AMGEN

TMS

OCUG 7th Annual Meeting
25 September 2002

Uses of Stemmer Algorithms, Substitutions, and interMedia in TMS Search Object Design

AUTO ENCODER

Sunil G. Singh
DBMS Consulting

Donna Caruso
Amgen, Inc.

Acknowledgements

- Thanks to the OCUG for the opportunity to present.
- Thanks to Andy Alasso, Kim Renjdrup & Dominique Farinaux-Dumas for their technical insights into the content of this presentation.
- Special thanks to Amgen, Inc. for their sponsorship as this functionality was utilized to develop the Amgen Auto-Encoder for TMS.

AMGEN

Goals

- Gain an understanding of the tools used in search object design.
- Review research on stemming algorithms’ performance in information retrieval.
- Amgen’s Case Study for application of the tools
- Concluding observations based on research & practical application within the TMS environment.

Definitions

- TMS Search Objects:
  - Procedures containing algorithms for searching TMS dictionaries
  - Integrated with TMS through search object definition
  - Executed from TMS API calls
- Information retrieval in the context of TMS search objects:
  - The ability to retrieve & match verbatim terms (VTs) to dictionary terms by using search algorithms.

AMGEN

Definitions (2)

- Retrieval tools used in search algorithms:
  - Stemmer Algorithms:
    - Porter Stemmer
    - Oracle interMedia (Xerox Corporation’s iMT stemmer)
  - Substitutions:
    - Full words
    - Partial words
- Candidate Terms
  - List of dictionary terms retrieved in the search algorithm that are suggested dictionary matches used in manual classification.

Definitions (3)

- Morphological variants (word variations)
  - Unrecognizable in exact term-matching algorithms (cramp, cramps, cramping).
  - Similar semantic interpretations and can be treated as equivalents in information retrieval (cramps, cramping -> cramp).

Why Use Stemmers?

- Stemmers have been created for information retrieval to reduce terms to their root form for improved recognition by term-matching procedures.

Unstemmed Word	Stem
Blurry	Blur
Blurred	Blur
Blurring	Blur

Stemmer Scope

1. Traditional approach based on suffix removal:
   - Focus on the Porter Stemmer
2. Linguistic methods based on the Xerox Stemmer
   - Focus on Oracle interMedia using default English lexer (lexicon)
     - Search & retrieval capability for text
     - Concept searching
     - Theme analysis

Porter Stemmer

- The Porter stemming algorithm is a process for removing morphological variants & inflexional endings (suffixes) from words in English.
- It is mainly used as part of a term normalization process during information retrieval.

Xerox Stemmer

- Xerox's English lexical database can linguistically identify 77,000 base forms of 500,000 variant words with the following morphological tools:
  - Inflectional stemmer
  - Derivational stemmer

Xerox Stemmer (2)

- Inflectional Stemmer:
  - Identifies changes in word form due to case, gender, number, tense, person, mood, voice.
    - Nouns: children -> child
    - Verbs: understood -> understand
    - Adjectives: best -> good
    - Pronouns: whom -> who

Xerox Stemmer (3)

- Derivational Stemmer:
  - Reduces variant words to their derived form using suffix and prefix removal
  - Must preserve original meaning

Stemmer Analysis

- Impacts of Stemming:
  - Only a small improvement to retrieval performance
  - Although it does not hurt retrieval performance
- Traditional approach & linguistic methods perform equally as well.

Stemmer Analysis (2)

- Down side to suffix removal stemmer:
  - Lumps “general, generous, generation, generic” into “gener” root.
  - Does not find a root for “recognize, recognition”.
  - Creates roots that are not actual words making it difficult for dictionary information retrieval “genetic, genetically, geneticist, genetics” into “genet” root.

Research[1] Observations

- Some form of Stemming is beneficial; the average absolute improvement due to stemming ranges from 1-3%.
- Plural removal is very effective with small queries.
- No difference in average performance of Stemmers.
- Rules based suffix removal is beneficial is some cases, but not ideal in all cases.

