Applying the Seed-and-Extend Strategy to Text-Alignment

Master’s Defense
What is Text-Alignment?
What is Seed-and-Extend?

Plain Documents

The quick brown fox jumps over the lazy dog

A fast brown fox jumps over a sluggish pup
What is Seed-and-Extend?

Plain Documents | Sequences

Segmenting

d | d'

- the quick brown fox jumps over the lazy dog
- the quick fox jumps over the lazy dog
- a fast brown fox jumps over a sluggish pup
- a fast brown fox jumps over a sluggish pup
What is Seed-and-Extend?

Segmenting

Plain Documents

Seeding

Sequences

Matching seeds

the quick brown fox jumps over the lazy dog

brown fox
jumps over
fox jumps

brown fox
jumps over
fox jumps

a fast brown fox jumps over a sluggish pup
What is Seed-and-Extend?

Plain Documents → Segmenting → Sequences → Seeding → Matching seeds → Extension → Aligned passages

the quick **brown fox jumps over** the lazy dog

a fast **brown fox jumps over** a sluggish pup
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Model of Seeding

Plain Documents -> Seeding -> Matching seeds
Model of Seeding

- i.e. whitespace tokenizer, sentence or paragraph splitter
Model of Seeding

Plain Documents

Tokenizer

Filter

- i.e. wordlist filter, POS-tag filter

Seeding

Matching seeds
Model of Seeding

- Plain Documents
  - Tokenizer
  - Filter
  - Extractor

- Seeding

- Matching seeds

- i.e. plaintext, hypernyms, frequency or word vectors
Model of Seeding

Plain Documents → Seeding → Matching seeds

- i.e. n-grams or skip-grams
Model of Seeding

- i.e. exact or set match, Jaccard or cosine similarity
Model of Seeding

Plain Documents  \(d\),  \(d'\)  →  Seeding  →  Matching seeds
Model of Seeding

Plain Documents

Matching seeds

The quick brown fox jumps over the lazy dog

A fast brown fox jumps over a sluggish pup

\( S^\text{min}_d \quad S^\text{max}_d \quad S^\text{min}_{d'} \quad S^\text{max}_{d'} \)
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Seeder combination
Seeder combination

Plain Documents

Matching seeds
Seeder combination
Seeder combination
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Relaxation

the quick brown fox jumps over the lazy dog

a fast brown fox jumps over a sluggish pup

the quick brown fox jumps over the lazy dog

a fast brown fox jumps over a sluggish pup
Relaxation

\[ \Phi_1 \subseteq \Phi_2 \]

most specific seeder

most general seeder
Relaxation

most specific seeder

most general seeder
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Model of Extension
Model of Extension
Model of Extension
Model of Extension
Model of Extension
Model of Extension

Matching seeds

Extension

Aligned passages

Core Point: if at least \texttt{minPts} other points within \texttt{epsilon} radius

Border Point: within \texttt{epsilon} of a core point

Noise Point: if neither core nor border point
Model of Extension

\[ \phi_1 \]

- t: whitespace
- f: words
- e: text
- z: 5-grams
- m: exact

\[ \begin{array}{c}
0.5 \\
0.4 \\
0.8 \\
\end{array} \]

\[ \begin{array}{c}
F1 \\
rec. \\
prec. \\
\end{array} \]

epsilon 100

epsilon 500

epsilon 1000

minPts 2

minPts 4

minPts 6
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Hyperparameter estimation

- Collection estimation
  - Given the seeder, find the best parameters for all documents
  - Can be determined once with the ground truth

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>collection estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.54</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Hyperparameter estimation

- **Collection estimation**
  - Given the seeder, find the best parameters for all documents
  - Can be determined once with the ground truth

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Collection estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.54</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Hyperparameter estimation

- Collection estimation
  - Given the seeder, find the best parameters for all documents
  - Can be determined once with the ground truth

- Document estimation
  - Given the seeder and a pair of documents, find the best parameters for that particular pair
  - This can be learned
Hyperparameter estimation

- Determine the truth (best parameters) for a pair of documents via gridsearch
- Use length-frequency-histogram as features
- Get 10,000 examples per seeder from the PAN corpora
Hyperparameter estimation

- Determine the truth (best parameters) for a pair of documents via gridsearch
- Use length-frequency-histogram as features
- Get 10,000 examples per seeder from the PAN corpora

<table>
<thead>
<tr>
<th></th>
<th>prediction</th>
<th>collection estimate</th>
<th>per document optima</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi_1, \varphi_2 \rightarrow \varphi_1 \cup \varphi_2$</td>
<td>0.70</td>
<td>0.69</td>
<td>0.86</td>
</tr>
<tr>
<td>$\varphi_1, \varphi_3 \rightarrow \varphi_1 \cup \varphi_2 \cup \varphi_3$</td>
<td>0.65</td>
<td>0.69</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Contributions

- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation
Future Work

- Model of seeding
- Seeder combination
- Relaxation
- **Best Algorithm?**

- Model of extension
- Parameter estimation
- **Improved parameter learning**
Future Work

- Model of seeding
- Seeder combination (Breeding)
- Relaxation (Mutation)
- Best Algorithm
- Genetic optimization

- Model of extension
- Parameter estimation
- Improved parameter learning
Thanks!

- Model of seeding
- Seeder combination (Breeding)
- Relaxation (Mutation)
- Best Algorithm
- Genetic optimization
- Model of extension
- Parameter estimation
- Improved parameter learning