TERMINOLOGY IN INFORMATION EXTRACTION AND TECHNOLOGY FORECASTING

Adam Meyers
New York University
Funding and Collaboration

• Collaborators at NYU
  • Zachary Glass
  • Ralph Grishman
  • Yifan He
  • Giancarlo Lee
  • Shasha Liao
  • Angus Grieve-Smith

• NYU was a subcontractor of BAE under FUSE
  • FUSE was IARPA’s Foresight and Understanding from Scientific Exposition program
  • There has been occasional interest in extending the system
    • I am currently extending the system to better-handle legal text
Outline

• What is Terminology?
• 2 Types of Terminology in the FUSE program
  • Term Tokens in Information Extraction
  • Term Types in Technology Forecasting
• NYU’s Terminology Extraction System
  • The Termolator: 🕶️
    • System and Evaluation
    • Open Source Distribution
• Concluding Remarks
What is Terminology?

• **Webster’s II New Collegiate Dictionary Definition**
  • The vocabulary of technical terms and usages appropriate to a particular field, subject, science, or art.

• Operational Definitions:
  • Keyword sequences for Information Retrieval (IR)
    • Need not be technical, e.g., *wheat, barley, white mouse*, in genetics
  • Items to define in Technical Glossaries
  • **Items to track for Technology Forecasting (TF)**
  • **Arguments of Information Extraction (IE) Relations**

• Noun Terminology:
  • Technical word sequence headed by noun
  • Vast majority of all terminology
  • Non-noun terminology exists, but not included in this research
The Termolator: 2 Subsystems

- **In-Line Term System**: Finds instances of terms (tokens)
  - Finite State Machine based on dictionaries and POS tags
  - Finds terminology for Information Extraction
  - Identifies term tokens, instances of terms in sentences
    - 500 term tokens occur in document X—50 are instances of H5N1
  - Used for Relation Extraction in FUSE
  - Limited previous work in this area

- **Distributional Term System**: Finds term types
  - Counts instances of term types
    - 30 term types occur in document X—H5N1 occurred 500 times
  - Ranks term types by characteristic-ness to a particular topic
  - Top N term types are kept, the rest are discarded
  - Our FUSE team used term types for Terminology Forecasting
  - Our Distributional Term System Uses In-Line Terms as Input
    - Previous Work: N-grams or Noun-Groups as Input
    - Previous Work used for creating lists of key search terms & glossaries
In-Line Term Tokens are used for IE

• Information Extraction (IE)
  • domain: patents, technical articles, Web of Science abstracts

• Relation arguments are often term tokens

• Entities:
  • Documents (article citations, patents, URLs, standards, self-references like “we” or “our”)
  • People (Inventors, Researchers, etc.)
  • Organizations (Funding Agencies, Patent Holders, …)
  • Term Tokens (topic words, inventions, discoveries, etc.)

• Relations: ABBREVIATE, ORIGINATE, EXEMPLIFY, BASED_ON, CONTRAST, CORROBORATE, BETTER_THAN, PRACTICAL, STANDARD, …
Sample Relations

• **Originate**
  • *Eagle's minimum essential media* and *DOPG* was obtained from *Avanti Polar Lipids*
    • **Originate**(*Eagle, Eagle’s minimum essential media*)
    • **Originate**(*Avanti Polar Lipids, Eagle’s minimum essential media*)
    • **Originate**(*Avanti Polar Lipids, DOPG*)

• **Contrast**
  • *necrotrophic effector system* that is an exciting contrast to the *biotrophic effector models*
    • **Contrast**(*necrotrophic effector system, biotrophic effector models*)

• **Better Than**
  • *Bayesian networks* hold a considerable advantage over *pairwise association tests*
    • **Better than**(*Bayesian networks, pairwise association tests*)
More Sample Relations

• Significant (sentiment-like, author = implied arg)
  • Anaerobic SBs are an emerging area of research and development
    • Significant(Anaerobic SBs)

• Practical (sentiment-like, author = implied arg)
  • The gene proteins used in this experiment
    • Practical(gene proteins)

• Alias
  • Silver behenate, also known as CH3-(CH2)20-COOAg
    • Alias(Silver behenate, CH3-(CH2)20-COOAg)
Defining In-Line Terms for IE Tasks