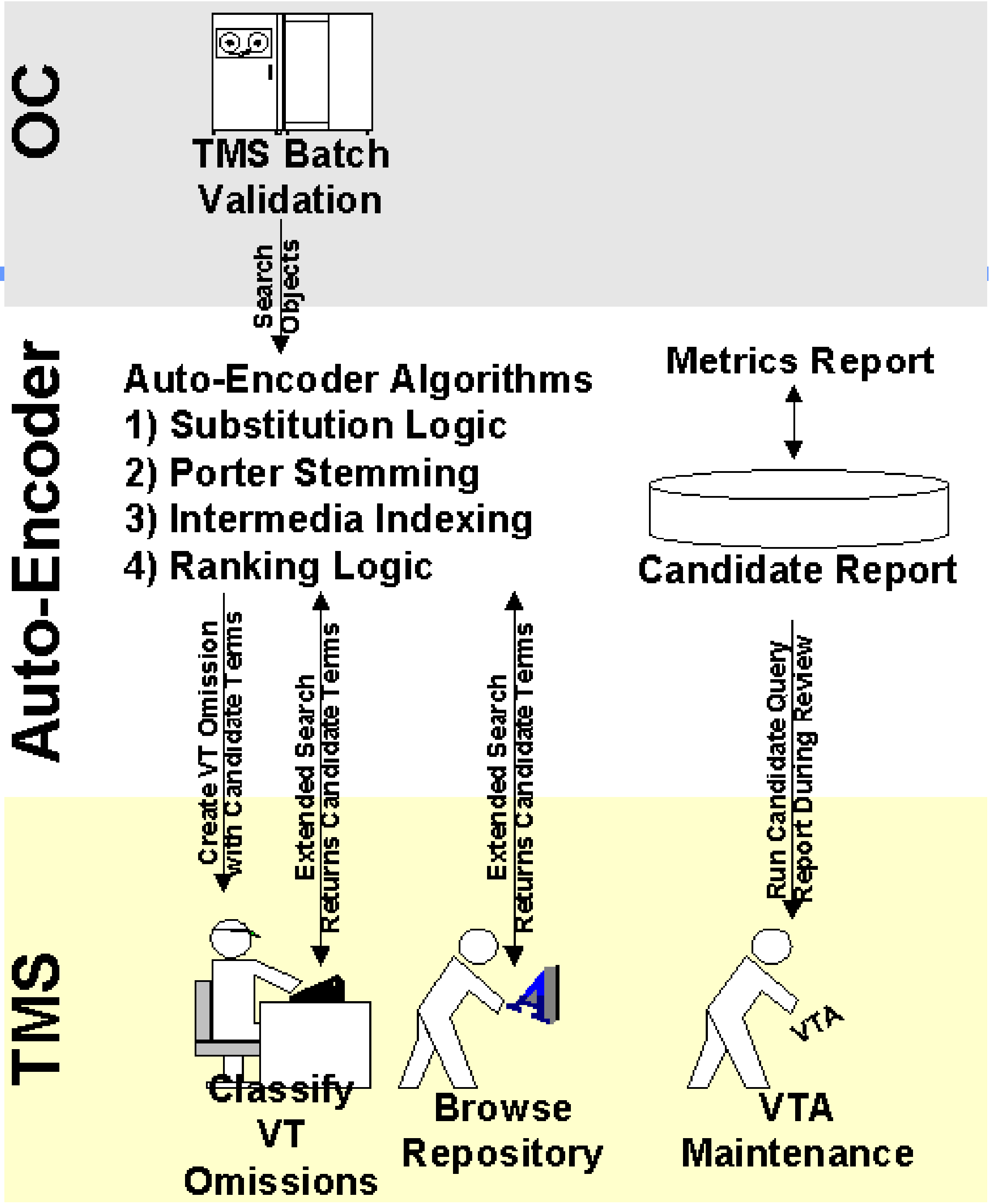
[1] Researchers from Rank Xerox Research Centre, France used the SMART text retrieval system developed at Cornell University to examine the performance of 5 different stemming algorithms.

Research Observations (2)

- Linguistic methods are limited based on the content of the lexicon; unable to correct stem words which are not contained in the lexicon.
- Linguistic root words are not always optimal for information retrieval.
  - “English” based lexicon is most effective for “English” words and their definitions.

Amgen Case Study

- VTO Creation
- Coding Workflow
- Review Workflow



OC

TMS Batch Validation

Search Objects

Auto-Encoder

Auto-Encoder Algorithms
1) Substitution Logic
2) Porter Stemming
3) Intermedia Indexing
4) Ranking Logic

Metrics Report

Candidate Report

Create VT Omission with Candidate Terms

Extended Search Returns Candidate Terms

Extended Search Returns Candidate Terms

Run Candidate Query Report During Review

TMS

Classify VT Omissions

Browse Repository

VTA Maintenance

Business Opportunities

- Improve the process of manually classify verbatim terms to dictionary terms.
- Improve accuracy & consistency in the dictionary coding process.

