Paper 180-2012

Error Reduction and Report Automation Approaches for Textually Dense Pharmaceutical Regulatory Conformance Incident Data

Barry deVille, Mark Wolff, SAS Institute Inc., Cary, NC

ABSTRACT

The Periodic Safety Update Report (PSUR)[1] is a significant feature of the post-trial, in-market monitoring of drug efficacy. The manufacturer's response to a drug-use incident is comprehensive: new literature must be searched, a wide range of interactions with related agents must be examined, and finally, a complete clearance must be claimed. This response is a huge information-processing burden due to the many factors at play and because so much of the information – starting with the original exception report (including numerical data) – is delivered in unstructured textual format.

This presentation discusses a proof of concept carried out on behalf of a major pharmaceutical manufacturer that deployed a variety of text content identification and manipulation techniques to eliminate manual error and automate the report production.

INTRODUCTION

Text Analytics in the reporting domains of health and life sciences often involves the analysis and preparation of data for various pharmacovigilance applications such as PSUR and Vaccine Adverse Event Reporting System (VAERS [2]), as well as for the collection and summarization of treatment histories ranging from episodic, sometimes serial, medical procedures and hospitalization histories.

As described in Wikipedia (http://en.wikipedia.org/wiki/Text_analytics), the overarching goal of text analytics is

> *"… to turn text into data for analysis via application of natural language processing (NLP) and analytical methods."*
>
> *"The term also describes that application of text analytics to respond to business problems, whether independently or in conjunction with query and analysis of fielded, numerical data. It is a truism that 80% of business-relevant information originates in unstructured form, primarily text. "*

Our work in this area is a first-class example of the text analytic nature of pharmacovigilance:

- Much of the data – in the range of 80% – is unstructured.
- Linguistic approaches, such as NLP, are required to identify drug treatments, procedures, and effects.
- A number of analytic approaches that involve numerical manipulation of various kinds are required to identify sums of treatment events, ranges of dosage, onset of conditions, and so on.
- Advanced linguistics coupled with the numerical assessment of ranges and differences are required to detect such nuances as "somewhat elevated" and "10% increase."

In the course of our work, we have identified a number of basic operations that serve as building blocks which, taken individually (or combined together), transform raw inputs into semi-finished information products fit for rapid human review and storage for subsequent retrieval and publication.

Our goal here is to describe these building blocks and so contribute to the development of a syntax and semantics of a text analytic language suitable for describing and summarizing medical incident and procedure reports.

ELEMENTARY OPERATIONS AS BUILDING BLOCKS TO RESULTS

The basic techniques in report construction involve the construction of facilities that can:

- detect events (and associated timing and other details)
- detect linked objects (and associated attributes)

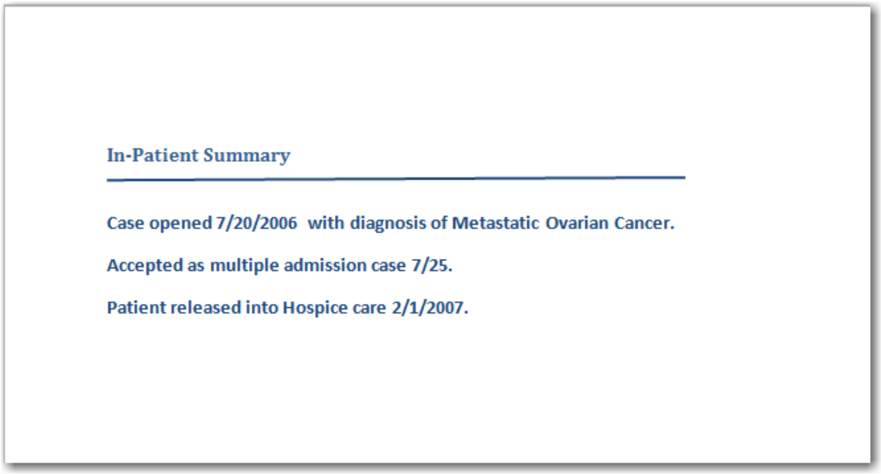
[1] Examples of PSUR reports are available at www.mhra.gov.uk/home/groups/pl-a/documents/.../con2014940.doc, www.arzneimittelinstitut.de/download/Vol9en.pdf

[2] Wikipedia, available at (http://en.wikipedia.org/wiki/Vaccine_Adverse_Event_Reporting_System)

EVENT AND OBJECT RECOGNITION AND SUMMARIZATION

Event and object recognition is shown in Figure 1. This example, taken from preparatory work in the application of text analytics to medical records (in-patient hospitalization records analysis), we see a treatment summary. The basic content of the summary has been identified by scanning hospitalization admission and treatment files (merged together using a patient identifier field).

- The earliest date on file, coupled with an analysis of the “Notes” field, enables us to establish the admission date.
- An analysis of the “Diagnosis” field enables us to establish the formal medical diagnosis.
- Scanning the events records enables us to count and sum the number of treatments. For example, we establish that this patient has had multiple admissions, so we can produce a notation to that effect.
- The “Notes” field on the latest data entry enables us to establish the disposition of the patient and associated date.



In-Patient Summary

Case opened 7/20/2006 with diagnosis of Metastatic Ovarian Cancer.

Accepted as multiple admission case 7/25.

Patient released into Hospice care 2/1/2007.

Figure 1: Example Summary Record (Hospitalization)

Running behind the scenes is the operation of linguistic and computational procedures that are a standard part of SAS text analytics.

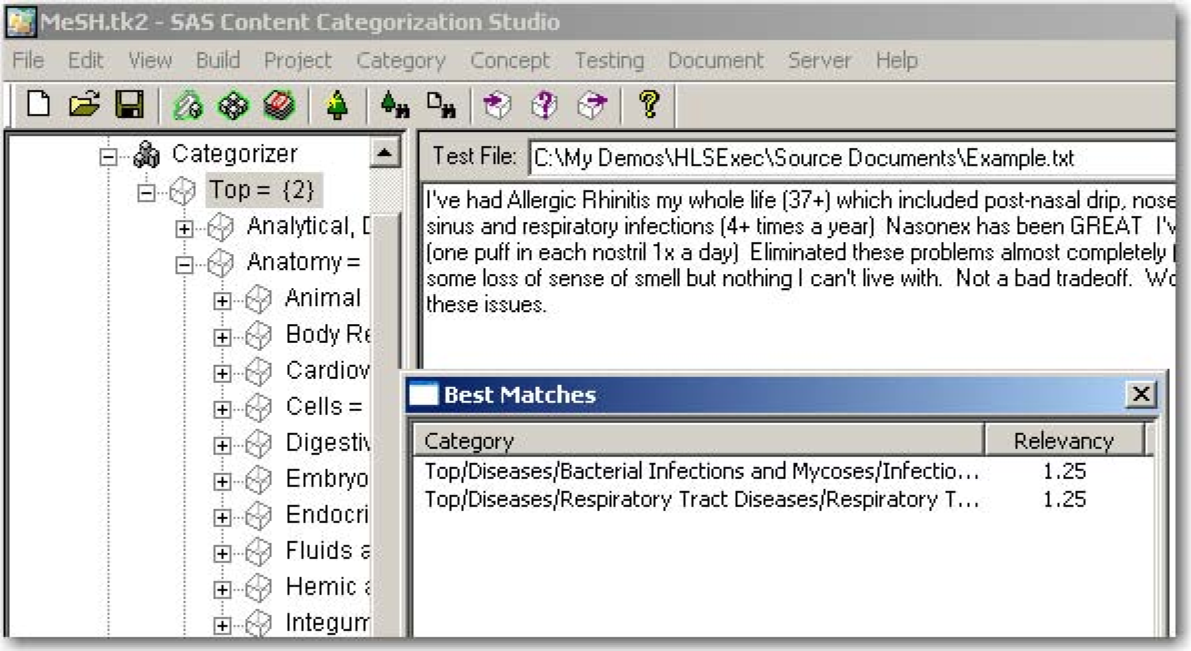
ELEMENTARY OPERATIONS

Categories, Concepts, and Event-Object Summarization and Arithmetic

The SAS®Text Analytics toolset enables us to identify a number of situations that arise in PSUR report production.

