Methods for Retrieving Alternative Contract Language Using a Prototype

Silviu Pitis
spitis@gatech.edu

Retrieval by Prototype

1

Given a prototype

"The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement."

2

Retrieve similar provisions

1. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

2. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

3. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

...

3

Organize the results effectively

The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

13 matches 100% View all

The Company will use ~~its best~~ commercially reasonable efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

26 matches, 94% View all

The Company will use its best efforts (I) if the Securities have been rated prior to the initial saleof such Securities, to confirm ~~that the~~such ratings ...

74 matches, 72% View all

...

Outline of Presentation

1

Given a prototype

- 3 Motivating Scenarios
  - Contract negotiation
  - Legal admin & due diligence
  - Education

2

Retrieve similar provisions

- Ranked Retrieval
  - What doesn't work
- Potential Approaches
- Empirical Comparison

3

Organize the results effectively

- Novelty Detection & Search Result Clustering
- Essential Features

About me

- J.D., Harvard Law School, 2014
- Licensed to practice law in New York
- Junior transactional lawyer @ Kirkland & Ellis, 2014-2016
  - Worked on public & private M&A, capital markets, and fund formation
  - Used spare time to learn programming & develop legal automation applications
- M.S. in Computer Science Candidate, Georgia Tech, 2016-2017
  - Currently working on deep learning, natural language processing and reinforcement learning

Goal for this project:

To make something that would have improved my life as a junior attorney.

Scenario #1: Contract Negotiation

- Counter-party proposes language that is either unfavorable or unfamiliar
- Very common problem
- Disagreements over language can bring negotiations to a halt

- Consider:

> "Material Adverse Effect" means any event ... that has a material adverse effect on ... the Company; provided, however, that none of the following ... will constitute ... a Material Adverse Effect:
>
> …
>
> (iv) a failure of the Company to meet any published or internally prepared projections, budgets, plans or forecasts of revenues, earnings or other financial performance measures or operating statistics,
>
> ...

If you are unfamiliar with the proposed carve-out, how do you respond?

Scenario #1: Contract Negotiation

- Would greatly benefit from a search function that can quickly and reliably identify the following added language:

> "Material Adverse Effect" means any event ... that has a material adverse effect on ... the Company; provided, however, that none of the following ... will constitute ... a Material Adverse Effect:
>
> …
>
> (iv) a failure of the Company to meet any published or internally prepared projections, budgets, plans or forecasts of revenues, earnings or other financial performance measures or operating statistics **(it being understood that the facts and circumstances underlying any such failure that are not otherwise excluded from the definition of a "Material Adverse Effect" may be considered in determining whether there has been a Material Adverse Effect)**,
>
> ...

Scenario #2: Legal Administration & Due Diligence

- **Investor contracts (fund formation)**:
  - 100s of investors with nearly identical provisions
  - Need to satisfy *Most Favored Nations* clause
- **Supplier contracts (in-house counsel perspective):**
  - 100s of suppliers with nearly identical provisions
  - Need to catalog licensing rights for business reasons
- **Supplier contracts (due diligence perspective):**
  - 100s of suppliers with nearly identical provisions
  - Need to catalog change in control provisions for business reasons, and to satisfy due diligence obligation

Scenario #3: Attorney Education

A few common questions that would otherwise require years of experience:

- What alternatives exist?
- Is this proviso common?
- Are these two clauses related (is it always the case that they show up together)?

Outline of Presentation

1

Given a prototype

- 3 Motivating Scenarios
  - Contract negotiation
  - Legal admin & due diligence
  - Education

2

Retrieve similar provisions

- Ranked Retrieval
  - What doesn't work
- Potential Approaches
- Empirical Comparison

3

Organize the results effectively

- Novelty Detection & Search Result Clustering
- Essential Features

Ranked Retrieval

- **Objective**: Retrieve *relevant* provisions by prototype.
- **Problem**: What is relevant? How to rank by relevance?

Ranked Retrieval

- **Objective**: Retrieve *relevant* provisions by prototype.
- **Problem**: What is relevant? How to rank by relevance?