• Not all Noun Groups (NGs) can be IE arguments
  • NGs include *table top, large number, first step, other diagram*, ...
  • A narrower classification reduces errors for IE patterns
    • just as selection restrictions reduce attachment errors

• If We Run Distributional System with NGs and use only High-Ranking Terms
  • Too few NGs are considered
  • Many relation arguments will be missed

• IE arguments are a subset of NGs and a superset of high-ranking terms
Our Inline Term Extraction System

• Our POS tagset
  • Refines some PTB POS classes and collapses others
  • Uses dictionaries, word lists and morphological rules
  • Classes include Out-of-Vocabulary Nouns, Technical Adjectives, Person Names, …

• A Finite-State-Machine (FSM)-based chunker identifies potential terms (PTs)
  • Uses B/I/E/O tag sequences in style of (Ramshaw and Marcus 1995) to represent states corresponding to each word W in sentence
  • State(W) depends on: POS(W), POS(W-1) and State(W-1)
    - $PT = E \lor BI^* \lor BI^*E$

• Filter makes final selection of inline terms
  • Similar well-formedness filter in Distributional System
In-Line Term Extractor

Input Text

A semiconductor device which includes a semi...

FSM Chunker

Potential Terms
semiconductor device surface stiffener semiconductor chip...

Inline-Terms
semiconductor device semiconductor chip...

Filters
FSM Identifies Potential Terms:

• A *semiconductor device* which includes: a *semiconductor chip* bonded to a *surface* of a solid *device*; and a *stiffener* surrounding the *periphery* of the *semiconductor chip*.

• Differs from standard noun group chunking because some premodifiers are excluded (determiners, some adjectives)
Filters Remove Unlikely Candidate Terms

• Accepts Terms from previous slide which each contain an O-NOUN (Out-of-vocabulary NOUN)
  • semiconductor/O-NOUN device
  • semiconductor/O-NOUN chip (2 instances)
• Also accepts Terms containing technical adjectives or nominalizations
  • thermal/TECH-ADJ stress
  • fabrication/NOM process
• Rejects Terms because they contain no O-NOUNs, technical adjectives, or other qualifying words
  • surface
  • device
  • stiffener
  • periphery
• Other Non-Terms (e.g., morphological properties, status as NE, etc.)
  • T
  • 212-345-8888
  • No.
  • New York
Supplementary patterns for identifying Terms

- Arguments of Abbreviation relations
  - Not organizations or places
  - Aligns words before parentheses with word in parentheses
    - *already been chewed* (ABC)
    - XML (Extensible Markup Language)
    - *third variable loop* (V3)
    - *D. melanogaster gene Muscle LIM protein at 84B* (abbreviated as *Mlp84B*)

- Terms Matching Regexp Patterns
  - Gene Sequences: *AACAAGGTGGCGCAGTT*
  - Chemical Formulas: *Ag2CrO4*
Evaluation of Inline Term System

- 2 Annotators Manually Annotated Inline terms in 3 documents
- Adjudicated the Results
- Scored annotators against adjudicated annotation
- Scored system against adjudicated annotation
- Compared annotator vs system performance
Annotation

• Setup
  • 2 annotators annotated the same three documents
  • Annotator 2 Adjudicated
  • Annotator 1’s score against Annotator 2 may be a good Upper Bound for evaluating the Automatic System (assumes the adjudication is biased in favor of Annotator 2).

• Defining Inline Term for Annotator
  • Single or multi-word nominal expression specific to technical discipline
  • It can be conventionalized by defining or abbreviating it early in the document and by reusing the term
  • Is term specific to technical discipline, i.e., is it obscure enough?
    • Would a naïve adult (like Homer Simpson) know the term?
    • Is it found in the Juvenile subcorpus of the Corpus of Contemporary American English (http://corpus.byu.edu/coca/)?
Corpora and Systems Tested

• Corpora
  • A Speech Recognition Patent (SRC)
  • A Sun Screen Patent (SUP)
  • A Journal Article about a Virus Vaccine (VVA)

• Systems Tested
  • Base 1: assume all noun groups minus determiners are terms
    • use MEMM chunker with Genia (Kim et al 2003) features
  • Base 2: baseline 1 system, but filtered by only keeping those Noun Groups that end with an O-NOUN
  • System without Filter: The chunking system as described, but without the filter
  • Final System

• Matching Criteria
  • Strict Match – The test term and answer key term are the same
  • Sloppy Match – The test term and answer key term overlap in extent.
## Inter Annotator Agreement