Category Identification

Figure 2 demonstrates how the SAS® Enterprise Content Categorization is used to identify the condition (category) reported in the text. As shown in the figure, “Respiratory Diseases” is identified as one of the primary characteristics of this condition report.



Test File: C:\My Demos\HLSExec\Source Documents\Example.txt

I've had Allergic Rhinitis my whole life (37+) which included post-nasal drip, nose sinus and respiratory infections (4+ times a year) Nasonex has been GREAT I'v (one puff in each nostril 1x a day) Eliminated these problems almost completely some loss of sense of smell but nothing I can't live with. Not a bad tradeoff. Wo these issues.

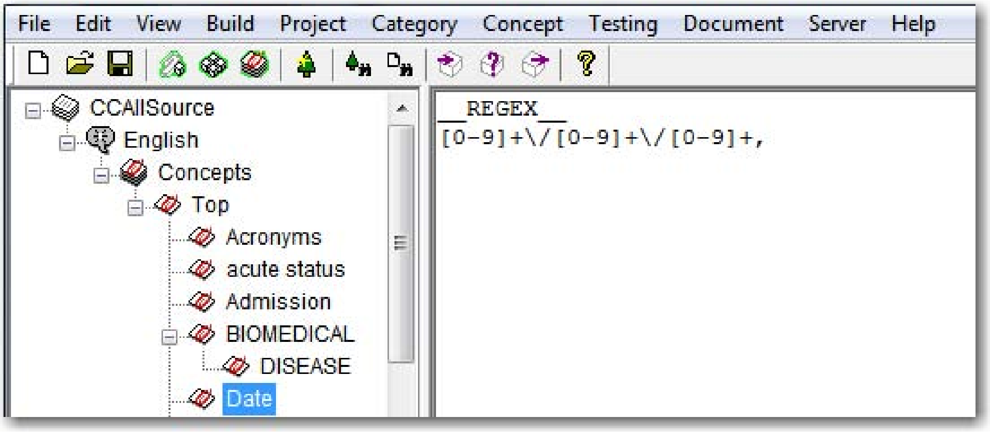
Category	Relevancy
Top/Diseases/Bacterial Infections and Mycoses/Infectio…	1.25
Top/Diseases/Respiratory Tract Diseases/Respiratory T…	1.25

Figure 2: Example of Text Analytics Categorization

Figure 2 provides an example of how various diseases and other medical conditions can be defined. Many medical taxonomies, such as the Medical Subject Headings (MeSH) (listed in the U.S. National Library of Medicine), are automatically updated by standards bodies, and are available for use in our applications. This ensures that we have the most recent means of identifying new drugs and compounds in the market.

Concept Identification and Conceptual Extraction

Figure 3 provides an example of how a regular expression can be used to identify a date field in the form of mm/dd/yy. As shown here, the SAS Text Analytics tools use the now-standard Perl regular expression syntax.

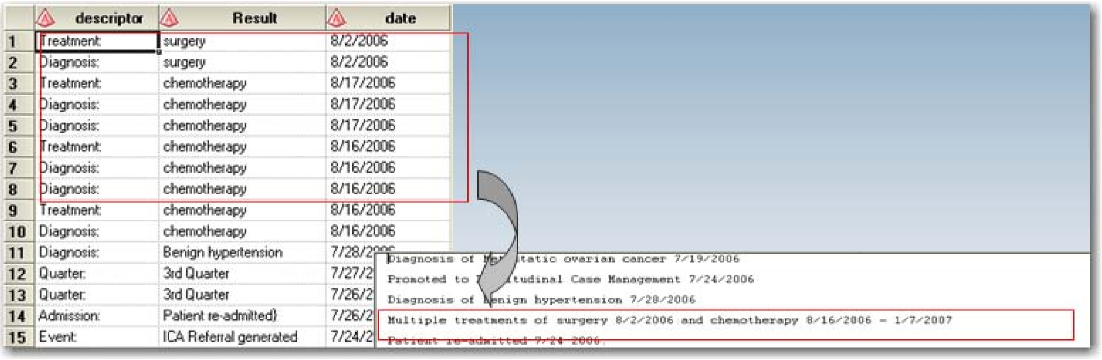


```
__REGEX__
[0-9]+\/[0-9]+\/[0-9]+,
```

Figure 3: Example Date Definition (Using Standard Perl Regular Expressions)

Event-Object Summarization and Arithmetic

Once the incoming stream of text can be identified, the results can be placed in a data store that can then be manipulated using standard SAS products such as SAS® Enterprise Guide®.



	descriptor	Result	date
1	Treatment:	surgery	8/2/2006
2	Diagnosis:	surgery	8/2/2006
3	Treatment:	chemotherapy	8/17/2006
4	Diagnosis:	chemotherapy	8/17/2006
5	Diagnosis:	chemotherapy	8/17/2006
6	Treatment:	chemotherapy	8/16/2006
7	Diagnosis:	chemotherapy	8/16/2006
8	Diagnosis:	chemotherapy	8/16/2006
9	Treatment:	chemotherapy	8/16/2006
10	Diagnosis:	chemotherapy	8/16/2006
11	Diagnosis:	Benign hypertension	7/28/20
12	Quarter:	3rd Quarter	7/27/2
13	Quarter:	3rd Quarter	7/26/2
14	Admission:	Patient re-admitted)	7/26/2
15	Event:	ICA Referral generated	7/24/2

```
Diagnosis of Metastatic ovarian cancer 7/19/2006
Promoted to Longitudinal Case Management 7/24/2006
Diagnosis of Benign hypertension 7/28/2006
Multiple treatments of surgery 8/2/2006 and chemotherapy 8/16/2006 - 1/7/2007
Patient re-admitted 7/24 2006
```

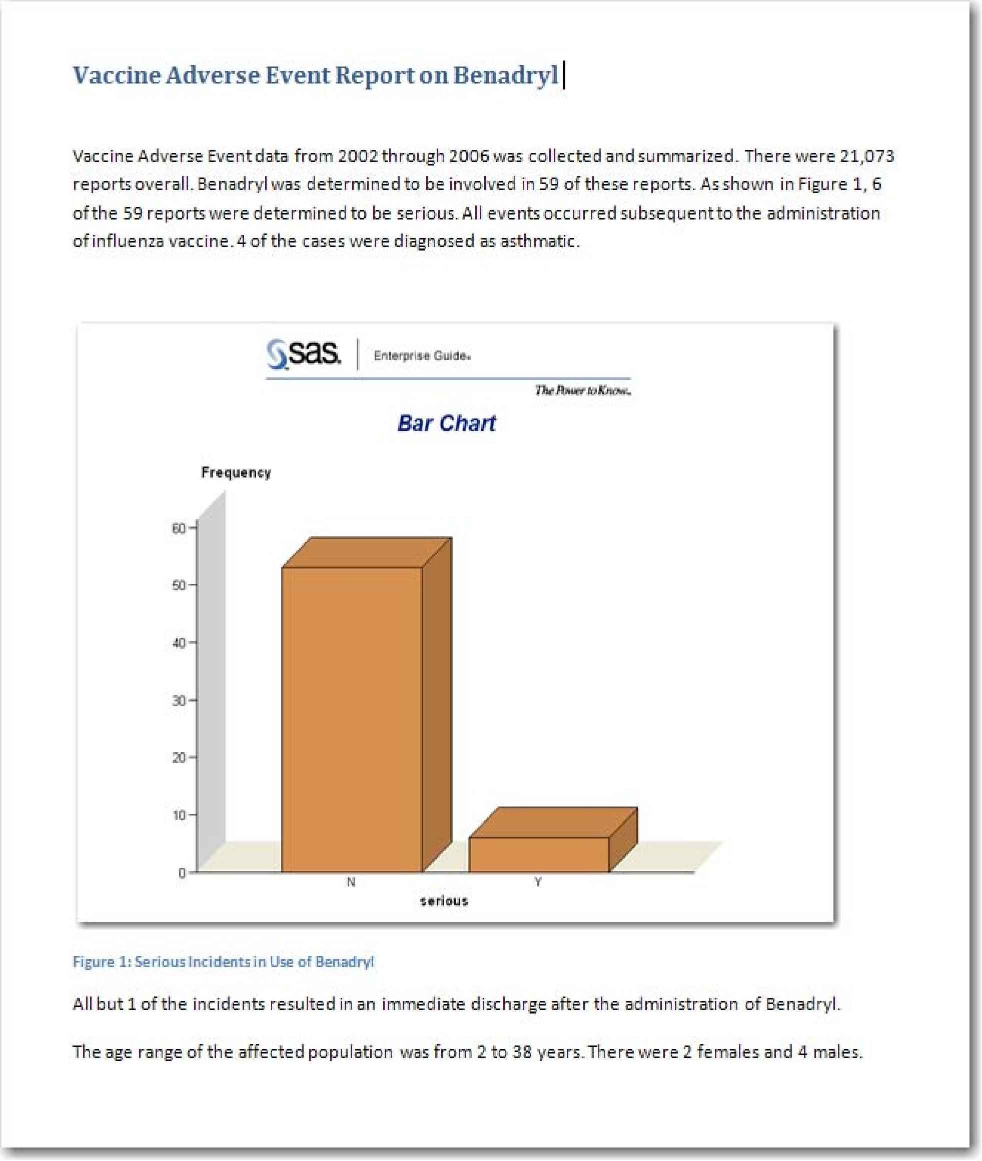
Figure 4: Illustration of the Relationship between Extracted Fields of Data and Associated Report Fragment