Roughly we want:

- Exact phrase matches first
- Partial phrase matches next
- Topical matches (shares topic/words, but not phrases)

Ranked Retrieval

Doesn’t quite work	**Why not?**
Phrase match	• Not useful: lawyer must carefully craft query

Ranked Retrieval

Doesn't quite work	**Why not?**
Phrase match	• Not useful: lawyer must carefully craft query
Generic ranked retrieval (TF-IDF) at the document level	• Scores results at document-level • Focus on individual, unordered terms cannot capture partial phrase matches

Aside:
How (Most) Generic Search Engines Work

Term Frequency

How often term appears in the document

Inverse Document Frequency
(think of as <u>overall rarity</u>)

How rarely term appears in *all documents*

$$\text{Document score} = \sum_{\substack{\text{query}\\\text{terms}}} \text{TF} * \text{IDF}$$

Ranked Retrieval

Doesn't quite work	**Why not?**
Phrase match	• Not useful: lawyer must carefully craft query
Generic ranked retrieval (TF-IDF) at the document level	• Scores results at document-level • Focus on individual, unordered terms cannot capture partial phrase matches
Duplicate detection (e.g., using shingling) Approximate string matching (e.g., indexing character n-grams)	• Will find closest (top) results, but: • Focus is too narrow and may omit relevant paraphrased language
Clause-level database / model precedent	• Requires ex ante decision over what clauses to index (does not support ex post queries) • Difficult to determine boundaries

Ranked Retrieval: What Might Work

- **TF-IDF on a bigram index** (cf. Song & Croft (1999))
  - Instead of searching based on single words, using word pairs (bigrams)
  - Bigrams capture word combinations and query structure
  - **Concern:** requires re-indexing; bigrams are sparse, producing large indices

- **TF-IDF passage retrieval** (e.g., Kaszkiel & Zobel (2001))
  - Score results as passages, instead of as documents
  - **Concern:** Passage-level scoring is much more computationally expensive

- **Heuristic measures** (e.g., Tao & Zhai (2007))
  - Augment TF-IDF with ad-hoc heuristic scores that capture proximity and structure
  - Propose two novel heuristic measures, both at the *document-level*
  - **Concern:** Heuristic

First Novel Heuristic: Position-adjusted Minimum Distance

- Inspired by Tao & Zhai's (2007) *MinDist*
- Heuristic bonus when *pairs* of consecutive query terms appear close to each other in a document, based on their ordered distance

First Novel Heuristic: Position-adjusted Minimum Distance

- Inspired by Tao & Zhai's (2007) *MinDist*
- Heuristic bonus when *pairs* of consecutive query terms appear close to each other in a document, based on their ordered distance
- Example computation:
  - **Query**: "material adverse effect"
  - **Document**: "An adverse material effect resulted from the adverse material condition."

  - Max bonus (for adjacent terms): 3
  - Penalty for each offset: 1
  - Sum score over all pairs of consecutive terms:
    - "material adverse":
    - "adverse effect":
  - **Total heuristic score:**

First Novel Heuristic: Position-adjusted Minimum Distance

- Inspired by Tao & Zhai's (2007) *MinDist*
- Heuristic bonus when *pairs* of consecutive query terms appear close to each other in a document, based on their ordered distance
- Example computation:
  - **Query**: "material adverse effect"
  - **Document**: "An adverse material effect resulted from the adverse material condition."

  - Max bonus (for adjacent terms): 3
  - Penalty for each offset: 1
  - Sum score over all pairs of consecutive terms:
    - "material adverse": 3 – 2 = 1
    - "adverse effect":
  - **Total heuristic score:**

First Novel Heuristic: Position-adjusted Minimum Distance

- Inspired by Tao & Zhai's (2007) *MinDist*
- Heuristic bonus when *pairs* of consecutive query terms appear close to each other in a document, based on their ordered distance
- Example computation:
  - **Query**: "material adverse effect"
  - **Document**: "An adverse material effect resulted from the adverse material condition."
  - Max bonus (for adjacent terms): 3
  - Penalty for each offset: 1
  - Sum score over all pairs of consecutive terms:
    - "material adverse": 3 – 2 = 1
    - "adverse effect": 3 – 1 = 2
  - **Total heuristic score: 3**

Second Novel Heuristic: Max ascending *m*-cover

- Inspired by cover and span-based heuristics (see, e.g., Clarke et al. (2000))
- Heuristic bonus for not-necessarily adjacent *ordered sequences* of query terms that appear within (2 * query length) terms of each other in the document

Second Novel Heuristic: Max ascending *m*-cover

- Inspired by cover and span-based heuristics (see, e.g., Clarke et al. (2000))
- Heuristic bonus for not-necessarily adjacent *ordered sequences* of query terms that appear within (2 * query length) terms of each other in the document
- Example computation:
  - **Query**: “material adverse effect”
  - **Document**: “An adverse material effect resulted from the adverse material condition, but the effect was carved out of the contract.”