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Doc</th>
<th>Terms</th>
<th>Matches</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Matches</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotator 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>798</td>
<td>1041</td>
<td>92.5%</td>
<td>92.0%</td>
<td>92.2%</td>
<td>1992</td>
<td>96.3%</td>
<td>92.0%</td>
<td>94.1%</td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>1809</td>
<td>1992</td>
<td>96.3%</td>
<td>92.0%</td>
<td>94.1%</td>
<td>2062</td>
<td>98.5%</td>
<td>95.2%</td>
<td>96.8%</td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>713</td>
<td>762</td>
<td>97.2%</td>
<td>82.9%</td>
<td>89.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Annotator 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>960</td>
<td>968</td>
<td>99.2%</td>
<td>85.6%</td>
<td>91.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>1999</td>
<td>2062</td>
<td>98.5%</td>
<td>95.2%</td>
<td>96.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>838</td>
<td>855</td>
<td>99.4%</td>
<td>93.0%</td>
<td>96.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Annotator 1 scores may be upper bounds for system results.
## Baseline Systems

<table>
<thead>
<tr>
<th></th>
<th>Doc</th>
<th>Terms</th>
<th>Matches</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Matches</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>602</td>
<td></td>
<td>24.3%</td>
<td>53.2%</td>
<td>33.4%</td>
<td>968</td>
<td>44.2%</td>
<td>96.8%</td>
<td>60.7%</td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>1367</td>
<td></td>
<td>36.5%</td>
<td>63.1%</td>
<td>46.2%</td>
<td>1897</td>
<td>50.6%</td>
<td>87.6%</td>
<td>64.2%</td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>576</td>
<td></td>
<td>28.5%</td>
<td>62.7%</td>
<td>39.2%</td>
<td>887</td>
<td>44.0%</td>
<td>96.5%</td>
<td>60.4%</td>
</tr>
<tr>
<td><strong>Base 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>66</td>
<td></td>
<td>24.9%</td>
<td>5.8%</td>
<td>9.5%</td>
<td>151</td>
<td>57.0%</td>
<td>13.4%</td>
<td>21.6%</td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>771</td>
<td></td>
<td>52.3%</td>
<td>35.6%</td>
<td>42.4%</td>
<td>1007</td>
<td>68.4%</td>
<td>46.5%</td>
<td>55.3%</td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>270</td>
<td></td>
<td>45.8%</td>
<td>29.4%</td>
<td>35.8%</td>
<td>392</td>
<td>66.5%</td>
<td>42.6%</td>
<td>51.9%</td>
</tr>
</tbody>
</table>

- Base 1 (all noun groups): results in high recall/low precision
- Base 2 (must end in O-NOUN): too severe a filter.
## System Results

<table>
<thead>
<tr>
<th></th>
<th>Doc</th>
<th>Terms</th>
<th>Matches</th>
<th>Strict</th>
<th>Sloppy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pre</td>
<td>Rec</td>
</tr>
<tr>
<td><strong>No Filter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>932</td>
<td></td>
<td>39.0%</td>
<td>82.4%</td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>1475</td>
<td></td>
<td>39.7%</td>
<td>68.1%</td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>629</td>
<td></td>
<td>27.8%</td>
<td>68.4%</td>
</tr>
<tr>
<td><strong>Final System</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>1131</td>
<td>669</td>
<td></td>
<td>69.0%</td>
<td>59.2%</td>
</tr>
<tr>
<td>SUP</td>
<td>2166</td>
<td>1193</td>
<td></td>
<td>64.7%</td>
<td>55.1%</td>
</tr>
<tr>
<td>VVA</td>
<td>919</td>
<td>581</td>
<td></td>
<td>62.1%</td>
<td>63.2%</td>
</tr>
</tbody>
</table>

Final System gets the highest F-score
Term Types Used in Technology Forecasting

- **Technology Forecasting (TF)** includes tracking the distribution of instances of the same term
  - A **term type** is a set of instances of the same term
- A Topic can be represented by a set of term types that “characterize” that topic.
  - Term types with more instances in topic X than in general
- Changes in a topic over time can be indicated by
  - Changes in the set of term types characterizing the topic
  - Changes in the frequency of those term types
- Changes in frequencies of term types in a topic over time
  - Can indicate changes in the “prominence” of these terms
- Approximately the same terms used for: search keys and glossaries (in previous work)
Our Distributional Term System

