including a default expression for match failures. The most specific expressions matched using heading identifiers or sentence constructs. More general expressions looked for verbs or introductory wording, such as *In this study…*Patterns allowed for possible tense and modal variations in the verbs described. Thus, the Purpose Extractor Algorithm includes a range of all possible patterns that could result from combining verbs and triggers, controlled for case-sensitivity. The default value for those cases that did not follow the standard format relied solely on the verb set described.

For length normalization, a maximum count of 450 characters (including spaces) was added to the Purpose Extractor Algorithm. The maximum length rule applied only in cases of multiple sentences extracted from the Purpose description, so that:

- If the key purpose sentence matched by the regular expression exceeded 450 characters, the entire sentence was extracted despite its length;
- If the entire Purpose Description text consisted of 3 sentences or less, then the number of sentences overrode the character length parameter, and the entire description was extracted, irrespective of length;
- If the 450th character fell in mid-sentence (for descriptions containing more than 3 sentences), then the extracted text was trimmed back to the preceding sentence delimiter; and
- For numbered or bulleted lists of multiple purposes, all items were extracted and the maximum length rule did not apply.

To improve regexp performance and ensure that extraction occurred in complete sentences, Grok [13], a freely available open source Java NLP software, was used. Grok uses maximum entropy modeling techniques to perform tasks including sentence boundary detection. A new model was trained using the entire set of Purpose Descriptions from the ClinicalTrials.gov XML documents, as of May 2005 (about 13,800 records). The training corpus was annotated using a combination of processing techniques via a Perl script, with human post-editing to correct errors from the script. Once the model was trained, the sentence boundary results were validated using the full set of 27,489 XML documents, as of February 2006.

3 Description of the Algorithm

The Purpose Extractor Algorithm includes sentence boundary detection (3.1), pattern matching (3.2), and a series of checks and filters (3.3) to ensure semantic and syntactic cohesion in the extracted text.

3.1 Sentence Boundary Detection

The Purpose Extractor Algorithm relies heavily on sentence boundary information. Determining whether to include the full text of the English-language Purpose Description is based on the total number of sentences, regardless of character count.

Sentence boundary detection problems were compounded by two phenomena:

- Punctuation errors, including missing periods, run in sentences, and the like;
- Punctuation other than periods: colons, slashes, semicolons, and question marks.

As a result, post-processing rule-based logic was added to correct errors in sentence boundary detection by Grok.

The step-wise logic of the Purpose Extractor Algorithm is as follows:

- The algorithm breaks the full clinical trial Purpose Description into individual sentences using Grok, keeping paragraph boundary markers.
- Post-editing mechanisms in the code correct poorly formed text, as when there are no spaces and no periods between two sentences, or a period is included at the end of a sentence and no space is left between that period and the next sentence. In these cases, the algorithm looks for a capital letter signalling the beginning of a new sentence and includes the missing boundary markers.
- For Purpose Descriptions that consist of 3 sentences or less, the entire description is returned as the extract.

3.2 Pattern Matching

Each regexp (starting with the most specific one – see Table 2) was tested for a match on every sentence in the Purpose Description:

- More specific regular expressions are given a higher weighting. Thus, if a sentence matching the more specific patterns is found anywhere in the Purpose text, it is considered for extraction before more general matches occurring earlier in the text.
- If sentence matching fails, the default mechanism is triggered, and a match using the verb set is attempted. The first match is accepted, relying on text order in the Purpose description. This mechanism is the most general of all purpose statement patterns.
- Once a match is found, the complete matching or "anchor" sentence is extracted from the English-language Purpose Description.
- The preceding (leading) and following (trailing) sentences are extracted in a two-stage process, to create the a summarization description:
  - If no leading sentence exists, then two trailing sentences are extracted.

- o If no trailing sentence follows the anchor, then two leading sentences are extracted.

For semantic coherence, leading sentences must stay within the same paragraph, but anchor-trailing sentences may cross a paragraph boundary.

Table 2 provides a high-level overview of the purpose patterns by degree of specificity. Variables in all caps represent synonymous or categorial sets, so that THIS includes *this* and *the*; STUDY includes *study, trial, research, protocol, investigation*; and so forth.