Figure 4 shows the relationship between fields of data and the resulting text summary. Here, numerous “chemotherapy” diagnoses are detected (along with the associated dates), so the text summary reads:

“Multiple treatments of … chemotherapy between 8/2/2006 and … “. (The example shown here also includes summaries for surgeries and the associated date ranges.)

EXAMPLE OF ELEMENTARY OPERATORS

The three elementary operations used with in-patient records are categories, concepts, and event-object summarization and arithmetic. The operations are readily generalizable to other collections of data that have either structure fields or which have fields of data that can be re-formed in a structured fashion. An example is shown in Figure 5. Each element of this report – for example, event outcome – is either a field in the analysis or is an object that can be detected with linguistic scanning.



Vaccine Adverse Event Report on Benadryl

Vaccine Adverse Event data from 2002 through 2006 was collected and summarized. There were 21,073 reports overall. Benadryl was determined to be involved in 59 of these reports. As shown in Figure 1, 6 of the 59 reports were determined to be serious. All events occurred subsequent to the administration of influenza vaccine. 4 of the cases were diagnosed as asthmatic.

Figure 1: Serious Incidents in Use of Benadryl

All but 1 of the incidents resulted in an immediate discharge after the administration of Benadryl.

The age range of the affected population was from 2 to 38 years. There were 2 females and 4 males.

Figure 5: Example Report Applying Approach to VAERS Data

The associated Enterprise Guide project that produces this result is illustrated in Figure 6.

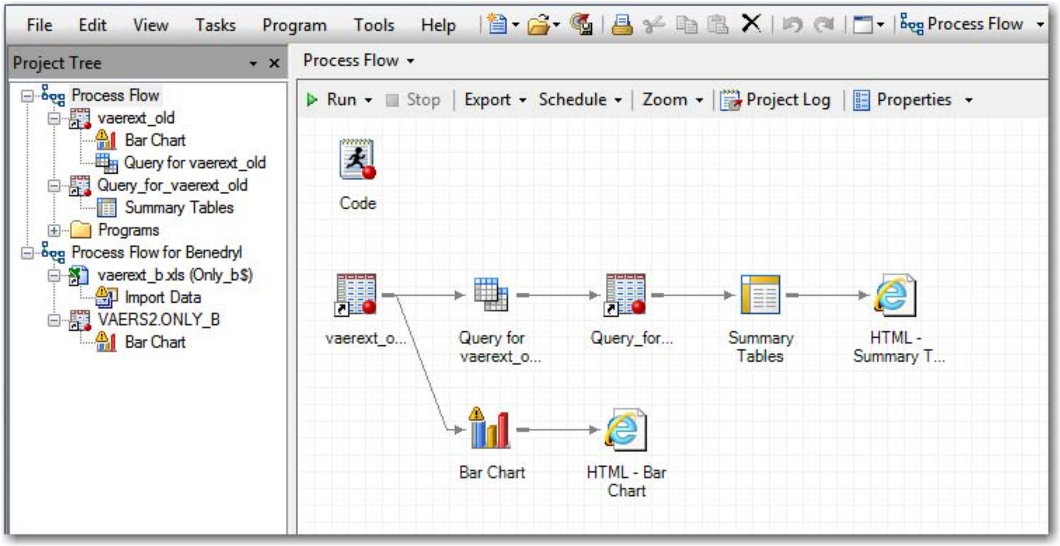


Figure 6: SAS Enterprise Guide Process Flow Diagram for VAERS Report

All results were produced using one of the three elementary operations:

1. Category identification
2. Concept recognition
3. Event-object summarization and arithmetic.
4. The other significant ingredient to this report is the use of SAS ODS output.

The ODS facility has the flexibility to produce a variety of output formats. In this case we are using the Microsoft Word display format as shown above.

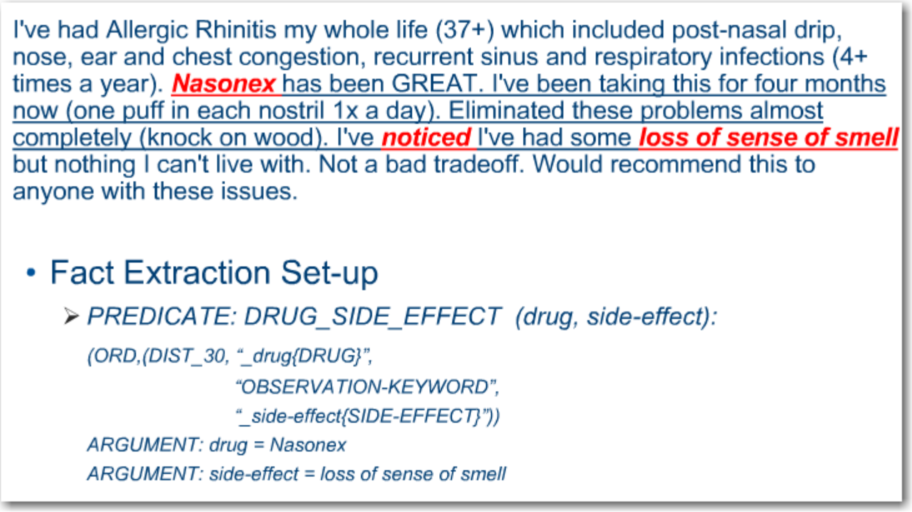
OTHER ELEMENTARY OPERATIONS

As shown above, event detection and object recognition are key features to the successful quantitative summarization of descriptive records.

Other required operations, illustrated below, include fact extraction, sentiment identification and, finally, conditional inference.

Fact Extraction

SAS Enterprise Content Categorization is capable of extracting specific information in context (facts), as shown in Figure 7. Here we detect a ***loss of sense of smell*** that is associated with the use of the drug.



I've had Allergic Rhinitis my whole life (37+) which included post-nasal drip, nose, ear and chest congestion, recurrent sinus and respiratory infections (4+ times a year). ***Nasonex*** has been GREAT. I've been taking this for four months now (one puff in each nostril 1x a day). Eliminated these problems almost completely (knock on wood). I've ***noticed*** I've had some ***loss of sense of smell*** but nothing I can't live with. Not a bad tradeoff. Would recommend this to anyone with these issues.

