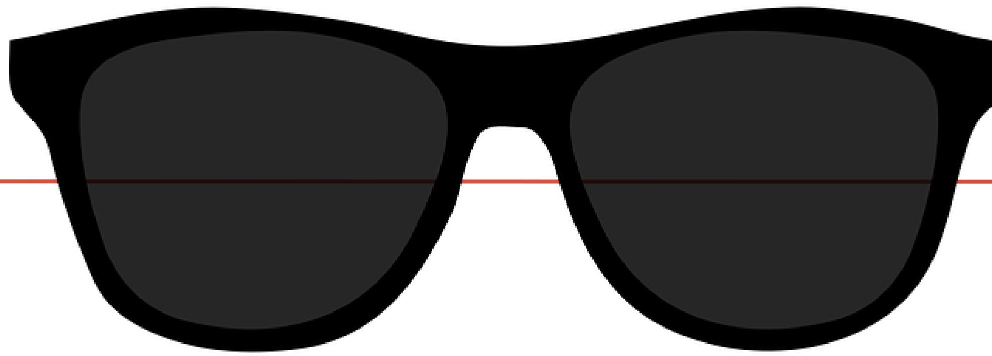
TERMINOLOGY IN INFORMATION EXTRACTION AND TECHNOLOGY FORECASTING



Adam Meyers

New York University

NYU

Funding and Collaboration

- Collaborators at NYU
  - Zachary Glass
  - Ralph Grishman
  - Yifan He
  - Giancarlo Lee
  - Shasha Liao
  - Angus Grieve-Smith
- NYU is a subcontractor of BAE under FUSE
  - FUSE is IARPA's Foresight and Understanding from Scientific Exposition program

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 a document—50 are instances of ***H5N1***
  - We use 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 a document—***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 uses for Terminology Forecasting
  - Our 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 ∨ BI* ∨ 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
…

Filters

Inline-Terms
semiconductor device
semiconductor chip
…

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**.*

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*$_{\text{NOUN/I}}$ *.*$_{\text{OTHER/O}}$

- 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**)*
  - *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 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
  - Determining if a term is specific to technical discipline

    - 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

				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

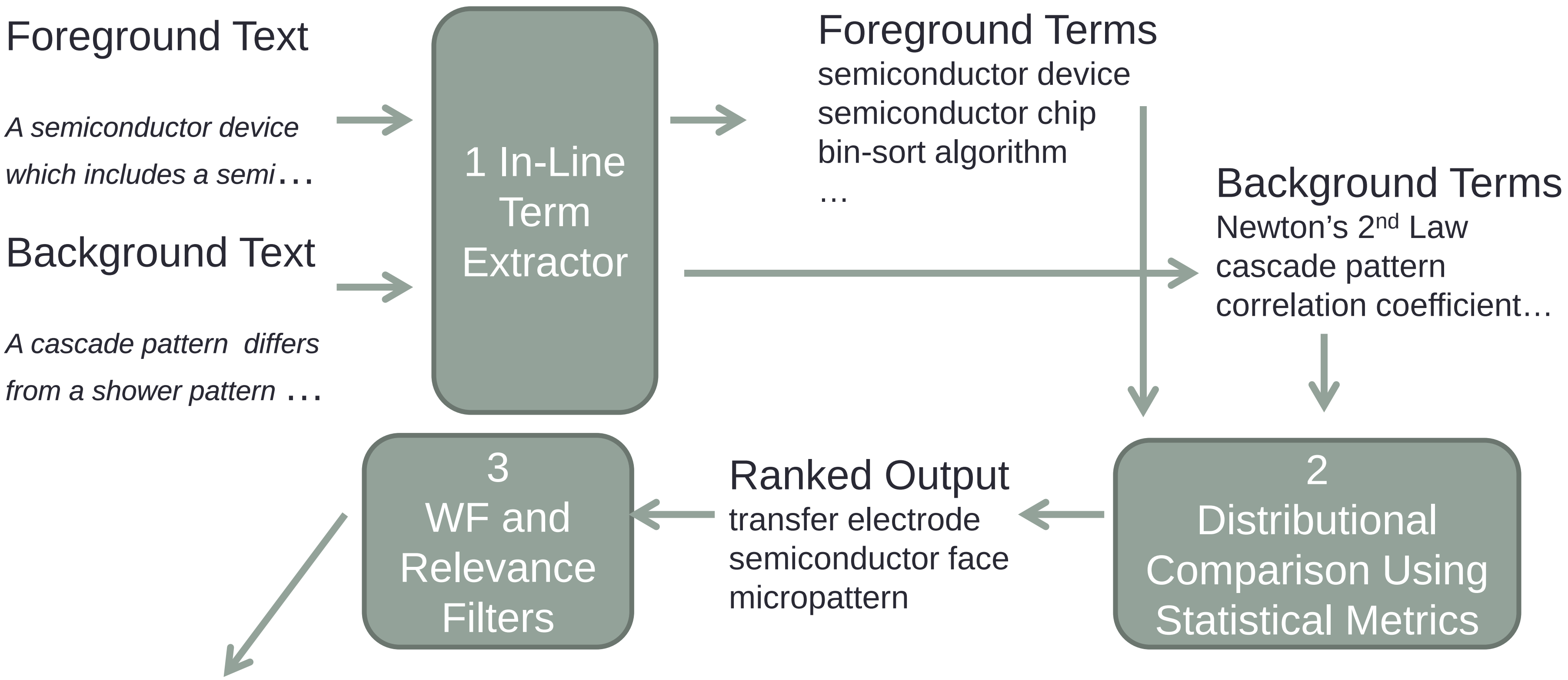
Term Types Used in Technology Forecasting