- Find In-line Terms for Foreground Corpus (or sample)
- Find In-line Terms for Background Corpus (or sample)
- Count instances of the same term
  - Allow for some variation
    - Implemented: (stemming) singular/plural, -ing endings, other
    - Partially Implemented:
      - Abbreviation/full-form
      - Noun mod alternations: Recognition of Speech → Speech Recognition
- Use Statistical Metrics to find terms that are:
  - More characteristic of Foreground than Background
- Rank terms by Metrics
- Rerank terms using additional metrics
  - Relevance Metric, based on a Yahoo Websearch
  - Well-formedness Metric: based on manual rules
- Take Top N terms
Distributional Term Extractor

Foreground Text

A semiconductor device which includes a semi...

A cascade pattern differs from a shower pattern ...

Background Text

1 In-Line Term Extractor

Foreground Terms
semiconductor device
semiconductor chip
bin-sort algorithm

2 Distributional Comparison Using Statistical Metrics

Background Terms
Newton's 2nd Law
cascade pattern
correlation coefficient...

3 WF and Relevance Filters

Ranked Output
transfer electrode
semiconductor face
micropattern

Final ReRanked Output: transfer electrode; semiconductor face; micropattern, through-connection, wavelength conversion chip, …
Metrics for Distributional Ranking

• A linear combination of 3 Measures comparing term distribution in Foreground (For) vs Background (Bac)
• Term Frequency Inverse Document Frequency (TFIDF)
  \[ TFIDF(t) = \frac{freq_{For}(t)}{freq_{Bac}(t)} \times \log\left(\frac{num_{BacDocs}}{num_{BacDocContains}(t)}\right) \]
• Document Relevance Document Consensus (DRDC)
  • (Navigli and Velardi, 2004)
  \[ DRDC(t) = \frac{freq_{For}(t)}{freq_{Bac}(t)} \times \sum_{d \in RDG} \frac{freq(t,d)}{freq_{For}(t)} \times \log\left(\frac{freq_{For}(t)}{freq(t,d)}\right) \]
  • Doc Relevance favors representative terms (like TFIDF)
  • Doc Consensus favors terms found in many documents
• Kullback-Leibler Divergence (KLD)
  • (Cover and Thomas, 1991; Hisamitsu et al., 1999).
  \[ KLD(t) = \log(freq_{For}(t)) - \log(freq_{Bac}(t)) \times freq_{For}(t) \]
  • Compares probability a term occurs in Foreground vs Background Corpus
Filters on Distributional Output

• 2 Filters that can be applied to our system or output of other term generation systems
  • In FUSE, they were applied to MITRE and BBN output
• Both scores are between 0 and 1, they are combined by multiplication

• Well-Formedness Filter
  • Many of the constraints are built into our chunker
    • Most terms have a score of 1
  • However, component of distributional System adds some common substrings of terms to output, some of which are ill-formed

• Relevance Filter
  • We use a Yahoo search result and heuristics to score terms more highly if they are used in articles or patents
Well-FormedNess Filter

• A term is well-formed if it is:
  • An abbreviation
  • A set of words that is abbreviated somewhere in the corpus
  • A single out of vocabulary word
  • Matches a regular expression that finds chemical names, DNA sequences or paths (urls, bio paths, etc.) – although URLs can be documents, rather than terms.

• A term is also well formed if it obeys noun group rules (a sequence of adjectives and nouns ending in a noun) AND it contains at least one out-of-vocabulary word, nominalization or technical adjective

• The degree of ill-formedness is not so important as scores below 1 rarely apply to accepted terms.

• This filter is more important when applied to term lists not created by The Termolator (Mitre and BBN term lists in FUSE)
Relevance Filter

- Run on each term below some cutoff (typically 30K)
  - Time consuming (about .75 seconds per term)
- Yahoo search (Bing) for exact match of term
  - Currently use free version
  - Yahoo Api (Boss) would cost .18 cents per search
    - https://developer.yahoo.com/boss/search
- Relevance = $H^2T$
  - $H =$ score representing number of hits
    - $\min\left(\log_{10}(\text{numberHits}), 10\right)$
      - Minimized to 1 for non-hits ("showing results for" and similar)
  - $T =$ Percent of top 10 hits that are articles or patents
    - Based on key word search in title, url and summary
      - Key words = \{patent, article, proceedings, journal, dissertation, abstract, \ldots\}
Evaluation of Distributional Term System