  - Many 2-covers (two examples shown above)

Second Novel Heuristic: Max ascending *m*-cover

- Inspired by cover and span-based heuristics (see, e.g., Clarke et al. (2000))
- Heuristic bonus for not-necessarily adjacent *ordered sequences* of query terms that appear within (2 * query length) terms of each other in the document
- Example computation:
  - **Query**: “material adverse effect”
  - **Document**: “An adverse material effect resulted from the adverse material condition, but the effect was carved out of the contract.”

  - Many 2-covers
  - One 3-cover, but excluded because covers more than (2 * query length = 6) terms

  - **Total heuristic score: 2**

Empirical Comparison: Setup

- **Dataset:** 20,236 publicly available contracts filed with the SEC
- **Queries:** 20 diverse, complete or partially complete contract provisions
- **Methods compared:** 4 methods compared, each on a unigram & bigram index

Unigram Index:	1 Document-level BM25	2 Passage-level BM25	3 Position-adj. min distance	4 Max ascending *m*-cover
Bigram Index:	5 Document-level BM25	6 Passage-level BM25	7 Position-adj. min distance	8 Max ascending *m*-cover

- **Metric:** Normalized discounted cumulative gain (nDCG) for the <u>top 10 results</u>
- **Scoring:** Results for all methods aggregated, then scored blind by hand according to five relevance categories0

Empirical Comparison: Results

Method	nDCG
Document-level BM25 (unigram)	0.613 ± 0.106
Document-level BM25 (bigram)	0.953 ± 0.030
Passage retrieval (unigram)	0.929 ± 0.057
Passage retrieval (bigram)	0.990 ± 0.007
Position-adj. min dist. (unigram)	0.950 ± 0.027
Position-adj. min dist. (bigram)	0.989 ± 0.011
Max ascending m-cover (unigram)	0.945 ± 0.034
Max ascending m-cover (bigram)	0.977 ± 0.021

Table 2: nDCG by retrieval method ($\pm$ 95% confidence)

Empirical Comparison: Remarks

Remarks

- **Best:** Passage retrieval on bigram index
  - **But**: slower, requires bigram index
- Position-adj min dist faster (my implementations), performs better on unigram index
- Document-level BM25 fastest, may have sufficient precision and speed to use as primary search, and then rerank top results

Method	nDCG
Document-level BM25 (unigram)	0.613 ± 0.106
Document-level BM25 (bigram)	0.953 ± 0.030
Passage retrieval (unigram)	0.929 ± 0.057
Passage retrieval (bigram)	0.990 ± 0.007
Position-adj. min dist. (unigram)	0.950 ± 0.027
Position-adj. min dist. (bigram)	0.989 ± 0.011
Max ascending m-cover (unigram)	0.945 ± 0.034
Max ascending m-cover (bigram)	0.977 ± 0.021

Table 2: nDCG by retrieval method (± 95% confidence)

Limitations

- Comparison focused on top 10 results, does not reflect recall of methods

Outline of Presentation

1

Given a prototype

- 3 Motivating Scenarios
  - Contract negotiation
  - Legal admin & due diligence
  - Education

2

Retrieve similar provisions

- Ranked Retrieval
  - What doesn't work
- Potential Approaches
- Empirical Comparison