- **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)
- FUSE paper about TF: Babko-Malaya, et. al. (2015)

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



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)
  - $TFIDF(t) = \frac{freqFor(t)}{freqBac(t)} * \log(\frac{numBacDocs}{numBacDocContains(t)})$
- Document Relevance Document Consensus (DRDC)
  - **(Navigli and Velardi,** 2004)
  - DRDC$(t) = \frac{freqFor(t)}{freqBac(t)} * \sum_{d \in RDG} \frac{freq(t,d)}{freqFor(t)} * \log\left(\frac{freqFor(t)}{freq(t,d)}\right)$
  - 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$(t) = \log\big(freqFor(t)\big) - \log(freqBac(t)) * \mathrm{freqFor(t)}$
  - 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)
- A Yahoo search (Bing) for exact match of term
  - We use the free version, but would pay .18 cents per search using Yahoo's API (https://developer.yahoo.com/boss/search/)
- ***Relevance = H^2T***
  - ***H = Score representing number of hits***
    - $\frac{\min(\log_{10}(numberHits), 10)}{10}$
    - ***Minimized for nonhits (keyed by "including/showing results for")***
  - ***T = Percent of top 10 hits that are either patents or articles***
    - ***As determined by key word search on url, title and summary***
      - *patent, article, proceedings, journal, dissertation, abstract, …*

Evaluation of Distributional Term System

- 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
- 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

Evaluation Distribution System Slide 2

- 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 or contained any spurious word.

Example Evaluations

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

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*]]