• Foreground Corpus: 2500 patents about optical systems
• Background Corpus: 2500 randomly selected patents
• Years: 1997-2007
• Ran the Distributional System and Ordered the Terms
  • Confidence_1 = Percentile X Well-Formedness
    • Uses the Percentile Ranking based on the distributional score, but filters out ill-formed terms
• Took the top 30K out of 219K terms and reranked using:
  • Relevance only; and
  • Confidence_2 = Percentile X Well-Formedness X Relevance
    • Uses Relevance on 30K terms due to time constraints
• Took Top 5000 terms ranked each of 3 ways and Scored for Precision
  • Confidence$_1$ Precision = 71%
  • Relevance Precision = 82%
  • Confidence$_2$ Precision = 86%

• For each ranked set, we took samples of 100 terms:
  • 20 from first 20%, 20 from second 20%, … 20 from 5$^{th}$ 20%.

• We manually evaluated the samples:
  • Terms were deemed correct if the term was deemed a valid keyword, was not missing any crucial modifier, and did not contain any spurious word.
## Example Evaluations

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>D</th>
<th>W</th>
<th>R</th>
<th>Total</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>stimulable phosphor</td>
<td>.866</td>
<td>1</td>
<td>.174</td>
<td>.151</td>
<td>Yes</td>
</tr>
<tr>
<td>104</td>
<td>ion beam profile</td>
<td>.889</td>
<td>1</td>
<td>.117</td>
<td>.126</td>
<td>Yes</td>
</tr>
<tr>
<td>346</td>
<td>x-ray receiver</td>
<td>.906</td>
<td>1</td>
<td>.099</td>
<td>.089</td>
<td>Yes</td>
</tr>
<tr>
<td>533</td>
<td>wavelength-variable</td>
<td>.838</td>
<td>1</td>
<td>.091</td>
<td>.076</td>
<td>Yes</td>
</tr>
<tr>
<td>556</td>
<td>irradiation time t</td>
<td>.460</td>
<td>1</td>
<td>.163</td>
<td>.075</td>
<td>No</td>
</tr>
<tr>
<td>1275</td>
<td>quadrupole lens</td>
<td>.460</td>
<td>1</td>
<td>.113</td>
<td>.052</td>
<td>Yes</td>
</tr>
<tr>
<td>1502</td>
<td>evolution</td>
<td>.439</td>
<td>1</td>
<td>.109</td>
<td>.048</td>
<td>No</td>
</tr>
<tr>
<td>1581</td>
<td>proximity correction</td>
<td>.451</td>
<td>1</td>
<td>.103</td>
<td>.046</td>
<td>Yes</td>
</tr>
<tr>
<td>1613</td>
<td>dfb laser</td>
<td>.943</td>
<td>1</td>
<td>.049</td>
<td>.046</td>
<td>Yes</td>
</tr>
<tr>
<td>1685</td>
<td>asymmetric stress</td>
<td>.493</td>
<td>1</td>
<td>.067</td>
<td>.033</td>
<td>Yes</td>
</tr>
<tr>
<td>3834</td>
<td>panoramagram</td>
<td>.483</td>
<td>1</td>
<td>.056</td>
<td>.027</td>
<td>Yes</td>
</tr>
<tr>
<td>4203</td>
<td>crystal adjacent</td>
<td>.316</td>
<td>1</td>
<td>.080</td>
<td>.025</td>
<td>No</td>
</tr>
<tr>
<td>4244</td>
<td>single-mode optical fiber</td>
<td>.875</td>
<td>1</td>
<td>.029</td>
<td>.025</td>
<td>Yes</td>
</tr>
<tr>
<td>4467</td>
<td>total reflection plane</td>
<td>.988</td>
<td>1</td>
<td>.024</td>
<td>.024</td>
<td>Yes</td>
</tr>
<tr>
<td>4879</td>
<td>photosensitive epoxy resin</td>
<td>.286</td>
<td>1</td>
<td>.079</td>
<td>.022</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Sample Incorrect Terms

- *irradiation time t*
  - A variable, not a term (without \( t \), it would be a term)
- *evolution*
  - This word has entered the common vocabulary
- *crystal adjacent*
  - This word sequence includes two words at a constituent boundary
    - a noun phrase followed by a modifying adjective phrase, e.g.,
    - [[*a liquid crystal*] [*adjacent to the lower alignment layer*]]
Informal Observations about Recall

