Terminology Extraction

Adam Meyers
New York University

IR and Related Applications

2020

1st a Demo

- Choose an instance of terminology that is likely to be mentioned in Wikipedia.
- I will go through these steps now and hope that an automatically generated glossary will be completed before the end of this talk.
- If the system takes too long, I will email the results or show them to you next class.

Terminology Extraction

- Given
  - set of documents about a topic (foreground)
  - set of documents about diverse topics (background)
- Find ranked list of terms (words, n-grams, etc.)
  - that are more characteristic of foreground than background
- Uses:
  - terms for previously described tasks, search terms, terms for glossary, terms to track for technology forecasting (predicting technological emergence), etc.

Termolator 

- Open Source Teminology Extraction for Chinese, French & English
  - http://nlp.cs.nyu.edu/termolator/
- Created under a goverment contract as part of the Foresight and Understanding from Scientific Exposition (FUSE). Subsequent development was supported under PRediction of Emergent SCIENce & Technology (PRESCIENT).
- Collaborators at NYU: Zachary Glass, Ralph Grishman, Yifan He, Giancarlo Lee, Shasha Liao, Angus Grieve-Smith, John Ortega, Yuling Gu, Leizhen Shi, Sandra Burlaud, Anand Tyagi and others

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

Examples of Terminology

- Juggling: ***cascade pattern, Mills mess, full shower***
- Real Estate: ***balloon mortgage, title search, full shower***
- Computer Science: **hidden markov model, genetic algorithm, top-down search**
- Knitting: ***gobelin stitch, half-treble, corner scallop***
- biology: ***myosin-ii, plasminogen activator, antizyme***

Termolator Work Flow

Foreground Text

A semiconductor device which includes a semi…

Background Text

A cascade pattern differs from a shower pattern …

1 In-Line Term Extractor

Foreground Terms

semiconductor device
semiconductor chip
bin-sort algorithm
…

Background Terms

Newton's 2nd Law
cascade pattern
correlation coefficient…

2 Distributional Comparison Using Statistical Metrics

Ranked Output

transfer electrode
semiconductor face
micropattern

3 WF and Relevance Filters

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

The Termolator: 2 Main Subsystems

- **In-Line Term System**: Finds instances of terms (tokens)
  - Finds noun/adjective sequences that obey constraints
  - Identifies term tokens, instances of terms in sentences
      - 500 term tokens occur in document X
      - 50 are instances of ***biotrophic effector models***
  - Limited previous work in this area
- **Distributional Term System**: Finds term types
  - Counts instances of term types
    - 30 term types occur in document X
      - ***biotrophic effector models*** occurred 500 times
  - Ranks term types by characteristic-ness to a particular topic
  - Top N term types are kept, the rest are discarded
  - Uses metrics similar to TF-IDF discussed in previous slides

Our In-line Term Extraction System

- Manual Rule Based "Chunker"
  - Identifies sequences of nouns and adjectives
    - part of speech tagger output
    - dictionaries
  - Technical words identified:
    - Out-of-Vocabulary (OOV) words – words not in dictionaries
      - ***semiconductor, biotrophic, gobelin***
    - Technical Adjectives – based on endings (-ic,-cal,-ous) and dictionaries
      - ***algebraic, amphibious, umbilical***
    - Nominalizations – based on endings (tion, etc.) and NOMLEX dictionary
      - ***conduction, vulcanization, accelerator***
- Well-formedness filter
  - eliminates ill-formed (too short, bad characters, etc.)
  - eliminates terms without OOV or technical words
  - eliminates words detected to be names of people or places

Example Identification of Technical Noun Adjective Sequences

- *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****.*

A$_{\text{DET/0}}$ *semiconductor*$_{\text{O-NOUN/B}}$ *device*$_{\text{NOUN/I}}$ *which*$_{\text{OTHER/O}}$ *includes*$_{\text{OTHER/O}}$ *a*$_{\text{DET/0}}$ *semiconductor*$_{\text{O-NOUN/B}}$ *chip*$_{\text{NOUN/I}}$ *bonded*$_{\text{VERB/O}}$ *to*$_{\text{PREP/O}}$ *a*$_{\text{DET/O}}$ *surface*$_{\text{NOUN/B}}$ *of*$_{\text{PREP/O}}$ *a*$_{\text{DET/O}}$ *solid*$_{\text{ADJ/O}}$ *device*$_{\text{NOUN/B}}$ *;*$_{\text{OTHER/O}}$ *and*$_{\text{OTHER/O}}$ *a*$_{\text{DET/O}}$ *stiffener*$_{\text{NOUN/B}}$ *surrounding*$_{\text{VERB/O}}$ *the*$_{\text{DET/0}}$ *periphery*$_{\text{NOUN/B}}$ *of*$_{\text{PREP/O}}$ *the*$_{\text{DET/O}}$