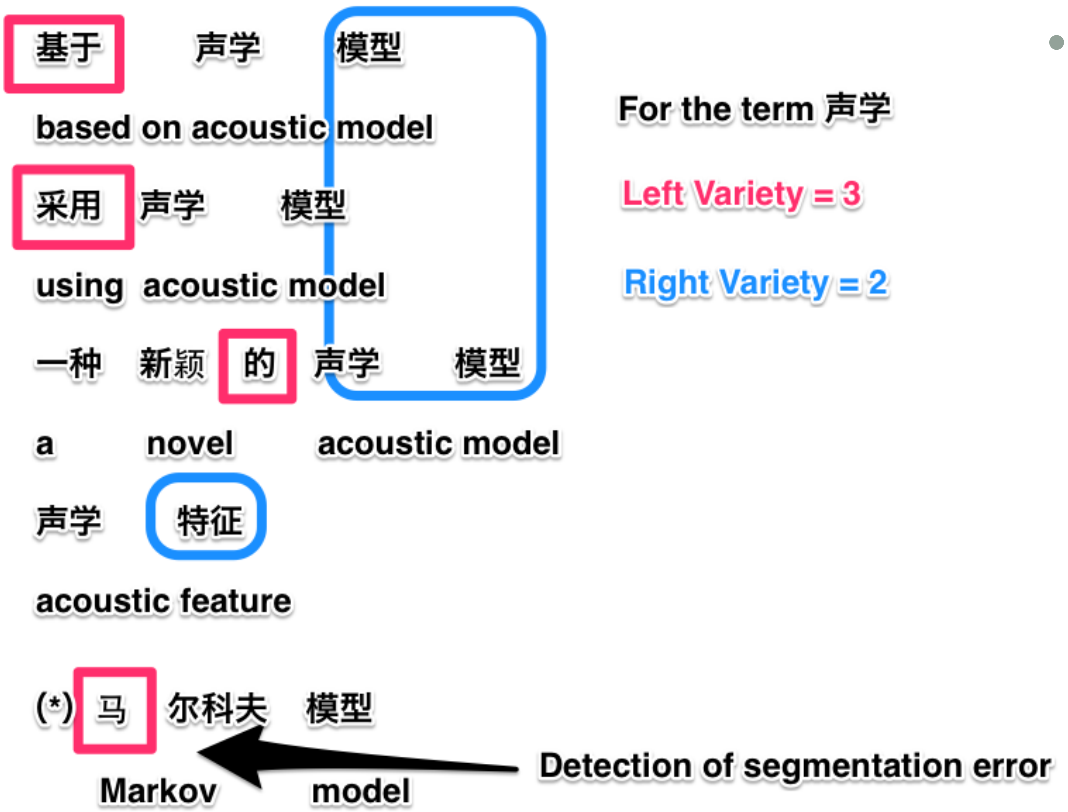
Informal Observations about Recall

- 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
- Coming Soon:
  - NYU's Website and Github
- Made to run on UTF-8 (including ASCII) and 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***

Our Papers on Terminology & NLP of Technical Literature

- A. Meyers, Y. He, Z. Glass and O. Babko-Malaya (2015). *The Termolator: Terminology Recognition based on Chunking, Statistical and Search-based Scores.* Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics.
- A. Meyers, , Z. Glass, A. Grieve-Smith, Y. He, S. Liao and R. Grishman (2014). *Jargon-Term Extraction by Chunking*. In Proceedings of COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language.
- A. Meyers , G. Lee, A. Grieve-Smith, Y. He and H. Taber (2014). *Annotating Relations in Scientific Articles*. In Proceedings of LREC 2014.
- Y. He and A. Meyers (2014). *Corpus and Method for Identifying Citations in Non-Academic Text*. In Proceedings of LREC 2014.
- A. Meyers (2013). *Contrasting and Corroborating Citations in Journal Articles*. In Proceedings of RANLP-2013.
- X. Li, Y. He, A. Meyers and R. Grishman (2013). *Towards Fine-grained Citation Function Classification*. In Proceedings of RANLP-2013.
- P. Thomas, O. Babko-Malaya, D. Hunter, A. Meyers and M. Verhagen (2013*). Identifying Emerging Research Fields with Practical Applications via Analysis of Scientific and Technical Documents*. In Proceedings of ISSI 2013.
- O. Babko-Malaya, A. Meyers, J. Pustejovsky and M. Verhagen (2013*). Modeling Debate within a Scientific Community*. In Proceedings of SOCIETY 2013.
- O. Babko-Malaya, P. Thomas, D. Hunter, A. Meyers, J. Pustejovsky, M. Verhagen, and G. Amis (2013). *Characterizing Communities of Practice in Emerging Science and Technology Fields*. In Proceedings of SOCIETY 2013.

Previous Work on Terminology Extraction

- Terms are most typically Noun Groups or Obey Other Linguistic Rules
  - K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. International Journal on Digital Libraries, 3(2):115–130.
  - Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.
- Comparing foreground & background documents to rank terms (many others)
  - Damerau, F. J. (1993). Generating and evaluating domain-oriented multiword terms from texts. *Information Processing and Management*, 29:433–447.
  - Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9: 99—115.
  - Navigli, R. and Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30.
  - Velardi, P., Missikoff, M., and Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies*. Workshop on Human Language Technology and Knowledge Management*.
- Finding Terminology via Relational Patterns
  - Y. Jin, M. Kan, J. Ng, and X. He. 2013. Mining scientific terms and their definitions: A study of the acl anthology. In EMNLP-2013.

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, analagous, 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

Previous POS	Current POS	Previous State	New State
	DET, PREP, POSS, OTHER		0
O-NOUN, C-NOUN, PLUR-NOUN	ROM-NUM	B-T, I-T	E-T
	PLUR-NOUN	B-T, I-T	I-T
	ADJ, CAP-ADJ	I-T	I-T
	NOUN, PER-NOUN, O-NOUN	B-T, I-T	I-T
O-NOUN	CAP-ADJ, TECH-ADJ, STAT-ADJ, NAT-ADJ	B-T, I-T	I-T
	CAP-ADJ, TECH-ADJ, NAT-ADJ, ING-VERB, ED-VERB, STAT-ADJ, NOUN, O-NOUN, PER-NOUN	E-T, O, S	B-T
TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ	TECH-ADJ, NAT-ADJ ADJ, CAP-ADJ	B-T, I-T	I-T
Else			O

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