- Recall or coverage is difficult to measure without an exhaustive amount of human annotation.
- The distributional system gets roughly the same precision for Noun Group input as Inline Term Group Input for the top N terms, where N is a small number.
- Using Inline Terms as input, we generate many more terms with high scores and thus seem to improve Recall by a large amount (at least a factor of 2).
  - But this is hard to measure.
- Rationale: Garbage In → Garbage Out
  - High F-scores for inline terms (vs NGs or N-grams).
  - Higher Quality terms are being ranked and so the high-ranked items are more likely to be correct.
The Termolator for Chinese

- Work by Yifan He
- Distributional System is the Same as English
- Uses Noun Group Chunker for input terms
- Accessor-Variety Filter (Feng et al., 2004)
  - Score Based on the Number of distinct words that appear before and after a particular term type
  - Low Scores indicate unlikely Chinese words
- 1100 terms extracted from 2000 speech recognition patents
  - 78% precision on top 50 terms
  - 85% precision on top 20 terms
Example of Chinese Term Filtering

- Examples for Access Variety based filtering
  - 尔科夫模型 (Markov model, with the first Chinese character 马 missing) is probably a boundary error
  - [Pic on left] 尔科夫模型 has the same character 马 on its left boundary thus its Left AV=1
  - [Pic on left] A correct term 声学 (acoustics) will have Left AV>=3
Open Source Distribution

• Open Source release of The Termolator
  – NYU’s Website: http://nlp.cs.nyu.edu/termolator/
  – Github:
    • English: https://github.com/AdamMeyers/The_Termolator
    • Chinese: https://github.com/ivanhe/termolator/

• English system for UTF-8 (including ASCII) & ISO-8859-1

• Tested on Public Domain Texts
  • Google Patents
  • Project Gutenberg
  • Open American National Corpus
Examples from Public Domain Texts

- Gutenberg: Chapters in a Book about knitting vs Other Docs
  - *open-work insertion, fine mesh, transverse stitching, empty scallop*

- Open American National Corpus (OANC) – Biology documents versus random documents
  - *myosin-ii, hsn3, intron, migration defect, sparc-null mice*

- Google Patents: Surgery patents (US Patent Class 606) vs Random Patents:
  - *fluid manifold, dissector arm, pedicle punch, balloon catheter*
Future Work: Initial Test in Legal Domain

- Foreground = Roe v Wade Court Opinion
- Background = 64,000 Supreme Court Opinions
- Adaption – Added words from several legal dictionaries to our dictionary
- Top 10 terms in system output (good):
  - medical-legal history, medical abortion practices, common-law prosecutions, roman catholic dogma, good-faith belief, definitional deficiencies, canon-law treatment, historical statutory development, common-law scholar, uniform abortion act
- Terms 21-30 (still mostly good)
  - abortion controversy, canon-law crime, executive assistant attorney, embryo, medical practice act, postconceptive, connecticut birth control law, texas abortion statute, texas to-wit, birth control pill
- Terms 101-110 (OK, but not as topical)
  - texas abortion, sensitive judge, medical advance, childe, abortion patient, gynecologic, therapeutic abortion, extradition provision, governmental invasion
- Terms 401-410 (bad, probably useless)
  - Law-suit, district court hearing, columbia statute, common-law precedent, pseudonym, health department, pasteurizer, pasteur, florida lime, licensure
Our Papers on Terminology & NLP of Technical Literature


Previous Work on Terminology Extraction

• Terms are most typically Noun Groups or Obey Other Linguistic Rules

• Comparing foreground & background documents to rank terms (many others)

• Finding Terminology via Relational Patterns
Concluding Remarks

• Statistical Comparison of terms in foreground/background is an established method of term extraction.
  • Previous methods use Noun Groups or N-grams as input
• Terminology tokens are often arguments of IE relations
  • Statistical methods cannot find most of these terms
  • Noun Groups produce noisy input for IE
  • Technical NGs, Noun Group-like phrases that include likely technical words (OOV words, technical adjectives, nominalizations, etc.), provide better input for IE
• Using Technical NGs as input to Statistical Term Extraction Results in More High Precision Terms
  • Better input yields more meaningful comparisons (Garbage In, Garbage Out)
• A web-search-based relevance filter improves results
  • Non-Terms are unlikely to be mentioned in technical documents accessed on WWW
• Results:
  • Top In-line Term System: 77.9% sloppy F measure (vs. human ~92% F-measure)
  • Top Distributional Terms System: 86% precision
Extra Slides