3

Organize the results effectively

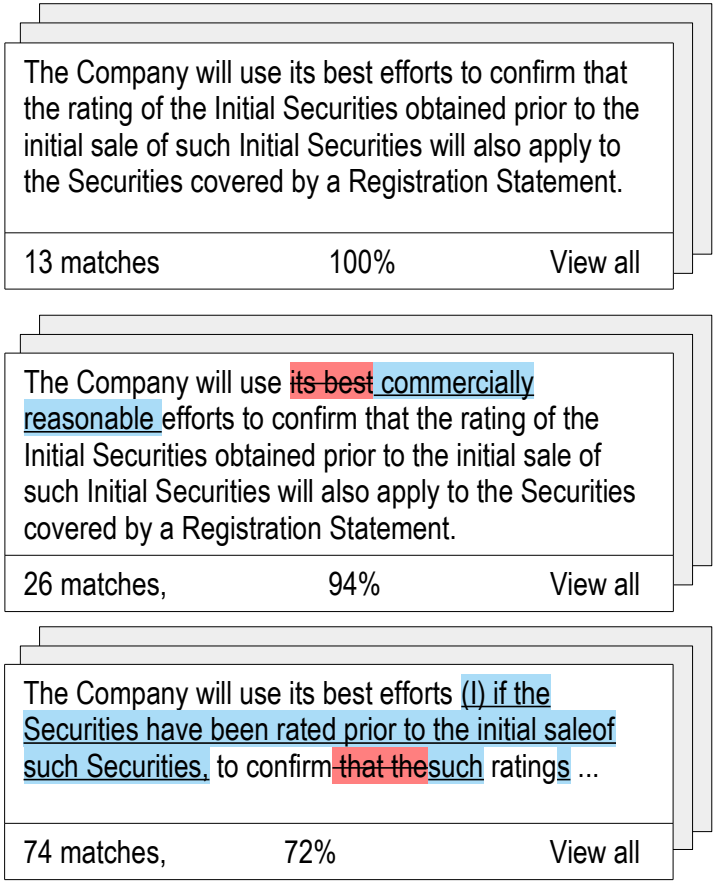
- Novelty Detection & Search Result Clustering
- Essential Features

Dynamic Result Clustering

- **Objective**: Hide redundant results
- **Problem 1:** Eliminating exact text reuse does not eliminate redundancy.
- **Problem 2:** Exact copies relevant, cannot eliminate!
- **Solution**:

Dynamic Result Clustering

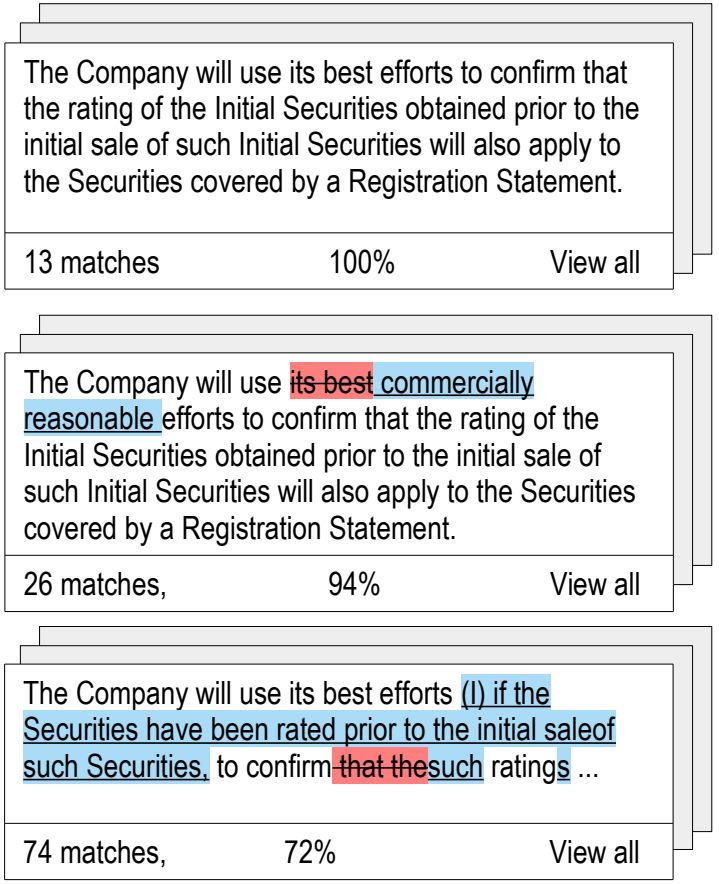
- **Objective**: Hide redundant results
- **Problem 1:** Eliminating exact text reuse does not eliminate redundancy.
- **Problem 2:** Exact copies relevant, cannot eliminate!
- **Solution**: Group redundant results into clusters



Dynamic Result Clustering

- **Objective**: Hide redundant results
- **Problem 1:** Eliminating exact text reuse does not eliminate redundancy.
- **Problem 2:** Exact copies relevant, cannot eliminate!
- **Solution**: Group redundant results into clusters