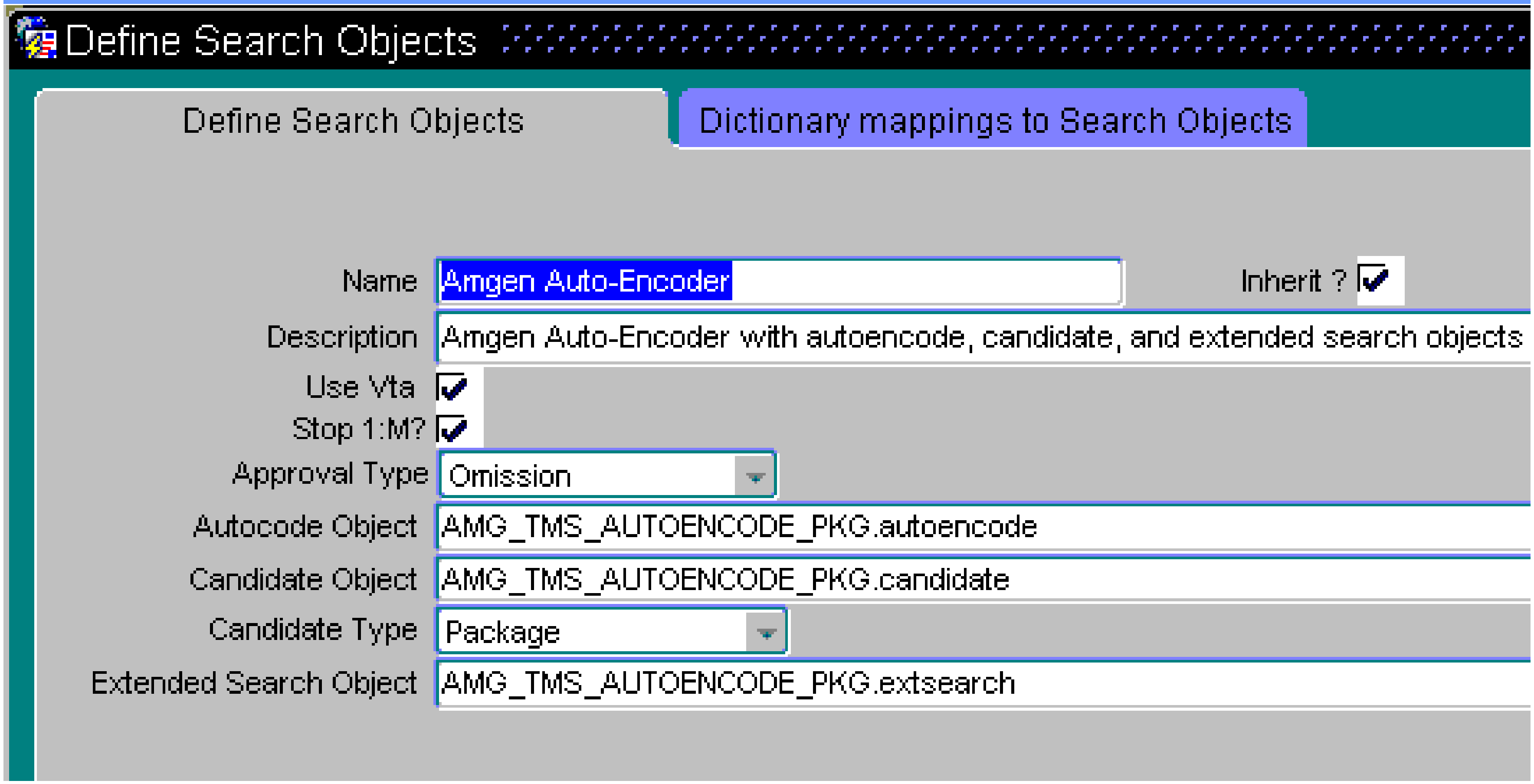
Directives

- Utilize existing TMS functionality to define & execute custom algorithms (no additional GUIs/Forms).
- Utilize complex search procedures to create a list of candidate terms to assist, not change, the existing dictionary coding and peer review workflow.

Directives (2)

- Optimize the search procedure performance by executing during TMS batch validation, not during the dictionary coding process; leverage machine time vs. person time.
- Utilize the existing TMS Classify VT Omissions form to display the list of candidate terms in “best match” sort order.
- Utilize the English lexicon, even though interMedia can support many languages.

Define Search Objects



Name	Amgen Auto-Encoder
Inherit ?	☑
Description	Amgen Auto-Encoder with autoencode, candidate, and extended search objects
Use Vta	☑
Stop 1:M?	☑
Approval Type	Omission
Autocode Object	AMG_TMS_AUTOENCODE_PKG.autoencode
Candidate Object	AMG_TMS_AUTOENCODE_PKG.candidate
Candidate Type	Package
Extended Search Object	AMG_TMS_AUTOENCODE_PKG.extsearch

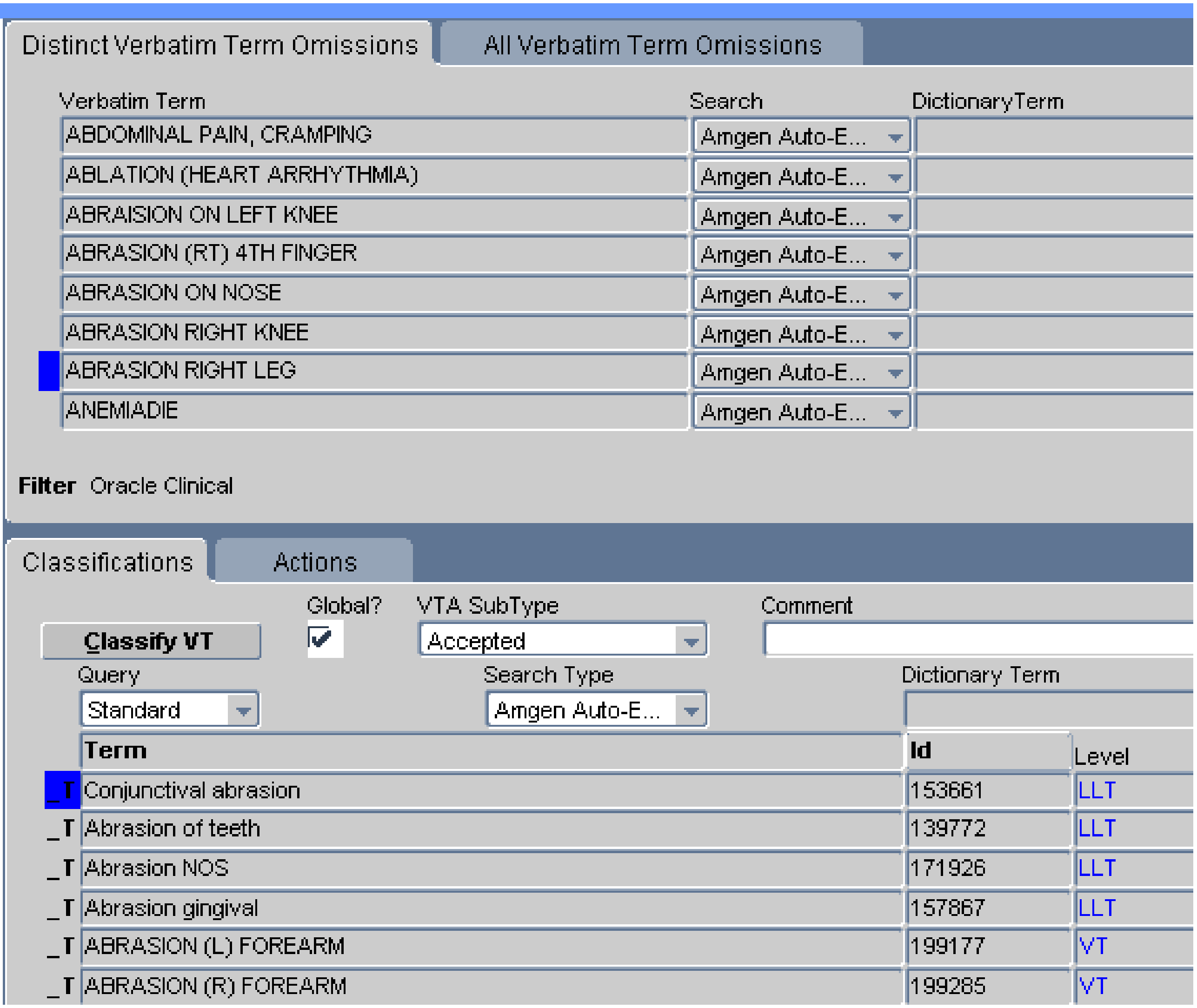
TMS Search Objects

- autoencode
  - Runs automatically during the TMS procedure in batch validation.
- candidate
  - Displays a list of suggested dictionary matches in Classify VT Omissions. Provides the ability to filter the search criteria to display a subset of the candidate terms.
- extsearch
  - Runs On-the-Fly during the auto-encoder search in Extended Search.