- Fact Extraction Set-up
  - *PREDICATE: DRUG_SIDE_EFFECT (drug, side-effect):*

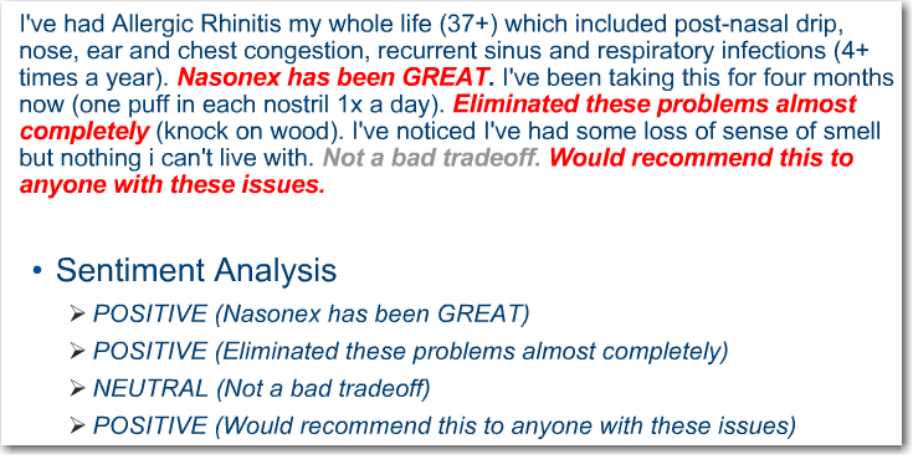
(ORD,(DIST_30, "_drug{DRUG}",
"OBSERVATION-KEYWORD",
"_side-effect{SIDE-EFFECT}"))
ARGUMENT: drug = Nasonex
ARGUMENT: side-effect = loss of sense of smell

Figure 7: Example Fact Extraction (Information in Context)

Sentiment Extraction

Figure 8 illustrates the extraction of sentiment in cases where named drug brands are present. As shown in the example, we have an overall positive assessment of the drug (Nasonex has been GREAT). SAS®Sentiment Analysis is included in the suite of SAS Text Analytics products and is specifically equipped to:

- identify various attributes of assessment that are associated with text targets
- assign a positive, negative, or neutral emotional charge associated with the specific attribute
- create an overall sentiment score that depends on the weight of the attributes and associated emotional charge



I've had Allergic Rhinitis my whole life (37+) which included post-nasal drip, nose, ear and chest congestion, recurrent sinus and respiratory infections (4+ times a year). ***Nasonex has been GREAT.*** I've been taking this for four months now (one puff in each nostril 1x a day). ***Eliminated these problems almost completely*** (knock on wood). I've noticed I've had some loss of sense of smell but nothing i can't live with. ***Not a bad tradeoff.*** ***Would recommend this to anyone with these issues.***

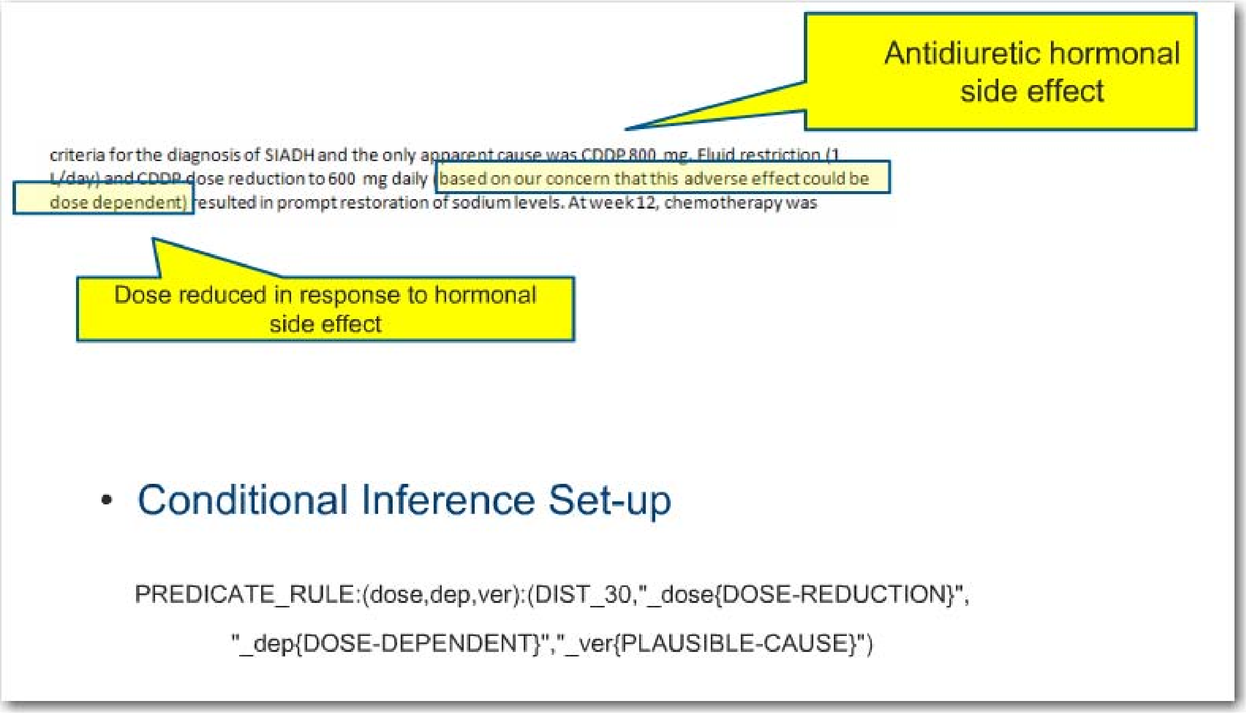
- Sentiment Analysis
  - *POSITIVE (Nasonex has been GREAT)*
  - *POSITIVE (Eliminated these problems almost completely)*
  - *NEUTRAL (Not a bad tradeoff)*
  - *POSITIVE (Would recommend this to anyone with these issues)*

Figure 8: Example of Sentiment Extraction

Further information about the Sentiment Analysis capability is reported in (Albright and Lakkaraju, 2010).

Conditional Inference

Conditional inference is based on a standard predicate rule structure which, in turn, is based on first-order logic (Hodges, 2001). This logic enables us to make both formal and informal inferences based on logical relationships.



Antidiuretic hormonal side effect

criteria for the diagnosis of SIADH and the only apparent cause was CDDP 800 mg. Fluid restriction (1 L/day) and CDDP dose reduction to 600 mg daily (based on our concern that this adverse effect could be dose dependent) resulted in prompt restoration of sodium levels. At week 12, chemotherapy was

Dose reduced in response to hormonal side effect

- Conditional Inference Set-up

```
PREDICATE_RULE:(dose,dep,ver):(DIST_30,"_dose{DOSE-REDUCTION}",
        "_dep{DOSE-DEPENDENT}","_ver{PLAUSIBLE-CAUSE}")
```