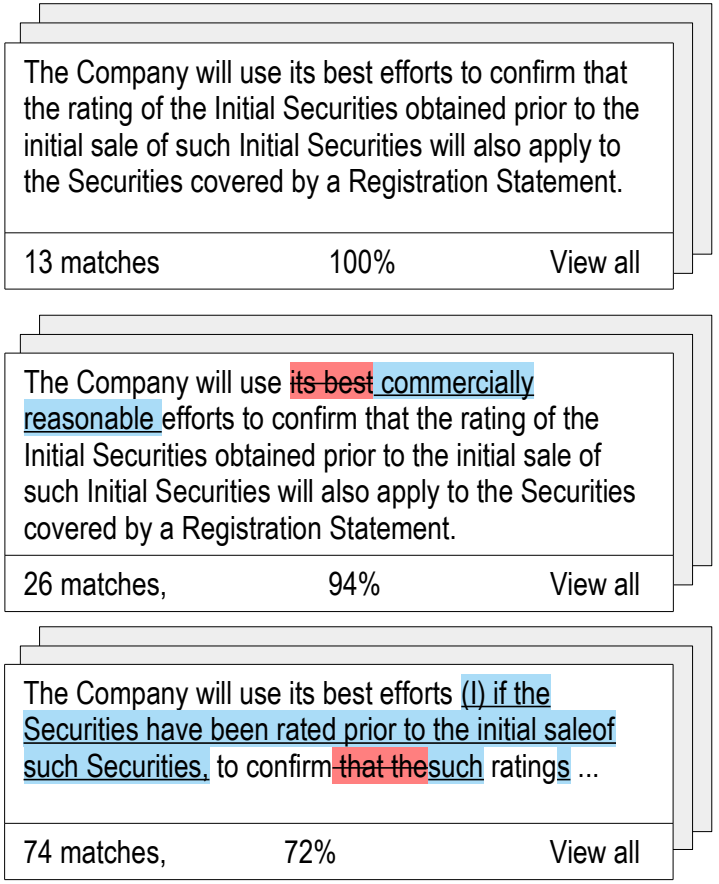
- **Implementation:**
  - Initialize first cluster with first result
  - For each result:
    - For each cluster:
      - If difference between result and cluster (as defined by some Δ function) is smaller than some threshold, group them



...

Dynamic Result Clustering

- **Objective**: Hide redundant results
- **Problem 1:** Eliminating exact text reuse does not eliminate redundancy.
- **Problem 2:** Exact copies relevant, cannot eliminate!
- **Solution**: Group redundant results into clusters



- **Implementation:**
  - Initialize first cluster with first result
  - For each result:
    - For each cluster:
      - If difference between result and cluster (as defined by some Δ function) is smaller than some threshold, group them

- **Problems:**
  - What threshold to use?
  - What Δ function to use?

Dynamic Result Clustering

- **Problem**: What threshold to use? What is redundant?
- **Nuance**: Same search can have different intent --- sometimes single words or even punctuation will matter; sometimes lawyers are looking for entire clauses.
- **Solution**:

Dynamic Result Clustering

- **Problem**: What threshold to use? What is redundant?
- **Nuance**: Same search can have different intent --- sometimes single words or even punctuation will matter; sometimes lawyers are looking for entire clauses.
- **Solution**: Dynamic (i.e., *user-tunable*) clustering

Δ = 5%

The Company will use ~~its best~~ [commercially] reasonable efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

26 matches, 94-97% View all

Δ = 2%

The Company will use ~~its best~~ reasonable efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

4 matches, 97% View all

The Company will use ~~its best~~ commercially reasonable efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

