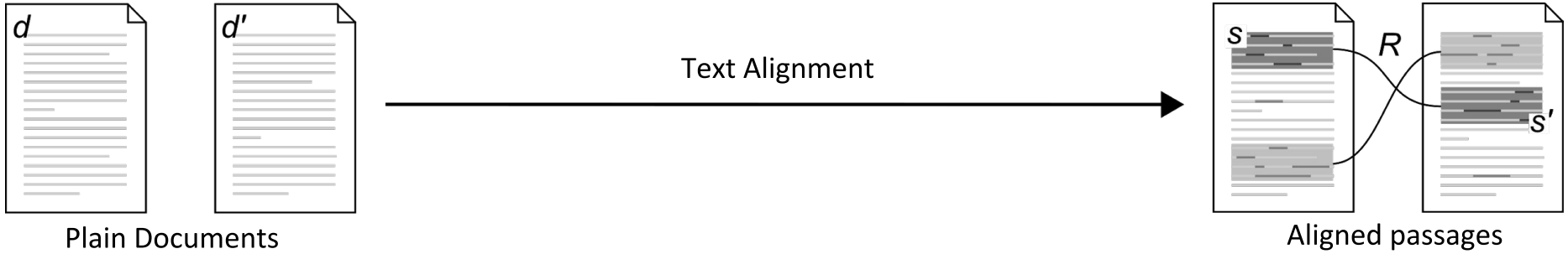
Applying the Seed-and-Extend Strategy to Text-Alignment

Master's Defense

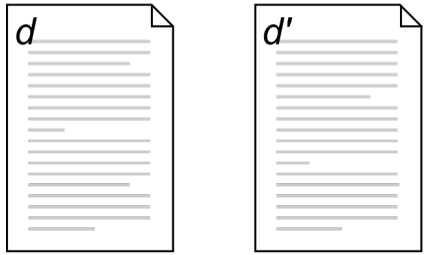
Matti Wiegmann
weimar.webis.de

Bauhaus-Universität
Weimar

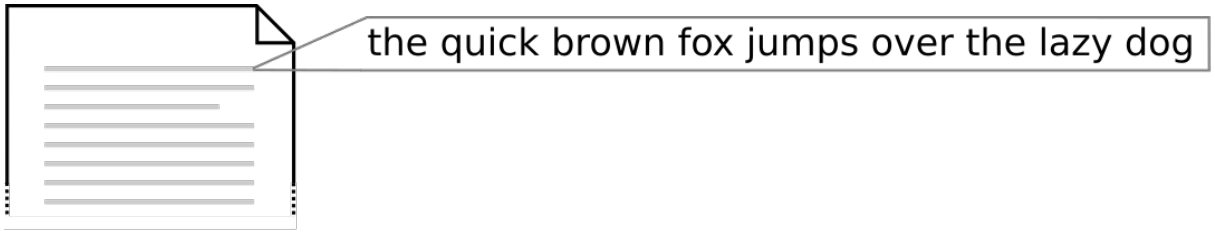
What is Text-Alignment?



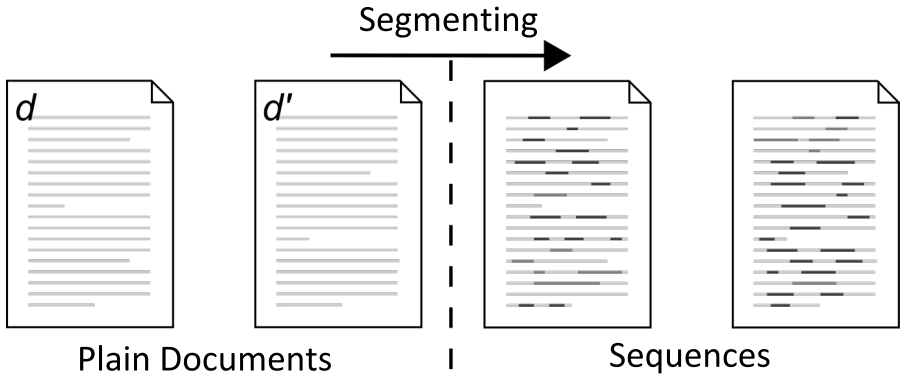
What is Seed-and-Extend?