Table 2. Purpose patterns used by the Purpose Extractor Algorithm by decreasing specificity.

Regular Expressions Patterns	Description
PURPOSE	String literal, all-cap
THIS OBJECTIVE of THIS STUDY	Introducing the goal
THIS STUDY(MODAL|AIM|TENSE)? VERB_SET	Specific information
(In)THIS STUDY (MODAL|AIM|TENSE)? VERB_SET	Action of this study
THIS OBJECTIVE AIM VERB_SET	Study goal in action
THIS PART STUDY AIM	Phrase specific aim
THIS STUDY AIM	Study's aim
To VERB_SET	Action, sent.-initial
STUDY (TENSE) VERB_SET	Study action
In THIS STUDY	Actions in study
THIS (STUDY)? STUDY	Study meta-reference
We (TENSE) VERB_SET	Researcher's actions
VERB_SET	Default rule

3.3 Semantic and Syntactic Checks and Filters

- The extracted purpose summary is checked to ensure that the text does not exceed 450 characters, including white space.
- Purpose-specific numbered or bulleted lists are extracted in their entirety, regardless of character count.
- Leading sentences marked for extraction that are part of a bigger discourse (e.g., *To accomplish this, Despite the above, Therefore, Thus,* and the like) are flagged. These discourse markers are clear indicators that extra-sentential information is needed for the semantic processing of the text. As the algorithm currently does not include reference resolution, leading sentences with these markers are discarded. In these particular instances, the extracted text consists of the anchor sentence and up to two trailing sentences.

For Spanish-language display, the extracted text was run in batch mode through a machine translation system, followed by manual post-editing. The post-editor received a file with three types of information: each trial’s unique identifier, the entire English-language Purpose Description, and the raw translation of the algorithmically extracted text. The post-editor used this information to determine the relevance and appropriateness of the extracted text.

The decision points in the Purpose Extraction Algorithm are summarized in the flowchart (Fig. 4):