• Slides Useful for Providing Extra Examples or Answers to Questions
Customized Parts of Speech 1

• Types of nouns (POS tagger marks NN or NNS)
  • O-NOUN: word is not in our lexicon (Comlex Syntax, lists of person names, lists of specialize vocabulary, e.g., chemical names)
  • PER-NOUN: word begins with a capital letter and is in our dictionary of first and last names
  • PLUR-NOUN: NNS nouns not marked O-NOUN or PER-NOUN
  • NOUN: Other NN nouns

• Types of adjectives (POS tagger marks JJ, JJR, JJS)
  • STAT-ADJ: first word in top ranked term in statistical system
  • TECH-ADJ:
    • adjective ends in -ic, -ous, -ary, and others
    • not in list of exceptions (basic, analogous, voluntary, …)
  • NAT-ADJ
    • adjectival form of country/state/city/continent: European, Indian, Peruvian
  • CAP-ADJ – adjective beginning with a capital letter
Customized Parts of Speech 2

- Verbs:
  - VBG = ING-VERB
  - VBN/VBD = ED-VERB
  - Other verbs are marked OTHER
- POS: possessive marker
- PREP: POS tagger marks TO or IN
- ROM-NUM: I, II, III, IV, …
- Det -- Determiner
- OTHER – all other parts of speech from tagger
Finite State Machine 1

• States
  • S = Start of word sequence
  • B-T = Beginning of Term
  • E-T = End of Term
  • I-T = Inside of Term
  • O = Other

• Transitions to new State is conditioned on:
  • Previous POS
  • Current POS
  • Previous State

• A Possible Term (PT) is:
  • a single E-T
  • B-T + zero or more I-T + zero or one E-T
## Finite State Machine 2

<table>
<thead>
<tr>
<th>Previous POS</th>
<th>Current POS</th>
<th>Previous State</th>
<th>New State</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET, PREP, POSS, OTHER</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>O-NOUN, C-NOUN, PLUR-NOUN</td>
<td>ROM-NUM</td>
<td>B-T, I-T</td>
<td>E-T</td>
</tr>
<tr>
<td></td>
<td>PLUR-NOUN</td>
<td>B-T, I-T</td>
<td>I-T</td>
</tr>
<tr>
<td></td>
<td>ADJ, CAP-ADJ</td>
<td>I-T</td>
<td>I-T</td>
</tr>
<tr>
<td></td>
<td>NOUN, PER-NOUN, O-NOUN</td>
<td>B-T, I-T</td>
<td>I-T</td>
</tr>
<tr>
<td>O-NOUN</td>
<td>CAP-ADJ, TECH-ADJ, STAT-ADJ, NAT-ADJ</td>
<td>B-T, I-T</td>
<td>I-T</td>
</tr>
<tr>
<td></td>
<td>CAP-ADJ, TECH-ADJ, NAT-ADJ, ING-VERB, ED-VERB, STAT-ADJ, NOUN, O-NOUN, PER-NOUN</td>
<td>E-T, O, S</td>
<td>B-T</td>
</tr>
<tr>
<td>TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ</td>
<td>TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ</td>
<td>B-T, I-T</td>
<td>I-T</td>
</tr>
<tr>
<td>Else</td>
<td></td>
<td></td>
<td>O</td>
</tr>
</tbody>
</table>
Term Filter

• Contains at least one noun.
• Is More than 1 character long
• Contains at least one word of all alphabetic characters.
• Does not end in abbrev from list: e.g., cf., etc., …
• No word violating morphological filter, ruling out various ID numbers, patent numbers, etc.
• Does not end in common ending of patent section headings
• Meets at least one of the following Conditions
  • Is a highly ranked topic term
  • Contains a highly ranked topic term
  • Contains at least one O-Noun
  • Is at least 4 words long and contains 3 words that are nominalizations (from NOMLEX) or TECH-ADJ
  • Is a nominalization and is at least 11 characters long
  • Is more than one word long, ends in a common noun and contains a nominalization
• Additional Filters to recognize NEs among PTs