autoencode & candidate

- **Autoencoded Terms**

- **Candidate List**



Verbatim Term	Search	DictionaryTerm
ABDOMINAL PAIN, CRAMPING	Amgen Auto-E...	
ABLATION (HEART ARRHYTHMIA)	Amgen Auto-E...	
ABRAISION ON LEFT KNEE	Amgen Auto-E...	
ABRASION (RT) 4TH FINGER	Amgen Auto-E...	
ABRASION ON NOSE	Amgen Auto-E...	
ABRASION RIGHT KNEE	Amgen Auto-E...	
ABRASION RIGHT LEG	Amgen Auto-E...	
ANEMIADIE	Amgen Auto-E...	

Filter Oracle Clinical

Global? ☑

VTA SubType: Accepted

Comment:

Query: Standard

Search Type: Amgen Auto-E...

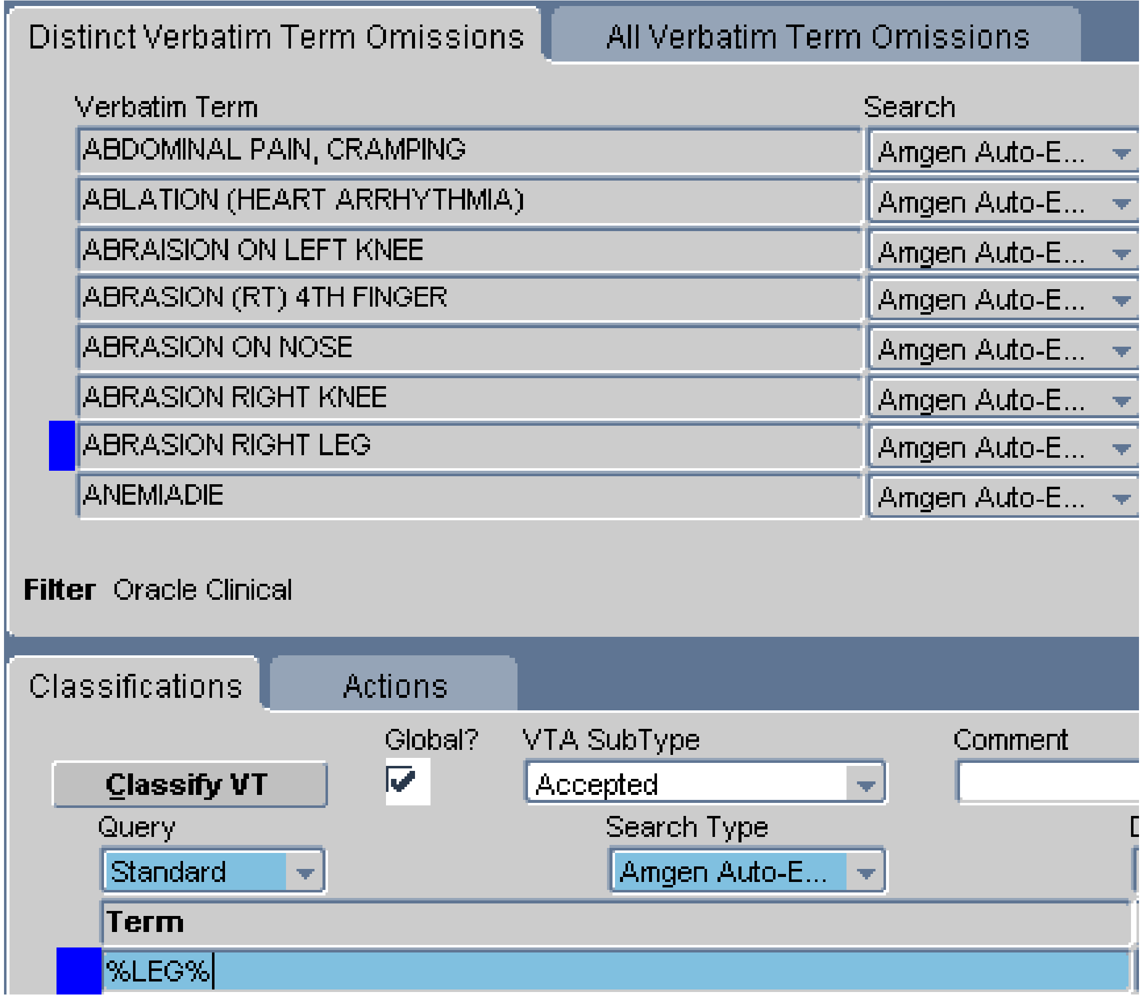
Dictionary Term:

	Term	Id	Level
_T	Conjunctival abrasion	153661	LLT
_T	Abrasion of teeth	139772	LLT
_T	Abrasion NOS	171926	LLT
_T	Abrasion gingival	157867	LLT
_T	ABRASION (L) FOREARM	199177	VT
_T	ABRASION (R) FOREARM	199285	VT

AMGEN

Apply Candidate Filter

- Search for a subset of candidate terms in the candidate list that contain the word “LEG”.