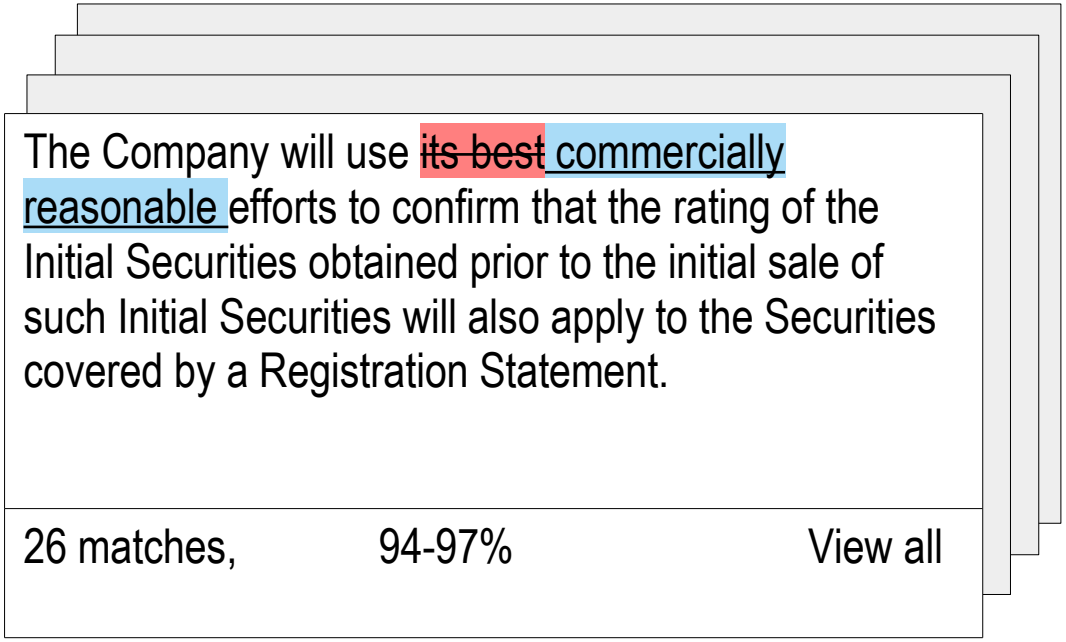
22 matches, 94% View all

Dynamic Result Clustering

- **Problem**: What Δ (difference) function to use?
- **Nuance**: Lawyer must be able to reason about what could be contained in clusters.
  (can click into cluster, or adjust threshold, but time consuming)
- **Example**:

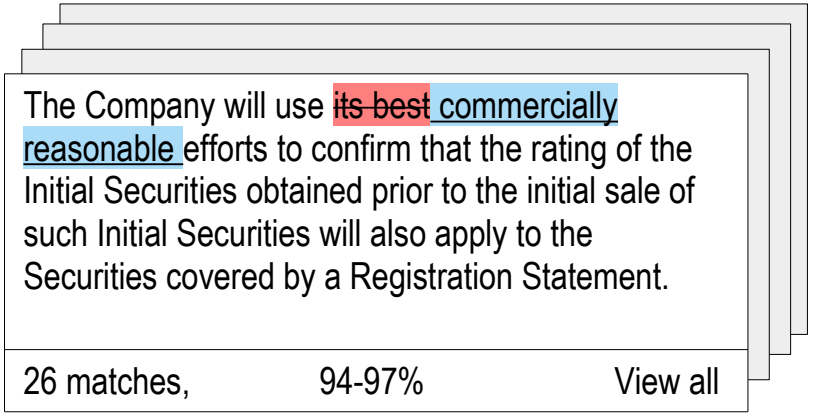
Dynamic Result Clustering

- **Problem**: What Δ (difference) function to use?
- **Nuance**: Lawyer must be able to reason about what could be contained in clusters.
  (can click into cluster, or adjust threshold, but time consuming)
- **Example**: What is the range of the 3%?



Dynamic Result Clustering

- **Problem**: What Δ (difference) function to use?
- **Nuance**: Lawyer must be able to reason about what could be contained in clusters.
  (can click into cluster, or adjust threshold, but time consuming)
- **Example**: What is the range of the 3%?



- **Solution:** Use Δ function that provides an *interpretable guarantee.*
- **Example:** Edit distance (integer)
  - Either at character or word level
  - If threshold == 5 characters,
    easy to see that “use reasonable efforts” not included in the cluster.

Retrieval by Prototype

1

Given a prototype

"The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement."

2

Retrieve similar provisions

1. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

2. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

3. The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

...

3

Organize the results effectively

The Company will use its best efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

13 matches 100% View all

The Company will use ~~its best~~ commercially reasonable efforts to confirm that the rating of the Initial Securities obtained prior to the initial sale of such Initial Securities will also apply to the Securities covered by a Registration Statement.

26 matches, 94% View all

The Company will use its best efforts (I) if the Securities have been rated prior to the initial saleof such Securities, to confirm ~~that the~~such ratings ...

74 matches, 72% View all

...