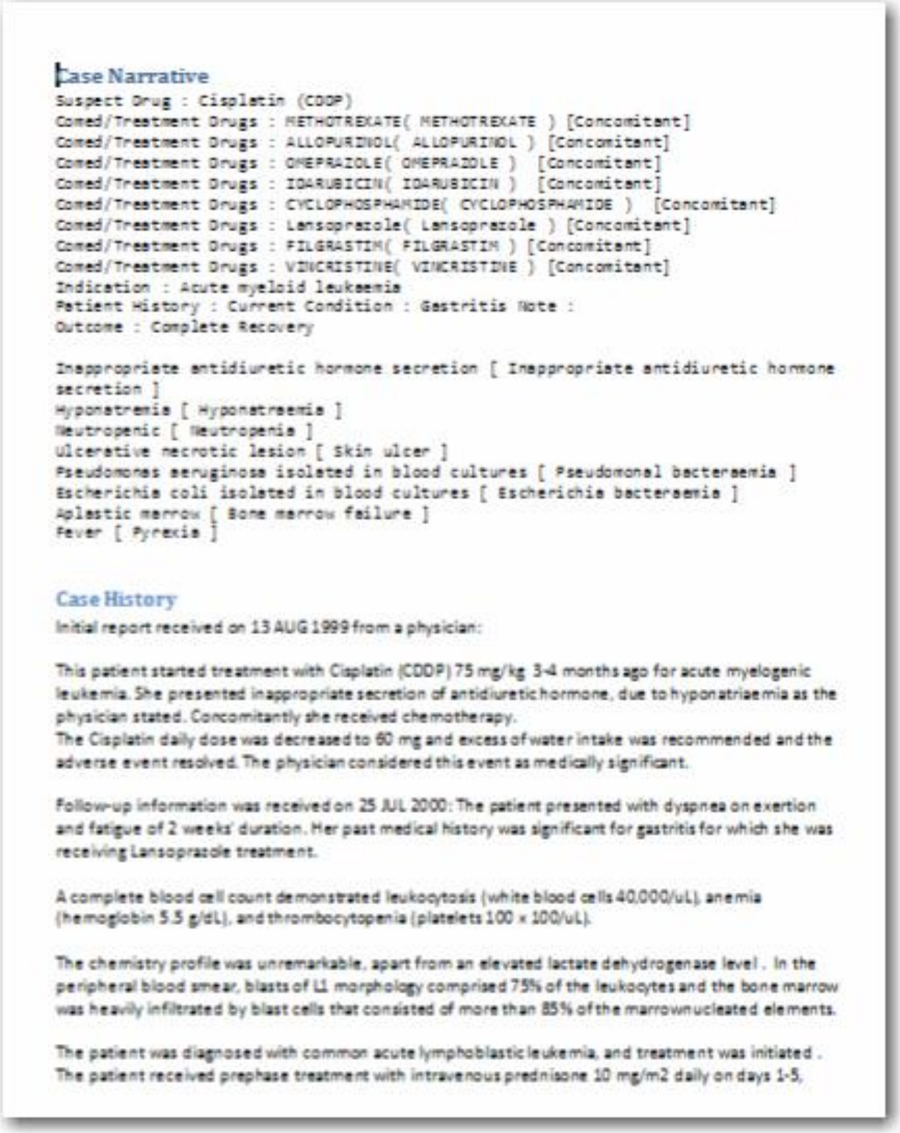
Figure 9: Example of Conditional Inference

In the example shown in Figure 9, we see that the detection syntax can be set up to look for instances of “dose reductions” and “dose-dependent side effects.” Taken in the context of other information, such as disease condition or brand name, this enables us to populate the record with the associated field.

Deployment

These forms of elementary operations that we have reviewed here are useful for automating a wide range of medical scenarios. Figure 10 (expanded in Figures 11 and 12) shows a typical pro-forma summary report that employs the various mechanisms described above.

Textually Dense Pharmaceutical Regulatory Conformance (continued)



Case Narrative

```
Suspect Drug : Cisplatin (CDDP)
Comed/Treatment Drugs : METHOTREXATE( METHOTREXATE ) [Concomitant]
Comed/Treatment Drugs : ALLOPURINOL( ALLOPURINOL ) [Concomitant]
Comed/Treatment Drugs : OMEPRAZOLE( OMEPRAZOLE )  [Concomitant]
Comed/Treatment Drugs : IDARUBICIN( IDARUBICIN )  [Concomitant]
Comed/Treatment Drugs : CYCLOPHOSPHAMIDE( CYCLOPHOSPHAMIDE )  [Concomitant]
Comed/Treatment Drugs : Lansoprazole( Lansoprazole ) [Concomitant]
Comed/Treatment Drugs : FILGRASTIM( FILGRASTIM ) [Concomitant]
Comed/Treatment Drugs : VINCRISTINE( VINCRISTINE ) [Concomitant]
Indication : Acute myeloid leukaemia
Patient History : Current Condition : Gastritis Note :
Outcome : Complete Recovery

Inappropriate antidiuretic hormone secretion [ Inappropriate antidiuretic hormone
secretion ]
Hyponatremia [ Hyponatraemia ]
Neutropenic [ Neutropenia ]
Ulcerative necrotic lesion [ Skin ulcer ]
Pseudomonas aeruginosa isolated in blood cultures [ Pseudomonal bacteraemia ]
Escherichia coli isolated in blood cultures [ Escherichia bacteraemia ]
Aplastic marrow [ Bone marrow failure ]
Fever [ Pyrexia ]
```

Case History

Initial report received on 13 AUG 1999 from a physician:

This patient started treatment with Cisplatin (CDDP) 75 mg/kg 3-4 months ago for acute myelogenic leukemia. She presented inappropriate secretion of antidiuretic hormone, due to hyponatriaemia as the physician stated. Concomitantly she received chemotherapy.
The Cisplatin daily dose was decreased to 60 mg and excess of water intake was recommended and the adverse event resolved. The physician considered this event as medically significant.

Follow-up information was received on 25 JUL 2000: The patient presented with dyspnea on exertion and fatigue of 2 weeks' duration. Her past medical history was significant for gastritis for which she was receiving Lansoprazole treatment.

A complete blood cell count demonstrated leukocytosis (white blood cells 40,000/uL), anemia (hemoglobin 5.5 g/dL), and thrombocytopenia (platelets 100 × 100/uL).

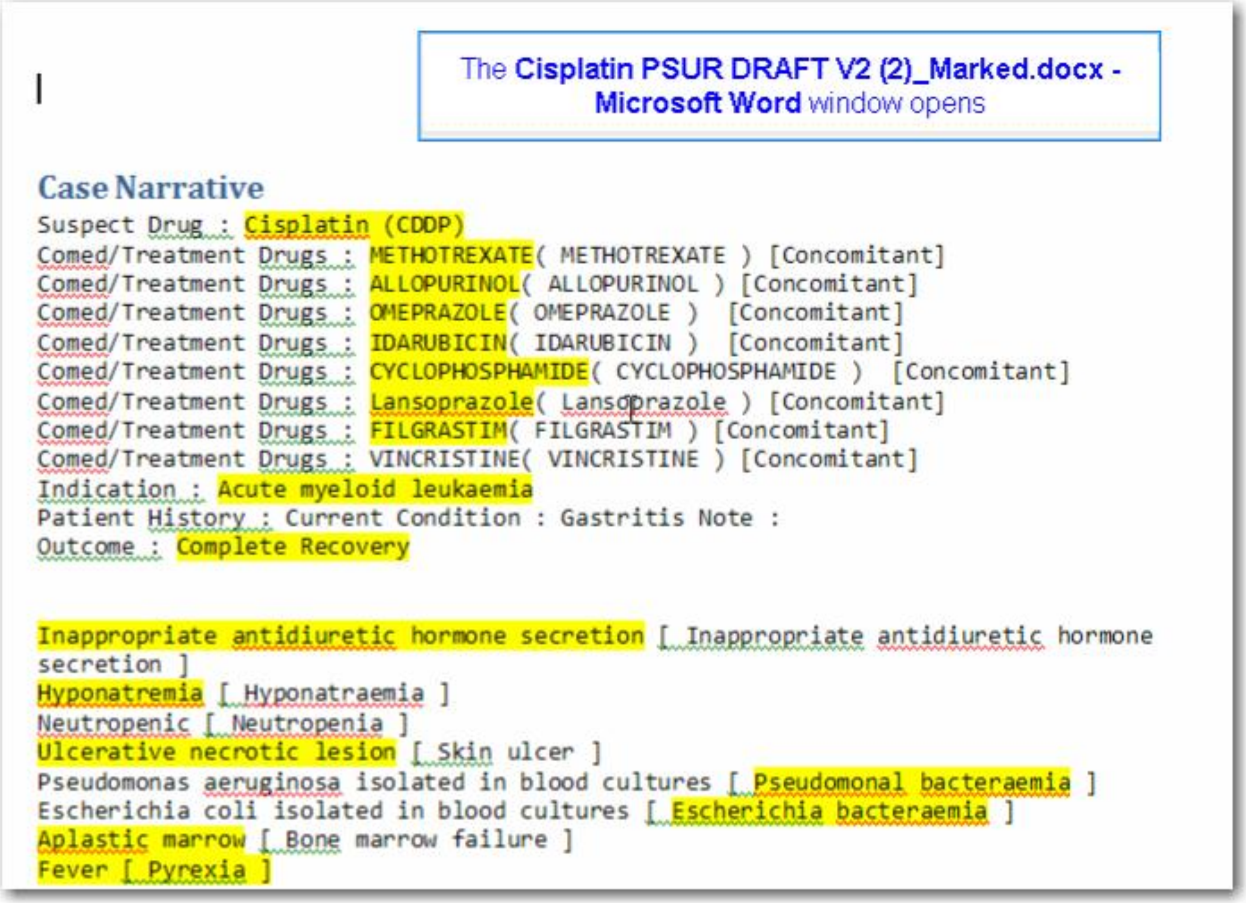
The chemistry profile was unremarkable, apart from an elevated lactate dehydrogenase level . In the peripheral blood smear, blasts of L1 morphology comprised 75% of the leukocytes and the bone marrow was heavily infiltrated by blast cells that consisted of more than 85% of the marrownucleated elements.