Verbatim Term	Search
ABDOMINAL PAIN, CRAMPING	Amgen Auto-E...
ABLATION (HEART ARRHYTHMIA)	Amgen Auto-E...
ABRAISION ON LEFT KNEE	Amgen Auto-E...
ABRASION (RT) 4TH FINGER	Amgen Auto-E...
ABRASION ON NOSE	Amgen Auto-E...
ABRASION RIGHT KNEE	Amgen Auto-E...
ABRASION RIGHT LEG	Amgen Auto-E...
ANEMIADIE	Amgen Auto-E...

Filter Oracle Clinical

Global? ☑

VTA SubType: Accepted

Comment:

Query: Standard

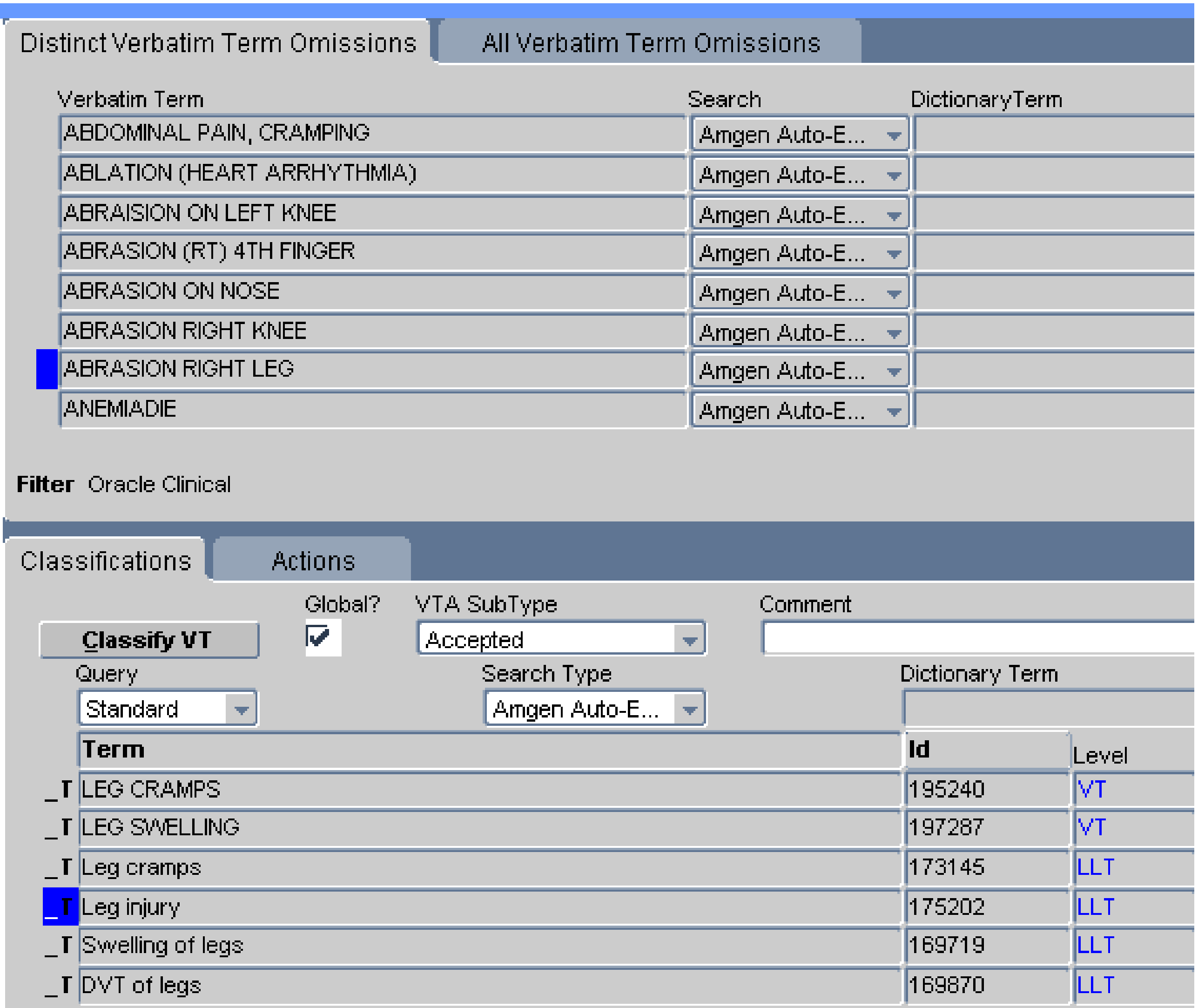
Search Type: Amgen Auto-E...

Term

%LEG%

Candidate Filter Results

- The Candidate filter retrieves a subset of candidate terms containing “LEG”.



Distinct Verbatim Term Omissions | All Verbatim Term Omissions

Verbatim Term	Search	DictionaryTerm
ABDOMINAL PAIN, CRAMPING	Amgen Auto-E...	
ABLATION (HEART ARRHYTHMIA)	Amgen Auto-E...	
ABRAISION ON LEFT KNEE	Amgen Auto-E...	
ABRASION (RT) 4TH FINGER	Amgen Auto-E...	
ABRASION ON NOSE	Amgen Auto-E...	
ABRASION RIGHT KNEE	Amgen Auto-E...	
ABRASION RIGHT LEG	Amgen Auto-E...	
ANEMIADIE	Amgen Auto-E...	