semiconductor$_{\text{O-OUN/B}}$ *chip*$_{\text{NOUN/I}}$ *.*$_{\text{OTHER/O}}$

Rules group yellow words (below) together resulting in blue sequences (above).

Filters Remove Unlikely Candidate Terms

- Accepts Terms which contain an Out-of-vocabulary (OOV) word
  - semiconductor/O-NOUN device
  - semiconductor/O-NOUN chip (2 instances)
- Accepts Terms containing technical adjectives or nominalizations
  - thermal/TECH-ADJ stress
  - fabrication/NOM process
- Rejects Terms because they contain no technical words
  - *surface*
  - *device*
  - *stiffener*
  - *periphery*
- Other Non-Terms removed for other reasons
  - 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**)*
  - *Schwartz and Hearst (2003)*
- 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 Adjudicated results 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: 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

				Strict				Sloppy		
	Doc	Terms	Matches	Pre	Rec	F	Matches	Pre	Rec	F
Annot 1	SRP	1131	798	70.8%	70.6%	70.7%	1041	92.5%	92.0%	92.2%
	SUP	2166	1809	87.5%	83.5%	85.5%	1992	96.3%	92.0%	94.1%
	VVA	919	713	90.9%	77.6%	83.7%	762	97.2%	82.9%	89.5%
Annot 2	SRP	1131	960	98.4%	84.9%	91.1%	968	99.2%	85.6%	91.9%
	SUP	2166	1999	95.5%	92.3%	93.8%	2062	98.5%	95.2%	96.8%
	VVA	919	838	97.4%	91.2%	94.2%	855	99.4%	93.0%	96.1%

Annotator 1 scores may be upper bounds for system results

Baseline Systems

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

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

System Results

				Strict				Sloppy		
	Doc	Terms	Matches	Pre	Rec	F	Matches	Pre	Rec	F
No Filter	SRP	1131	932	39.0%	82.4%	53.0%	1121	46.9%	99.1%	63.7%
	SUP	2166	1475	39.7%	68.1%	50.2%	1962	52.8%	90.6%	66.7%
	VVA	919	629	27.8%	68.4%	39.5%	900	39.8%	97.9%	56.6%
Final System	SRP	1131	669	69.0%	59.2%	63.7%	802	82.8%	70.9%	76.4%
	SUP	2166	1193	64.7%	55.1%	59.5%	1526	82.8%	70.5%	76.1%
	VVA	919	581	62.1%	63.2%	62.7%	722	77.2%	78.6%	77.9%

Final System gets the highest F-score

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 lemma as instances of the same term
  - singular/plural, -ing endings (stemming)
    - ***speech recognizers*** → ***speech recognizer***
  - abbrevation/full-form
    - ***html*** → ***hypertext markup language***
  - Noun mod alternations:
    - ***Recognition of Speech*** → ***Speech Recognition***
- Rank by Statistical Metrics similar to TF-IDF
  - finds terms more characteristic of foreground than background
- Rerank terms using Relevance Metric, based on a Yahoo Websearch
- Take Top N terms (e.g., N = 5000)

Statistical Metrics for Ranking Terms

- A linear combination of 3 Measures comparing the distribution of terms in the foreground (For) vs background (Bac)
- Term Frequency Inverse Document Frequency (TFIDF)
  - $TFIDF(t) = \frac{FreqFor(t)}{FreqBac(t)} \times \log(\frac{NumBacDocuments}{NumBacDocsContain(t)})$
- Document Relevance Document Consensus (DRDC)
  - Navigli and Velardi (2004)
  - $DRDC(t) = \frac{FreqFor(t)}{FreqBac(t)} \times \sum_{d \in Foreground} \frac{freqBac(t,d)}{freqFor(t)} \times \log(\frac{freqFor(t)}{freqBac(t,d)})$
  - Doc Relevance (1st factor) favors representative terms (like TFIDF)
  - Doc Censensus (2nd factor) favors terms found in many documents
- Kullback Leibler Divergence (KLD)
  - Cover and Thomas (1991), Hisamitsu, et. al. (1999)
  - $KLD(t) = (\log(freqFor(t)) - \log(freqBac(t))) \times freqFor(t)$
  - Compares Probability of term occurs in Foreground vs. Background
-