The patient was diagnosed with common acute lymphoblastic leukemia, and treatment was initiated .
The patient received prephase treatment with intravenous prednisone 10 mg/m2 daily on days 1-5,

Figure 10: Typical Pro-Forma Report Output

Typically, a Case Narrative presents a resume of the case. The information is typically presented in 2 - 3 paragraphs (often with an associated numerical table or Microsoft Excel worksheet). This is information that is otherwise normally contained in several pages of text.

In order to produce a good summary of this example case narrative, the software has to identify the terms that are highlighted in Figure 11. These include chemical compositions, drug names, diagnoses, as well as potential side effects. The technical components of these recognition requirements are readily identified using taxonomies such as the MeSH taxonomy referred to earlier.



The **Cisplatin PSUR DRAFT V2 (2)_Marked.docx - Microsoft Word** window opens

Case Narrative

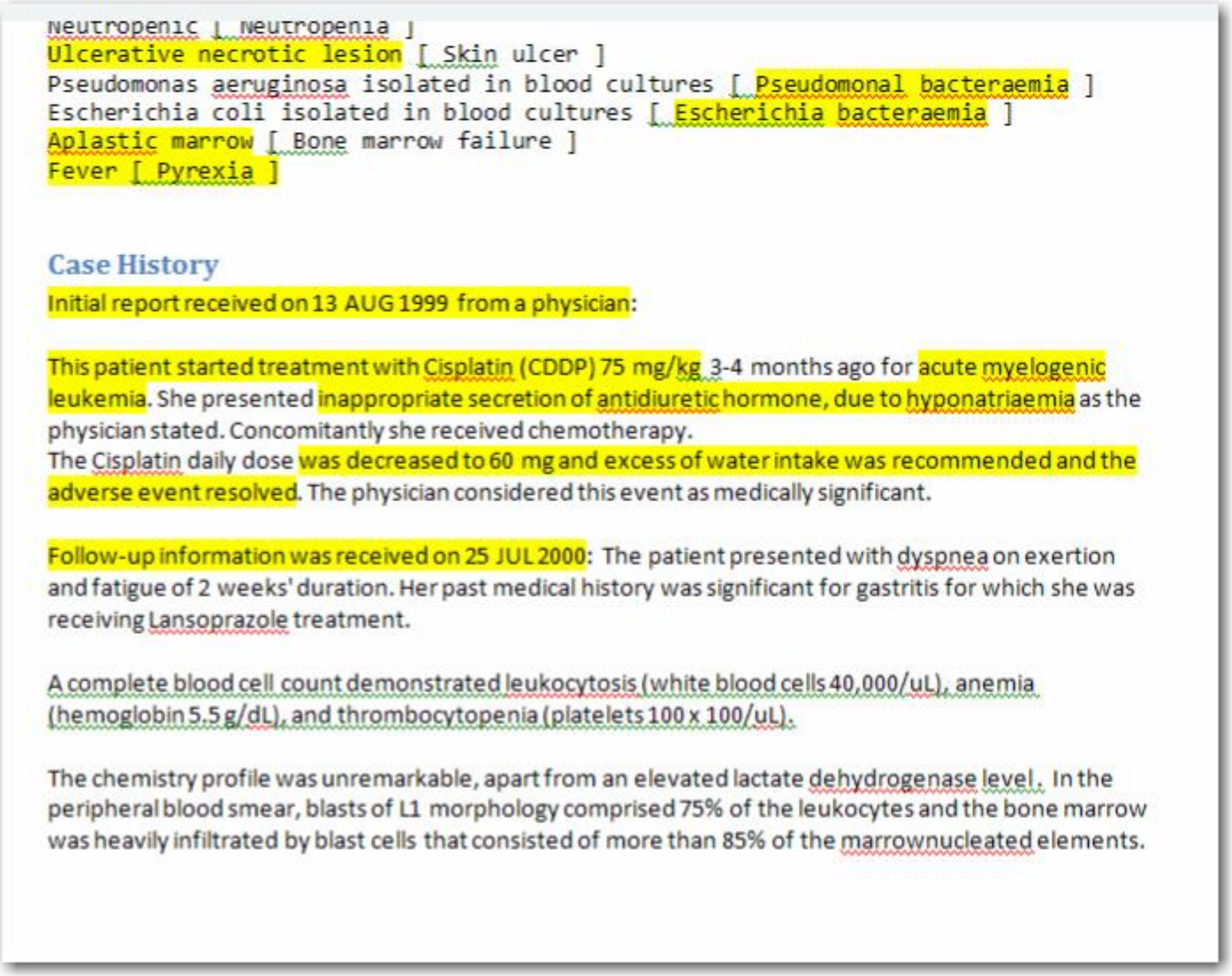
```
Suspect Drug : Cisplatin (CDDP)
Comed/Treatment Drugs : METHOTREXATE( METHOTREXATE ) [Concomitant]
Comed/Treatment Drugs : ALLOPURINOL( ALLOPURINOL ) [Concomitant]
Comed/Treatment Drugs : OMEPRAZOLE( OMEPRAZOLE )  [Concomitant]
Comed/Treatment Drugs : IDARUBICIN( IDARUBICIN )  [Concomitant]
Comed/Treatment Drugs : CYCLOPHOSPHAMIDE( CYCLOPHOSPHAMIDE )  [Concomitant]
Comed/Treatment Drugs : Lansoprazole( Lansoprazole ) [Concomitant]
Comed/Treatment Drugs : FILGRASTIM( FILGRASTIM ) [Concomitant]
Comed/Treatment Drugs : VINCRISTINE( VINCRISTINE ) [Concomitant]
Indication : Acute myeloid leukaemia
Patient History : Current Condition : Gastritis Note :
Outcome : Complete Recovery


Inappropriate antidiuretic hormone secretion [ Inappropriate antidiuretic hormone
secretion ]
Hyponatremia [ Hyponatraemia ]
Neutropenic [ Neutropenia ]
Ulcerative necrotic lesion [ Skin ulcer ]
Pseudomonas aeruginosa isolated in blood cultures [ Pseudomonal bacteraemia ]
Escherichia coli isolated in blood cultures [ Escherichia bacteraemia ]
Aplastic marrow [ Bone marrow failure ]
Fever [ Pyrexia ]
```

Figure 11: Example of Raw Input and Recognition Features

As shown above, the recognition features (shown in yellow) are examples of category, concept, or object-event recognition operations. It should be easy to see how these text strings are identified and echoed in the report production process.