Filter Oracle Clinical

Classifications | Actions

Classify VT

Global? ☑

VTA SubType: Accepted

Comment:

Query: Standard

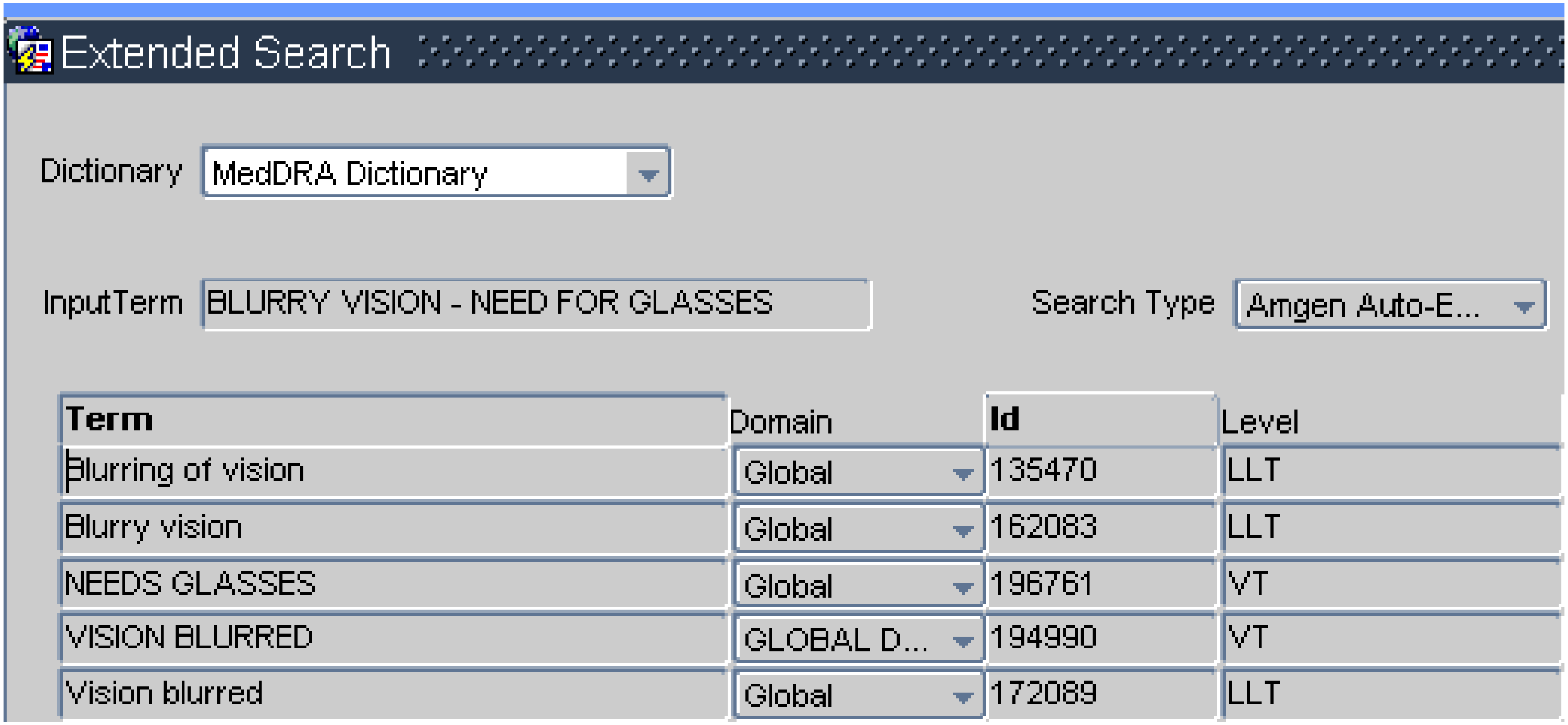
Search Type: Amgen Auto-E...

Dictionary Term:

	Term	**Id**	Level
_T	LEG CRAMPS	195240	VT
_T	LEG SWELLING	197287	VT
_T	Leg cramps	173145	LLT
_T	Leg injury	175202	LLT
_T	Swelling of legs	169719	LLT
_T	DVT of legs	169870	LLT

AMGEN

extsearch



Extended Search

Dictionary: MedDRA Dictionary

InputTerm: BLURRY VISION - NEED FOR GLASSES

Search Type: Amgen Auto-E...

Term	Domain	Id	Level
Blurring of vision	Global	135470	LLT
Blurry vision	Global	162083	LLT
NEEDS GLASSES	Global	196761	VT
VISION BLURRED	GLOBAL D...	194990	VT
Vision blurred	Global	172089	LLT

- **Autoencode any type of term on-the-fly**
- **Autoencoder searches all levels of the dictionary**

Autoencoding Algorithm

- Breaks up a Multi-word Term into individual words.
- Executes procedures against individual words in the order defined in the reference codelist.
  - Full Word Substitutions
    - Remove stop words (“an, nd, st, of” to blank)
    - Create substitution synonym list (TYLENOL to ACETAMINOPHEN)
    - Remove frequent terms

Autoencoding Algorithm (2)

- Partial Word Substitutions
  - Remove punctuation & symbols (“; *” to blank)
  - Remove numeric values (“0 – 9” to blank)
- Porter Stemmer (TOOTH ABSCESSES to Tooth abscess) or (FALLS to Fall)