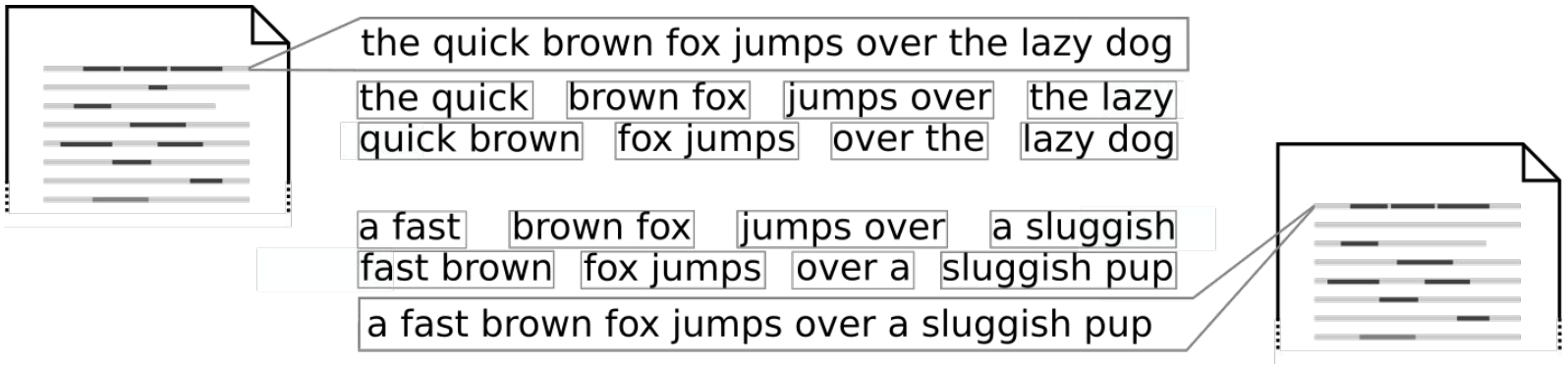


Plain Documents

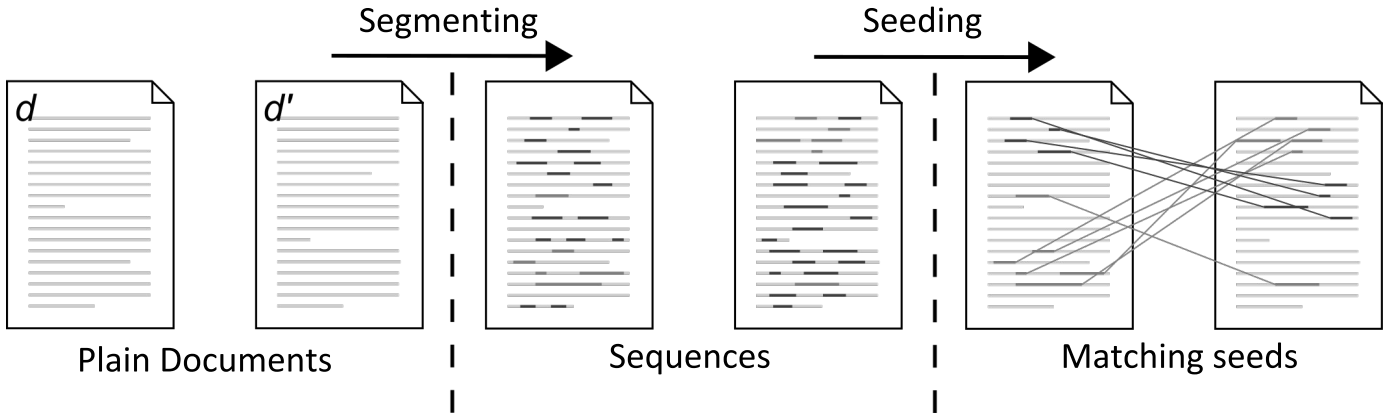


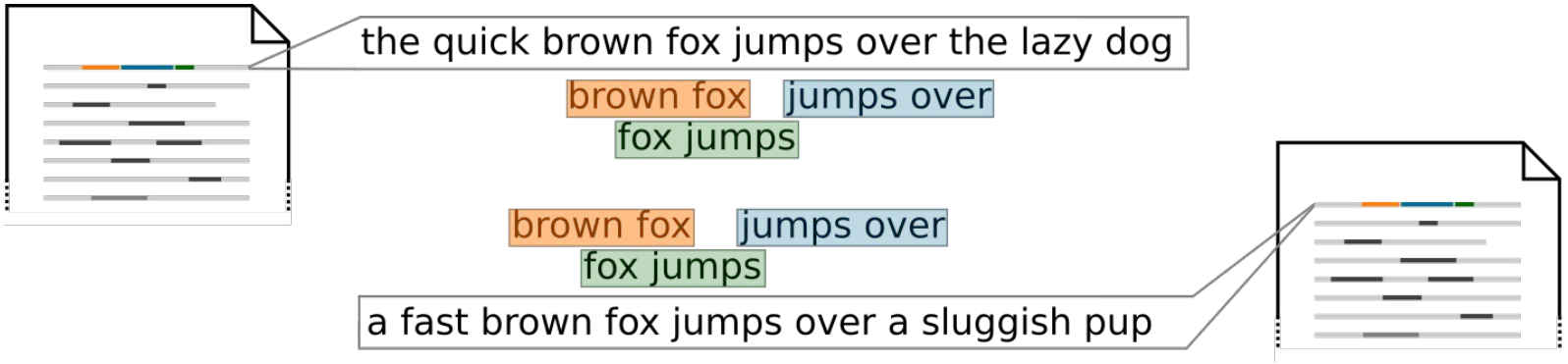
What is Seed-and-Extend?