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. (Sometimes favors terms with OOV words over terms with other technical words and no OOVs)
- 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
- Relevance = H^2T
  - H = 0 to 1 score based on number of hits
    - $\frac{min(\log_{10}(numberHits), 10)}{10}$
    - Minimized for non-hits (0 hits counts as 500 hits)
  - T = Percent of top 10 hits that are articles or patents
    - Based on key word search in title, url & summary
      - Key words = {patent, article, proceedings, journal, dissertation, abstract, ...}

Evaluation of Termolator Output

- Foreground Corpus: 2500 patents about optical systems
  - US Patent codes: 250, 349, 356, 359, 362, 385, 398 and 399
- Background Corpus: 2500 randomly selected patents
- Years: 1997-2007
- Took the top 30K out of 219K terms and reranked using:
  - Percentile X Well-Formedness X Relevance
- Manually evaluated 100 terms sampled from top 5000 terms
- A term was judged correct if
  - valid keyword
  - not missing crucial modifier
  - did not contain any spurious word.
- **The system achieved 86% Precision**
- Recall difficult to measure, but also produces more high-quality terms
- Competitive with other systems (main innovation is: inline terms)

Evaluation Details

- Sample Correct (sampled from the first 5000):
  - ***stimulable phosphor, ion beam profile, x-ray receiver, wavelength-variable, quadrupole lens, proximity correction, dfb laser, asymmetric stress, panoramagram, single-mode optical fiber, total reflection plane, photosensitive epoxy resin***
- Sample Incorrect
  - ***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*]]