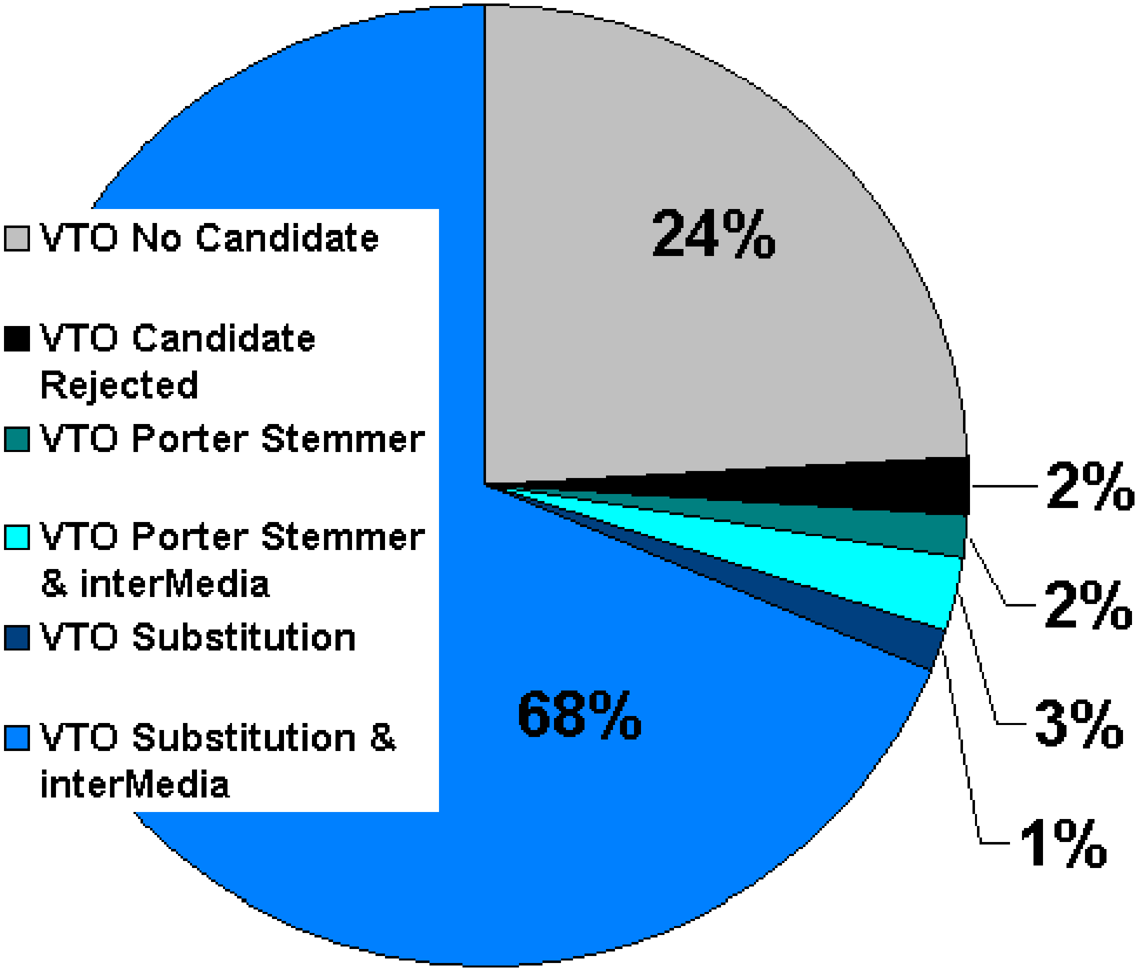
Autoencoding Algorithm (3)

- Reorders individual words with all possible permutations of a Multi-word Term (with limits).
- Searches the dictionary at the classification and verbatim term levels for matches and assigns a ranking value used to order the candidate list.

Autoencoding Algorithm (4)

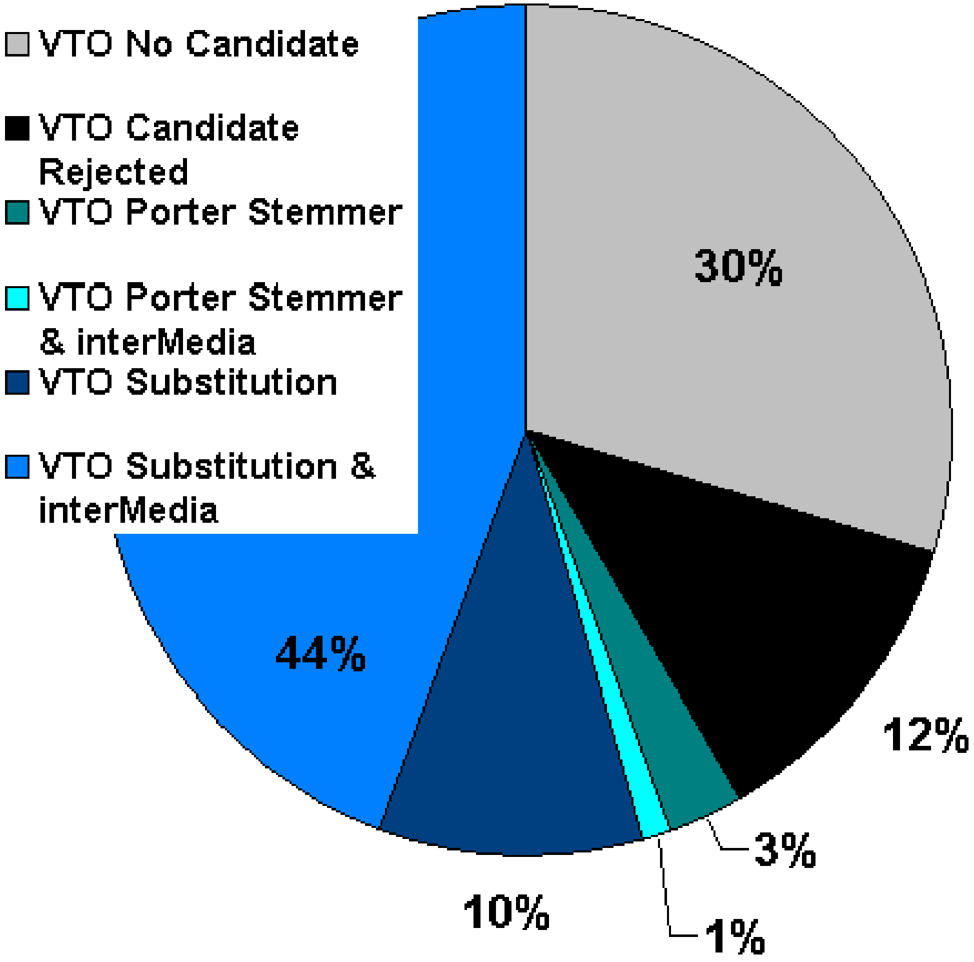
- Executes interMedia Logic and assigns a ranking value used to order the candidate list.
  - The interMedia Lexicon is English.
  - interMedia Indexing is used to perform the ‘CONTAINS’/ ‘ABOUT’ searches.
  - A default set of stop words is used in interMedia searches.