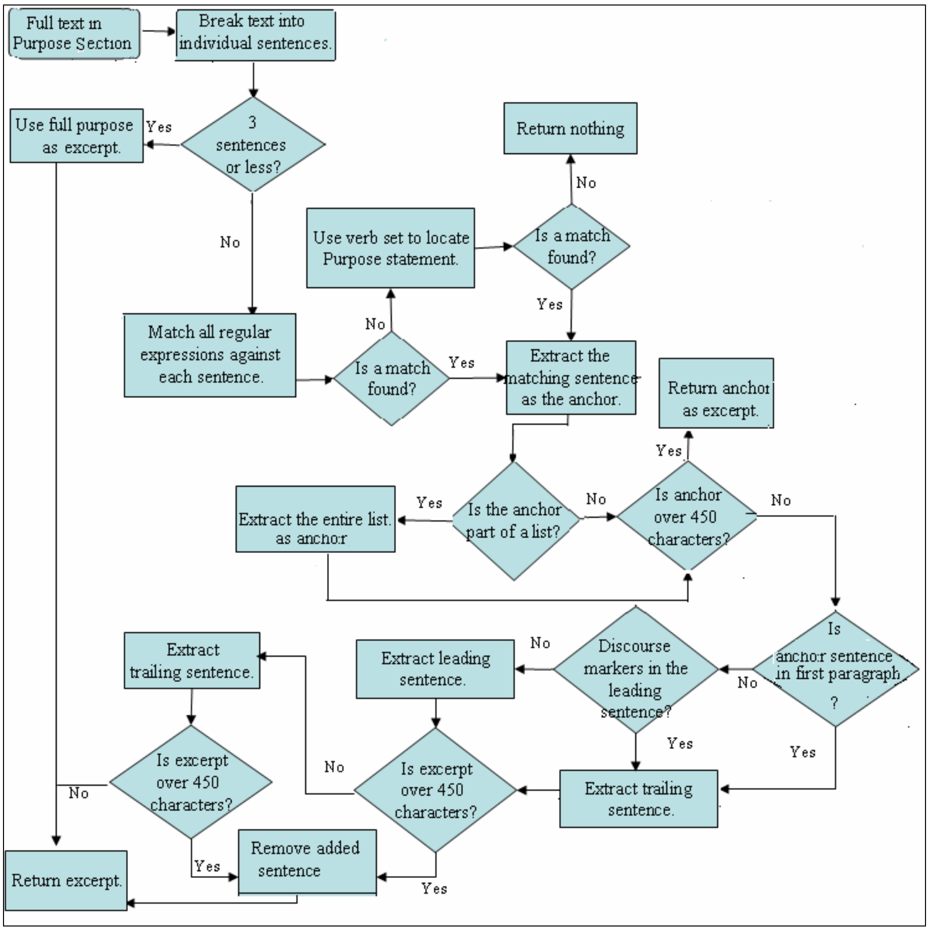

Full text in Purpose Section
Break text into individual sentences.
Use full purpose as excerpt.
Yes
3 sentences or less?
No
Return nothing
No
Use verb set to locate Purpose statement.
Is a match found?
Yes
Match all regular expressions against each sentence.
No
Is a match found?
Yes
Extract the matching sentence as the anchor.
Return anchor as excerpt.
Yes
Extract the entire list as anchor
Yes
Is the anchor part of a list?
No
Is anchor over 450 characters?
No
Is anchor sentence in first paragraph ?
No
Discourse markers in the leading sentence?
No
Extract leading sentence.
Extract trailing sentence.
Yes
Yes
Extract trailing sentence.
Is excerpt over 450 characters?
No
Is excerpt over 450 characters?
No
Yes
Remove added sentence
Yes
Return excerpt.

Fig. 4. Purpose Extractor Algorithm flowchart.

4 Evaluation

The Purpose Extractor Algorithm was applied to all 27,489 clinical trial records, as of February 2006. In 64 trials (0.2%) no excerpt was extracted due to ambiguous language or atypical verb usage. In 13,110 trials (48%), no further processing was necessary, as the Purpose Descriptions met the algorithm requirements for a short summary, and the entire text was returned as an excerpt (Figure 4). For the remaining 14,315 clinical trials, the compression rate of the extracted text averaged 30%.

For the evaluation, a random sample of 300 Purpose Descriptions was selected from the 14,315 clinical trials. For a stricter test, the validation set excluded the 13,110 trials with Purpose Descriptions that met the conditions for all text extraction,

and the 64 trials without extracted excerpts. Performance was evaluated in two ways: a) a human judge developed a Gold Standard for the 300 Purpose Description evaluation set, without access to extracts or summaries; the results of the Purpose Exctractor Algorithm were then compared against the Gold Standard, and
b) manual, multiple-annotator (n=3) evaluation comparing the algorithmically extracted Purpose excerpts (before translation) with their corresponding full-text Purpose Descriptions.

Gold Standard Evaluation. To derive a Gold Standard to measure the accuracy of the Purpose Extractor Algorithm, a two-column document was prepared with the unique number identifiers for the 300 clinical trials in the random sample, and the English-language Purpose Description text in the second column. A staff physician familiar with the clinical trial protocols was tasked with highlighting the key sentence that best represented the purpose or crux of the study within each Purpose Description. Since the algorithm extracts could be up to 3 sentences long, a "match" was defined as the selection by the human judge of any one of the sentences extracted automatically. Color-coding was used to distinguish between primary and secondary purposes, if present. To avoid any potential bias, the judge had no knowledge of the Purpose Extractor Algorithm and was not shown its output of 300 extracted excerpts. The first author determined whether or not the algorithmically extracted text matched the sentences marked by the human judge based on two criteria (Table 3):

Table 3. Degree of agreement between algorithmically extracted text and human judge ratings or Gold Standard (GS)

N=300 Extracts from Purpose Descriptions		
Criteria	Trials	%
1. Extraction met human judge criteria (GS)	269	90%
2. Extraction did not meet human judge criteria (GS)	31	10%

Failure analysis of extracted text that did not coincide with the Gold Standard resulted in three main error categories (Table 4). The most common category (70% of failures) resulted from language problems in the Purpose Description, including the following: (1) failure of the trial to state a purpose (usually observational trials); (2) providing several different purposes throughout the Purpose Description; or (3) stating the same purpose twice using similar but different wording (the algorithm selected one purpose sentence while the judge selected another).