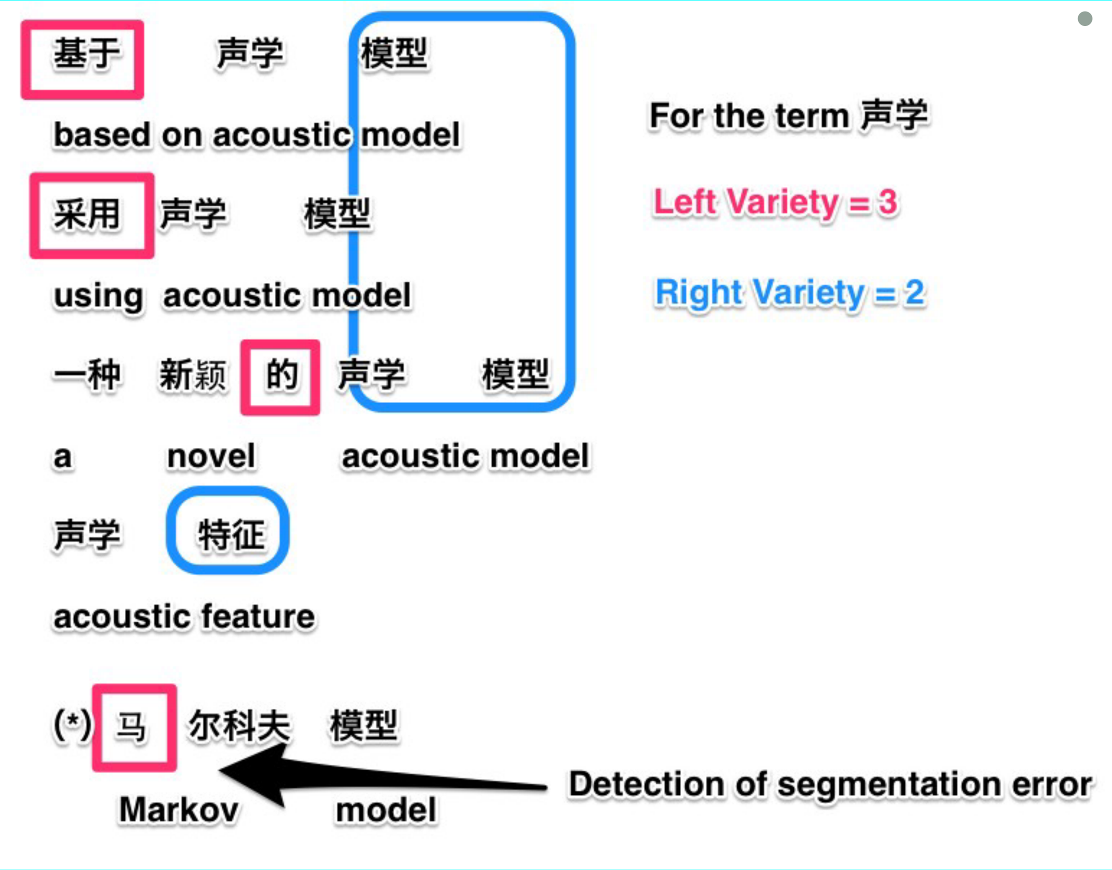
Informal Observations about Recall

- 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
- 2018-2019 research by students Y. Gu & L. Shi
  - Unifying components of English and Chinese System
  - Additional Chinese Components and Evaluation (in-progress)

Example of Chinese Term Filtering



- Examples for Access Variety based filtering
  - 科夫模型 (Markov model, with the first Chinese character 马 missing) is probably a boundary error
  - [Pic on left] 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
      - Future release planned of multi-lingual system
    - 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***

Subsequent Research on Termolator

Subsequent Work on Termolator

- Character-based language model: to eliminate unlikely paragraphs
  - Eliminates bibliographies, figures, noise
  - 5-gram model of sequences of, WHITESPACE, DIGIT. LETTER. PUNCTUATION, OTHER
- Merge Chinese and English system (Yuling Gu, Leizhen Shi, Echo Hong, Yuting Wang)
  - Some English components are not possible for Chinese
    - Use Brandeis noun chunker; no websearch filter
  - Chinese Character-based language model seems OK
- French system – preliminary, but all English components seem possible (Sandra Burland)
- System for getting foreground and background documents from **Wikipedia** (Sandra Burland, Anand Tyagi)

Current Work and Recent Papers

- Current Work:
  - Automatic Glossary for terminology output for English, French and Chinese, including evaluation of output
- Recent Papers
  - Tuning English system to court opinions
  - Evaluating the effect of different types of foreground and cibackground corpora

Termolator Text Front End and Glossaries

- Command Line Tool (from beginning of talk)
  - Check if completed
  - Running System
    - Keyword selection of Foreground from Wikipedia
    - Background based on Superclasses found in Wikipedia
- Creates "Glossary"

Glossary Continued

- Glossary of high score terms occurring in 3 or more documents
- Displays one “entry” for each term
  – Wikipedia First Paragraphs
  – 3 “diverse” sample paragraphs
- Examples:
  – Machine_learning – .edited_term_list and .summary
  – test_summaries
    - Category Theory (bk: all)
    - Printmaking (bk: "Artistic techniques" + "Visual arts media")

1st Paragraphs from Wikipedia

- If the whole string matches a Wikipedia Entry, that first paragraph is used, e.g., the first few matches for “Category Theory”
- If not, but substrings match Wikipedia, the substring 1st paragraphs are used, e.g., “anilox roll” under “Printmaking”

Selecting the Sample Usages

- Create vectors of TF-IDF scores for all the examples paragraphs containing the term.
- Compute a centroid vector – average of vectors.
- Form 3 clusters of vectors using, a simple clustering method:
  - iteratively merge most similar clusters one at a time until there are only 3 clusters left
- From each cluster, select the example that is the least similar to the centroid.

Glossary is Focus of Current Students

- Problem1: Further development of Chinese and French pipeline
- Problem2: Evaluation Method for English
- Problem3: Extend evaluation to Chinese and French.

Research Question: How Can We evaluate Glossaries?

- Some obvious things we can do first:
  - Evaluate sample 100 (or more) for Precision
    - Assume well-formed terms are correct
  - Evaluate sample 100 for Relevance
    - Of the correct terms, what percent is related to foreground topic?
- How do we measure recall for a topic?
  - Find previously created manual glossaries
  - Measure percent coverage
- How do we measure glossary text quality automatically?
  - Find existing online glossaries
  - Compare glossaries using existing NLG measures, e.g., BLEU or ROUGE
- Manual Measures of Glossary Utility
  - Likert Scales
    - Experts or non-experts?
    - What is the key question?
    - Inter-annotator agreement
- After English, how difficult would it be to replicate for French and Chinese?