Retrieval Tool Metrics - AEs



- Note: 3 week sampling of VTs autoencoded. Stemmer & Substitution % are based on selected candidates that are approved VTAs.

Retrieval Tool Metrics - Meds



- Note: 3 week sampling of VTs autoencoded. Stemmer & Substitution % are based on selected candidates that are approved VTAs.

Amgen’s Observations

- The most effective term-matching is a combination of substitutions & interMedia.
  - 68% for AEs
  - 44% for Meds
- “English” based lexicon is most effective for AEs but not as strong for Meds supporting existing research.
  - 71% for AEs
  - 45% for Meds

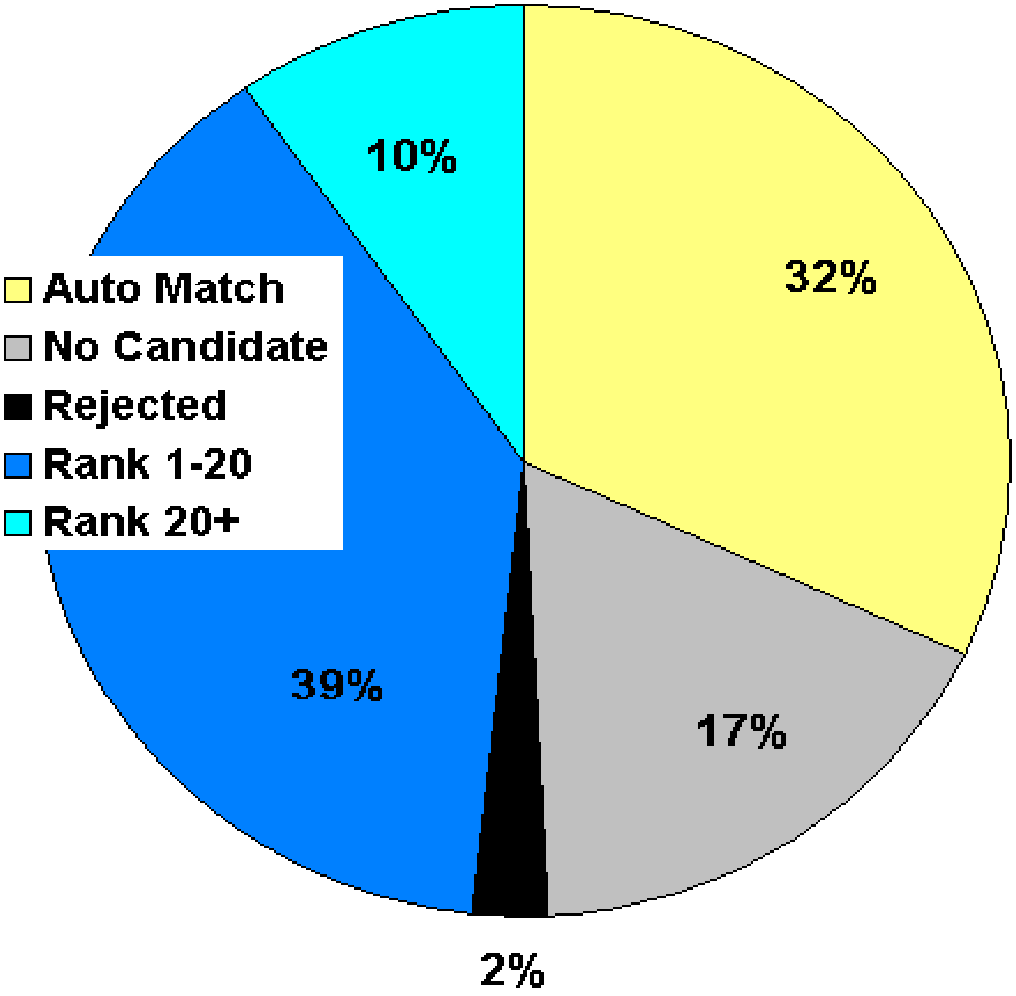
Amgen's Observations (2)

- Porter Stemmer retrieval performs within the expected range 1-3 % supporting existing research.
  - 2% for AEs
  - 3% for Meds
- A combination of Porter Stemmer & interMedia retrieval does not significantly increase term-matching.
  - 3% for AEs
  - 1% for Meds

Amgen’s Observations (3)

- The benefit to having the source code for the Porter Stemmer is being able to control more predictable results.
- Since source code is not available for the Xerox Stemmer, a strict algorithm definition is not available for interMedia.

Effectiveness Metrics



Conclusion

- Efficiency improvements of 39% gained when selecting candidates within the first 20 terms in the candidate list.
- Effective results of 70% are gained through auto matching (equal match) & manually selecting within the first 20 terms in the candidate list.

References

- M. Porter. An algorithm for suffix stripping. Program, 14(3):130-137, 1980. http://www.tartarus.org/~martin/PorterStemmer/index.html
- David A. Hull, Gregory Grefenstette. *A Detailed Analysis of English Stemming Algorithms.* January 31, 1996.
- Metalink. Oracle 8i interMedia Text 8.1.7 Technical Overview. May, 19 2002.
- Oracle 8i interMedia Text Reference Release 2 (8.1.6) December, 1999.

Q&A

?

AMGEN