Table 4. Failure Analysis: Purpose Extractor Algorithm compared to the Gold Standard.

N=31 Purpose Descriptions that Failed Gold Standard Criteria			
Error Category	Explanation	n=	%
Language issues (ambiguities); narrative not focused	Verb/noun ambiguity (e.g.: test); ambiguous language	3	71%
	Purpose not clearly stated, duplicated; many purposes stated throughout	16	
	Many verbs of vb_set used throughout	3	
Scope: Algorithm too narrowly defined	Strict CASE, should be relaxed	1	19%
	Verb/cue not included as marker	3	
	Purpose not sentence-initial	2	
Algorithm failure	Requires further analysis	3	10%

Annotator Evaluation. An independent evaluation on the same sample of 300 Purpose Descriptions was conducted by the first two authors and a physician not familiar with clinical trials. They each compared the extracts (before translation into Spanish) with the full Purpose Description text. As the intent of the extracted text was to facilitate user understanding of the gist of the study, internal coherence of the excerpts was considered, based on a 3-point scale:

- *Perfect extraction*: optimal performance of the algorithm, where the key purpose sentence was extracted from the Purpose Description text;
- *Appropriate extraction*: the extracted text did not describe the purpose but provided key study data; and
- *Extraction of wrong text*: the extracted text did not describe either the purpose of the study or key information.

Inter-annotator agreement using Cohen's kappa was considered fair (Kappa = 0.5436). Table 5 shows reconciled evaluation results for all evaluators:

Table 5. Reconciled annotator evaluation results (n=3) for the Purpose Extractor Algorithm.

N=300 Clinical Trials Purpose Descriptions		
	Trials	Ratio
Perfect extraction	266	89%
Appropriate extraction	22	7%
Extraction of wrong text	12	4%

4 Discussion

Although there was a high level of agreement (280 out of 300 trials) between the Gold Standard and annotator evaluations described in the preceding section, areas of divergence fell into two categories:

1) While the Gold Standard focused on accuracy, the annotators applied stricter guidelines by rating the algorithm extracts on coherence in addition to correct Purpose extraction. Thus, even though the algorithmically extracted text may have agreed with the Gold Standard, the annotators did not consider these cases of *perfect extraction* if:

- Discourse markers that referred to extra-sentential information introduced the key purpose sentence, and the reference was not resolved in the leading sentence(s);
- Acronyms contained in the key purpose sentence were not expanded in the anchor sentence or in the leading sentence(s). This resulted in lack of coherence or lack of clarity in the extracted text.

2) When multiple purposes were scattered throughout the Purpose Description without any indication of ranking (in terms of primary or secondary purposes), or the purpose was stated twice in the same study with some linguistic variation, the Gold Standard judge picked one of them, while the algorithm picked another. The annotators often considered these as cases of *Perfect extraction* because the judge's choice was not motivated by ranking, and either would have satisfied his criteria. Nevertheless, the extractions did not match the Gold Standard.

The high scores and high level of agreement in the evaluations is partly due to the format of many Purpose Descriptions, which largely facilitates the extraction task of the Purpose Extractor Algorithm: in 30% of the trials in our 300-study random sample (and 20% of the current data set of 42,000 trials), the Purpose Description is composed of two main capitalized labels, followed by a colon, "RATIONALE" and "PURPOSE", as in the first row of Table 1. The Purpose text can then be easily identified by the "PURPOSE" label. This label corresponds to the first regular expression pattern in Table 2. In contrast, the "OBJECTIVES" label in the last row of Table 1 is not very commonly found.

Lastly, the second evaluation by the Spanish-speaking participants in the user study on the CLIR prototype validated the performance of the Purpose Extractor Algorithm as an effective solution to their actual user need [8].