As we continue on the examination of the case narrative, we can see other information – also highlighted – that must be captured: date, doses, and associated symptoms or side effects.



```
Neutropenic [ Neutropenia ]
Ulcerative necrotic lesion [ Skin ulcer ]
Pseudomonas aeruginosa isolated in blood cultures [ Pseudomonal bacteraemia ]
Escherichia coli isolated in blood cultures [ Escherichia bacteraemia ]
Aplastic marrow [ Bone marrow failure ]
Fever [ Pyrexia ]
```

Case History

Initial report received on 13 AUG 1999 from a physician:

This patient started treatment with Cisplatin (CDDP) 75 mg/kg 3-4 months ago for acute myelogenic leukemia. She presented inappropriate secretion of antidiuretic hormone, due to hyponatriaemia as the physician stated. Concomitantly she received chemotherapy.
The Cisplatin daily dose was decreased to 60 mg and excess of water intake was recommended and the adverse event resolved. The physician considered this event as medically significant.

Follow-up information was received on 25 JUL 2000: The patient presented with dyspnea on exertion and fatigue of 2 weeks' duration. Her past medical history was significant for gastritis for which she was receiving Lansoprazole treatment.

A complete blood cell count demonstrated leukocytosis (white blood cells 40,000/uL), anemia (hemoglobin 5.5 g/dL), and thrombocytopenia (platelets 100 x 100/uL).

The chemistry profile was unremarkable, apart from an elevated lactate dehydrogenase level. In the peripheral blood smear, blasts of L1 morphology comprised 75% of the leukocytes and the bone marrow was heavily infiltrated by blast cells that consisted of more than 85% of the marrownucleated elements.

Figure 12: Results of Conditional Inference

In Figure 12 we also see information related to a decrease in dose (“Cisplatin … decreased to 60 mg”) that can be picked up through the operation of fact extraction. In this case “**Cisplatin**” (drug); “**60 mg**” (dose); “**decreased**” (decrease); and “**adverse event resolved**” (result) are all easily-recognized sentence-level events and attributes. These can be readily recognized and easily-mapped for post-processing and display.

We should resist the temptation to view the results presented here as fabricated; i.e. too simplistic for the “real world”. In fact, the authors who write these summaries are trained to create brief, to-the-point observations such as we see here. So the “translation” to computer-mediated reports is actually much easier as a result of this training and practice.

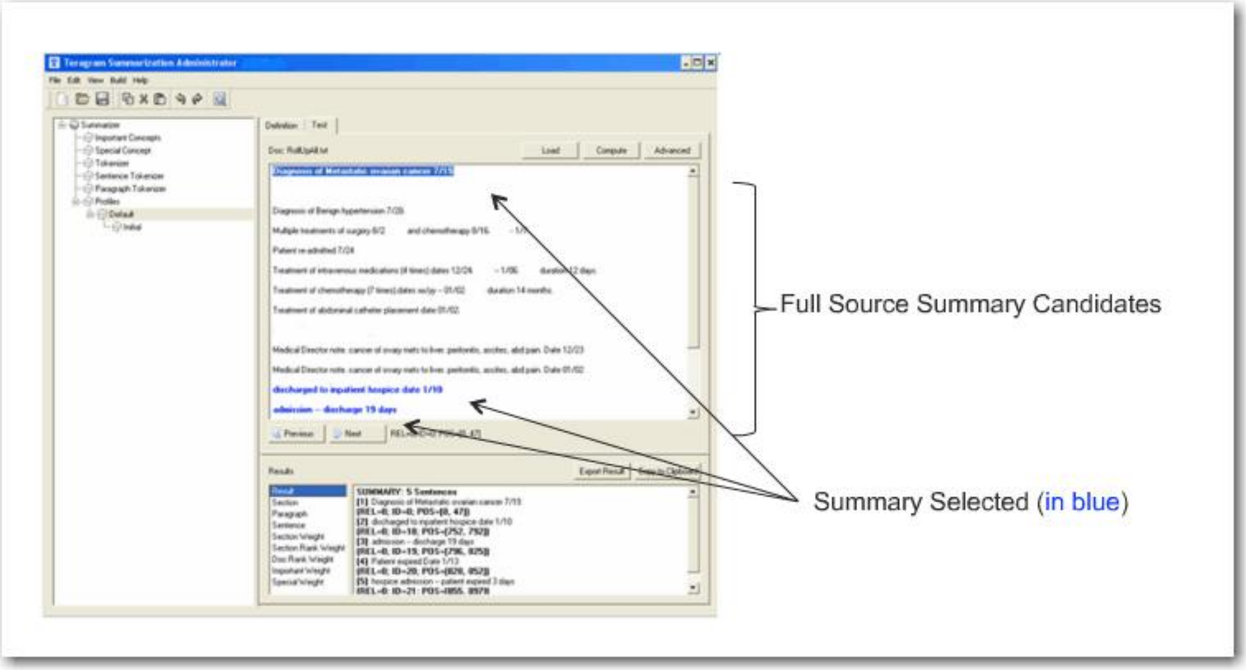
SUMMARIZATION

The extracted information and post-processed information can be combined together and put into a summarization task that can be managed by Text Summarization add-in (TSA) to SAS Enterprise Content Categorization.

A typical output is to prepare a case summary displaying in 2 or 3 paragraphs and sometimes displaying numerically, information that would otherwise normally take many pages of textual, tabular, and graphic output.

Once information is captured as a data table, it can be post-processed so that, in this case shown earlier in Figure 4, we can add the number of chemotherapy instances that have been identified. This can be used to generate a textual description that takes the number of summed chemotherapy instances as one of its inputs. The date range can also be collected and this, too, can be echoed in the textual description that is produced.

The extracted information and the post processed information can be combined together and can be put into a summarization task that can be used to produce a pro-forma summary of the information that has been captured for presentation.



Doc: RollUp48.txt

Diagnosis of Metastatic ovarian cancer 7/15

Diagnosis of Benign hypertension 7/26

Multiple treatments of surgery 8/2 and chemotherapy 8/16 - 1/7

Patient re-admitted 7/24

Treatment of intravenous medications (4 times) dates 12/24 - 1/06 duration 12 days

Treatment of chemotherapy (7 times) dates xx/yy - 01/02 duration 14 months

Treatment of abdominal catheter placement date 01/02

Medical Director note: cancer of ovary mets to liver, peritonitis, ascites, abd pain. Date 12/23

Medical Director note: cancer of ovary mets to liver, peritonitis, ascites, abd pain. Date 01/02