Publications

Termolator Papers

- A. Meyers, Y. He, Z. Glass, J. Ortega, S. Liao, A. Grieve-Smith, R. Grishman and O. Babko-Malaya (2018). "*The Termolator: Terminology Recognition based on Chunking, Statistical and Search-based Scores.*" Frontiers in Research Metrics and Analytics
  - https://www.frontiersin.org/articles/10.3389/frma.2018.00019
- N. Pham, L. Pham and A. Meyers (2021) "Legal Terminology Extraction with the Termolator". NLLP-2021.
  - https://aclanthology.org/2021.nllp-1.16/
- S. Nordquist and A. Meyers (2022). "On Breadth Alone: Improving the Precision of Terminology Extraction Systems on Patent Corpora". NLLP-2022
- Code (a few updates are forthcoming)
  - http://nlp.cs.nyu.edu/termolator/
  - https://github.com/AdamMeyers/The_Termolator
  - https://github.com/ivanhe/termolator/
-

Tuning Termolator to the Court Decisions

N Pham, L. Pham and A. Meyers (2021)

- Data: Supreme Court Database (SCDB) from Washington University School of Law
  - Spaeth et. al. 2013 http://supremecourtdatabase.org
  - Can be downloaded through Python's Textacy library
- Tuned Termolator to work on Supreme Court Decisions
  - Manually annotated categories from Washington University (about 8.4 K files with topic categories)
  - **Eliminate noise**: Regular expressions to identify citations (to other decisions) and the names of legislation.
  - D**ifferent search engine** (Harvard Case Law Access instead of Yahoo)

Experiments

- Data – 8.4K documents
  - 14 “broad” and 279 “narrow” issues (topics)
  - Class size varies from 1 to 1924 documents
- Tested on larger foreground sets
  - Broad issue 1: Criminal Procedures, 1924 cases
  - Broad issue 8: Economic Activity, 1667 cases
  - Broad Issue 5” Privacy, 110 cases
  - Narrow issue 10050: Search and Seizure, 238 cases
    - Subtopic of Broad 1
  - Narrow issue 80010: Antitrust 216 issues
    - Subtopic of Broad 8

Precision for Broad Issue 8

- Baseline system: 23%
- Parameter adjustments: 35%
- Case/legislation filter: 44%
- Additional filter for digits and hyphens: 50%
- Legal Search Customization: 63%
  - Absolute Improvement over baseline: 40%
  - Relative Improvement: 274%

Precision for all 4 tests

Issue	Generality	Freq	Baseline	Final
Criminal 01	Broad	1924	25%	65%
Economic 08	Broad	1667	23%	63%
Privacy 05	Broad	110	27%	40%
Search & Seizure 10050	Narrow	238	19%	30%
Antitrust 80010	Narrow	216	13%	28%

- Narrower & High Frequency Classes have better results
- Domain customization improves results

Background Selection

- **Nordquist and Meyers (2022)**
- Some Terminology systems assume the same background corpus for all foregrounds
  - **We show that this might not be the optimal strategy**
- Cooperative Patent Classification
  - ontology of topic codes for patents
  - Possible to identify classes and superclasses
- For foreground F, experiments with different backgrounds going from the most similar to F (an immediate superclass) to the least similar (a mixture of patents and non patents).
- Example on next 2 slides

Background Sets for SemiConductor Patents

- General = combination of **Base** and OANC
- **Base** = Randomly selected patents
- H = Electricity
- H/01= Basic Electric Elements
- H/01/L = Semiconductor Devices
- H/01/L/21 = **Foreground Patents**: Processes or apparatus adapted for the manufacture or treatment of semiconductor or solid state devices or of parts thereof

Precision of Semiconductor patents with Different Backgrounds

- Results
  - H/01/L = .63
  - H/01 = .61
  - **H = .72**
  - **Base = .7**
  - General = .45
- Best results for a "sweet" spot (the right level of superclass).
- Different types of terminology, depending on the background (e.g., general patent terminology with general corpus, more specific for H/01 then H, etc.)
- Similar results with other topics

Summary

- Termolator is a system for identifying multi-word terminological expressions based on foreground and background sets of text.
- CPC patents and Wikipedia provide ontologies that make it easy to find foreground/background sets.
- Termolator can also generate automatic glossaries of terms
- Extensions to Chinese, French and other languages are possible, especially given the availability of classified sets of documents within Wikipedia.
- Finding Appropriate Evaluation Methods is an Important next step.