Limitation. The evaluation was conducted on the Purpose Description text as opposed to the entire clinical trial record. For the Gold Standard judge, this was an important limitation in the few cases (about 4) where no purpose was stated in the Purpose Description. Some trial records offer some indication of purposive information in the other fields or sections of each record (e.g., title, study design), but this is not often the case. In order to conduct a parallel evaluation of the same data, the judge based his decisions entirely on the Purpose Description as presented in the two-column document prepared for the evaluation. Viewing the rest of the clinical trial record before disqualifying the trial for not indicating the purpose may have resulted in slightly different results.

4 Conclusions and Future Work

This paper illustrates the implementation of a pragmatic approach to summary extraction using regular expressions, with a high precision in purposive text extraction for ClinicalTrials.gov Purpose Description text. Regular expressions are well suited for capturing uniform patterns in natural language, as in the records of this clinical trials registry, where typically a single sentence often conveys the crux of the study. Medical information systems (or those in other domains) needing summary extractions for display or search may use this approach as an efficient, cost-effective mechanism, since patterns can be tailored to the natural language style of each system and specific information needs (e.g., "results" as opposed to "purpose").

Optimization of the algorithm will focus on extending and improving performance. In-depth failure analysis has effectively highlighted areas of improvement for wrong text or no text extraction, such as relaxing some of the algorithm conditions (e.g., eliminating case sensitivity for some key expressions, extending the verb set, or adding additional productive cue phrases) and including a mechanism for acronym expansion to increase semantic coherence and understanding. When the extracted text contains an acronym, its expansion is usually located in one of the first sentences of the first paragraph of the Purpose description, as part of the background information, while the purpose-specific text may be one or two paragraphs down. Similarly, anaphora resolution techniques not currently implemented may lead to further improvement. Other information such as part of speech tags, noun phrase boundaries, and concept information via the UMLS® may improve regular expression performance. For generalizability, future research must include validation on other medical information systems, ensuring that regular expressions are not based on sponsor-specific text for the clinical trial registry described in this project.

Acknowledgments. This work was supported by the Intramural Research Program, National Institutes of Health, National Library of Medicine. We would like to thank the two physicians that participated in the evaluation described in this paper and the two anonymous reviewers of this manuscript.

References

1. Afantenos, S. D., Karkaletsis, V., Stamatopoulos, P.: Summarization from Medical Documents: A Survey. Artificial Intelligence in Medicine. Vol.33. (2005) 157–177
2. Hovy, E.: Automated Text Summarization. In: Mitkov, R. (ed.): The Oxford Handbook of Computational Linguistics. (2005) 583–598
3. Edmundson, H. P.: New Methods in Automatic Abstracting. Journal of the Association for Computing Machinery. Vol.16(2). (1969) 264–285
4. Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. 18th ACM-SIGIR Conference Proceedings. (1995) 68–73
5. Abdalla, R., Teufel, S.: A Bootstrapping Approach to Unsupervised Detection of Cue Phrase Variants. In: Proc COLACL. (2006) 921–928
6. Paice, C.: The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-Indicating Phrases. In: Oddy, R. N., Rijsbergen, C. J., Williams, P. W. (eds.): Information Retrieval Research. (1996) 172–191
7. Rosemblat, G., Tse, T., Gemoets, D., Gillen, J.E., Ide, N.C. Supporting access to consumer health information across languages. ICML9. Brazil. (2005) Accessed at http://www.icml9.org/program/track6/
8. BearingPoint, Inc.: Focused Processed Evaluation to Assess the Usefulness and Effectiveness of a ClinicalTrials.gov Spanish Prototype System. (2006) Unpublished report

9. Rosemblat, G., Graham, L.: A Pragmatic Approach to Summary Extraction in Clinical Trials. Poster Abstract. In: Proc. HLT-NAACL BioNLP Workshop. New York (2006) 124
10. Teufel, S., Moens, M.: Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In: Mani, I., Maybury, M. T. (eds.): Advances in Automatic Text Summarization. (1999) 155–176
11. Rindflesch, T. C., Fiszman, M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. In: Journal of Biomedical Informatics. 36(6) (2003) 462–77
12. Jurafsky, D., Martin, J. H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. (2000)
13. Grok, OpenNLP project. Accessed at http://grok.sourceforge.net
14. UMLS® Knowledge Sources. Accessed at http://umlsks.nlm.nih.gov/