discharged to inpatient hospice date 1/10

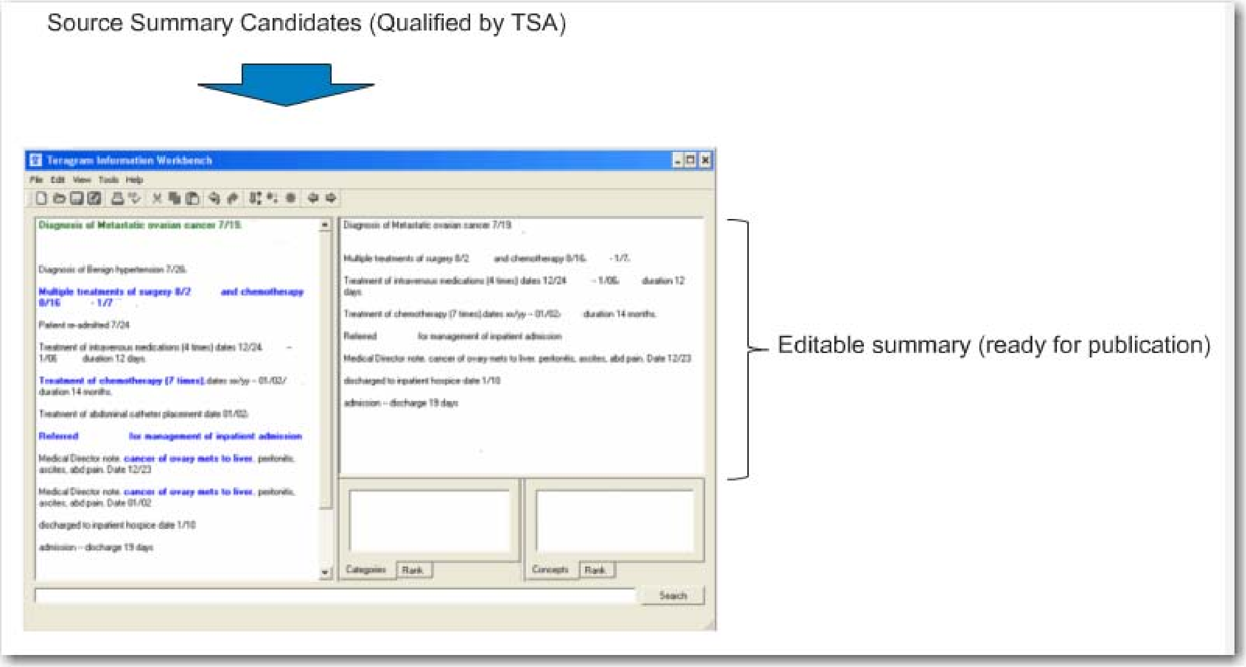
admission -- discharge 19 days

REL=0; ID=0; POS=(0, 47)

```
SUMMARY: 5 Sentences
[1] Diagnosis of Metastatic ovarian cancer 7/15
(REL=0; ID=0; POS=(0, 47))
[2] dischaged to inpatient hospice date 1/10
(REL=0; ID=18; POS=(752, 792))
[3] admission -- discharge 19 days
(REL=0; ID=19; POS=(796, 825))
[4] Patient expired Date 1/13
(REL=0; ID=20; POS=(828, 852))
[5] hospice admission -- patient expired 3 days
(REL=0; ID=21; POS=(855, 897))
```

Figure 13: Production of the Pro-Forma Summary

Finally, as illustrated in Figure 14, the pro-forma summary can be edited in a word-like document editor to produce a final review version of the desired presentation document.


Source Summary Candidates (Qualified by TSA)
Teragram Information Workbench
File Edit View Tools Help
Categories
Rank
Concepts
Rank
Search
Editable summary (ready for publication)

Diagnosis of Metastatic ovarian cancer 7/15

Diagnosis of Benign hypertension 7/26

Multiple treatments of surgery 8/2 and chemotherapy 8/16 - 1/7

Patient re-admitted 7/24

Treatment of intravenous medications (4 times) dates 12/24 - 1/06 duration 12 days

Treatment of chemotherapy (7 times), dates xx/yy - 01/02/ duration 14 months

Treatment of abdominal catheter placement date 01/02

Referred for management of inpatient admission

Medical Director note: cancer of ovary mets to liver, peritonitis, ascites, abd pain. Date 12/23

Medical Director note: cancer of ovary mets to liver, peritonitis, ascites, abd pain. Date 01/02

discharged to inpatient hospice date 1/10

admission -- discharge 19 days

Diagnosis of Metastatic ovarian cancer 7/19

Multiple treatments of surgery 8/2 and chemotherapy 8/16 - 1/7

Treatment of intravenous medications (4 times) dates 12/24 - 1/06 duration 12 days

Treatment of chemotherapy (7 times) dates xx/yy - 01/02 duration 14 months

Referred for management of inpatient admission

Medical Director note: cancer of ovary mets to liver, peritonitis, ascites, abd pain. Date 12/23

discharged to inpatient hospice date 1/10

admission -- discharge 19 days

Figure 14: Edited Summary Results

CONCLUSION

Here we have demonstrated six basic operations that provide a blend of linguistic and quantitative text analytic processing – together with associated report production capabilities that are suitable for the construction of an automated report production process in the preparation of PSUR reports.

These operations are:

- categorization
- conceptual Extraction
- event-object detection and quantitative summarization and arithmetic
- conditional inference
- fact extraction
- sentiment extraction

Our business analysis has demonstrated significant economies in the deployment of such a system. As shown below in Figure 15 in a provisional system architecture diagram all results are subject to human review.

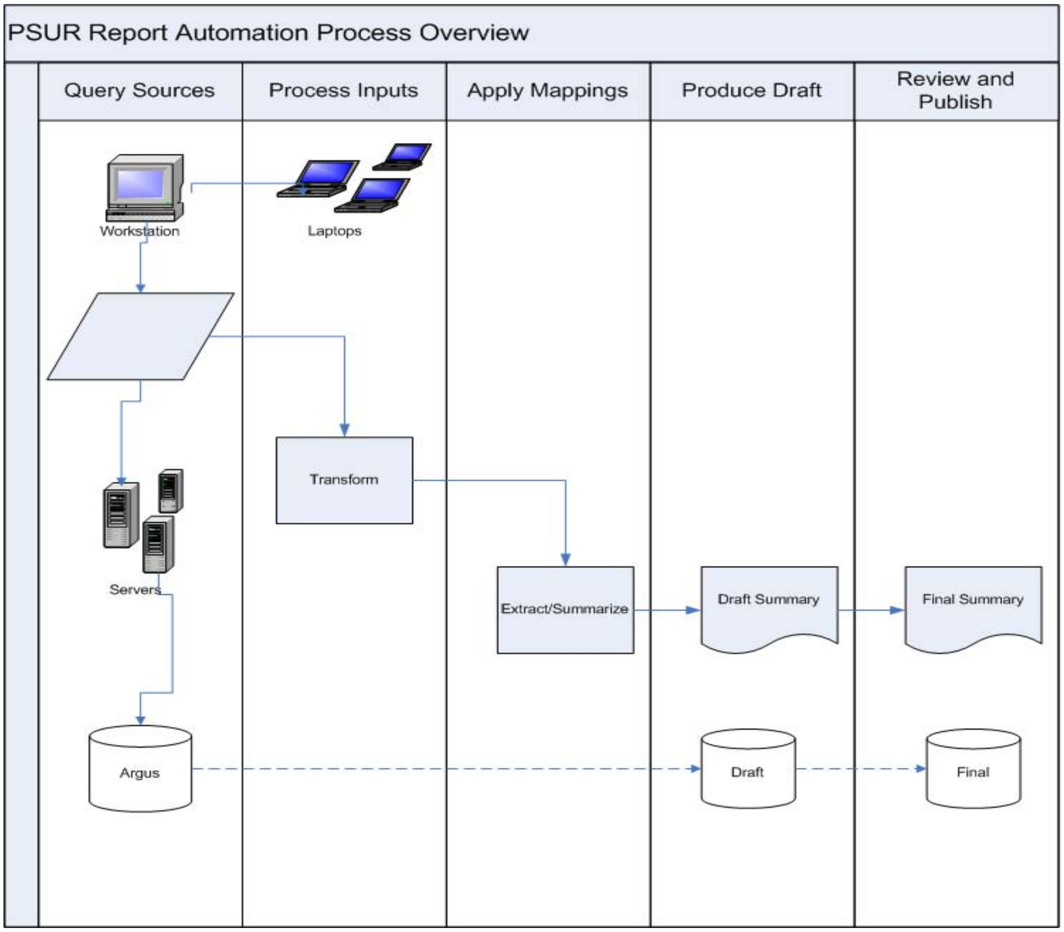


Figure 15: Provisional System Architecture

There are significant benefits in the development of a central repository of documents. Even greater economies – difficult to quantify – are delivered by virtue of the important quality control function that is provided by computation.

REFERENCES

SAS Institute Inc. 2010. SAS Institute white paper. “Combining Knowledge and Data Mining to Understand Sentiment – A Practical Assessment of Approaches.” http://www.sas.com/reg/wp/corp/2799.

U.S. National Library of Medicine, Medial Subject Headings, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. Available at http://www.nlm.nih.gov/mesh/

Wilfrid Hodges, "Classical Logic I: First Order Logic," Lou Goble, ed., **THE BLACKWELL GUIDE TO PHILOSOPHICAL LOGIC**. Blackwell, 2001.

ACKNOWLEDGMENTS

None of this work would have been possible without the help of many others, primarily:

Saratendu Sethi, Dr. Edwin Schaart, Phil DiMassimo

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Barry de Ville
SAS Campus Drive
SAS Institute Inc.
E-mail: Barry.deVille@sas.com

Mark Wolff
SAS Campus Drive
SAS Institute Inc.
E-mail: Mark.Wolff@sas.com