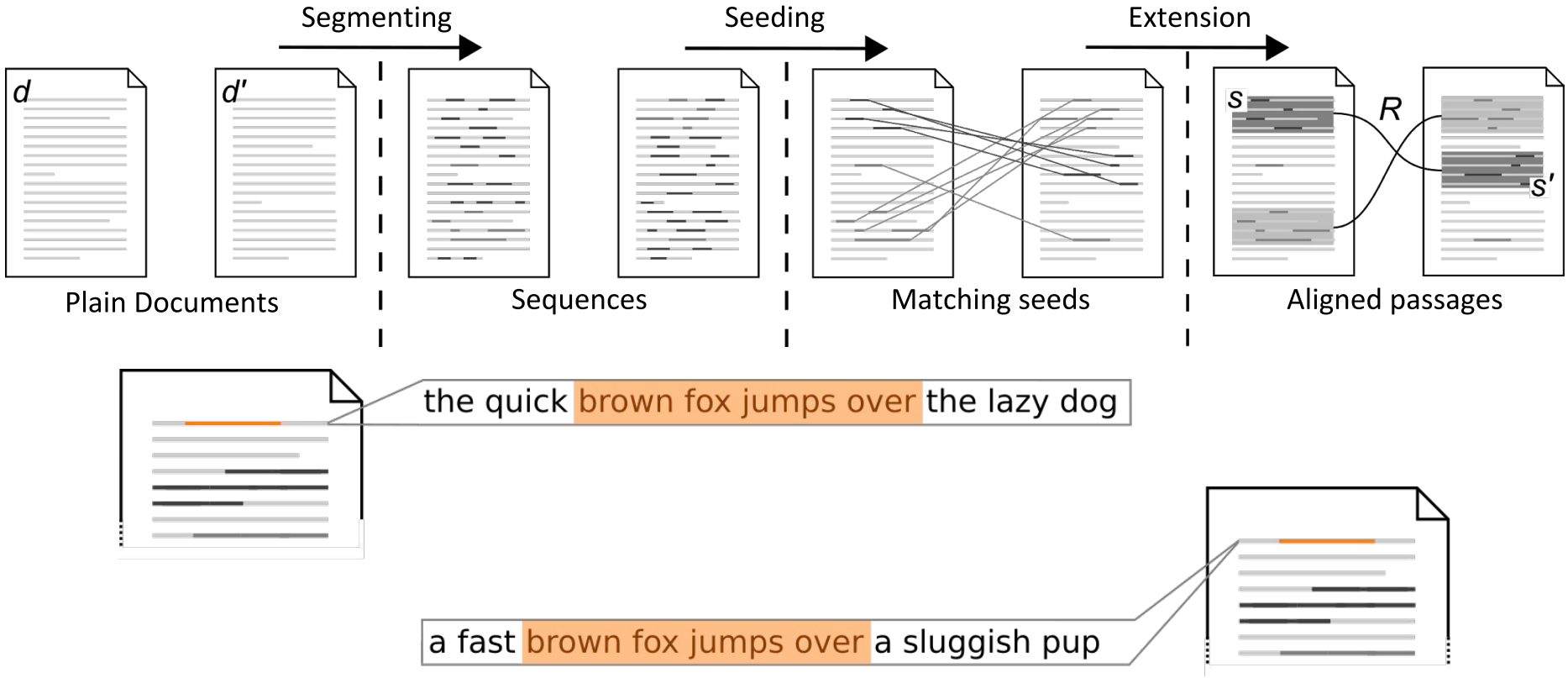


What is Seed-and-Extend?

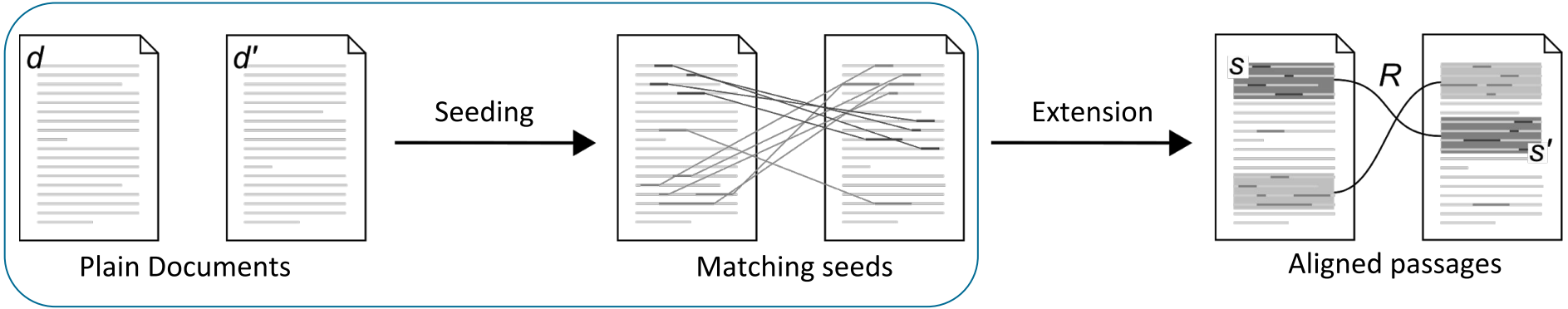




What is Seed-and-Extend?

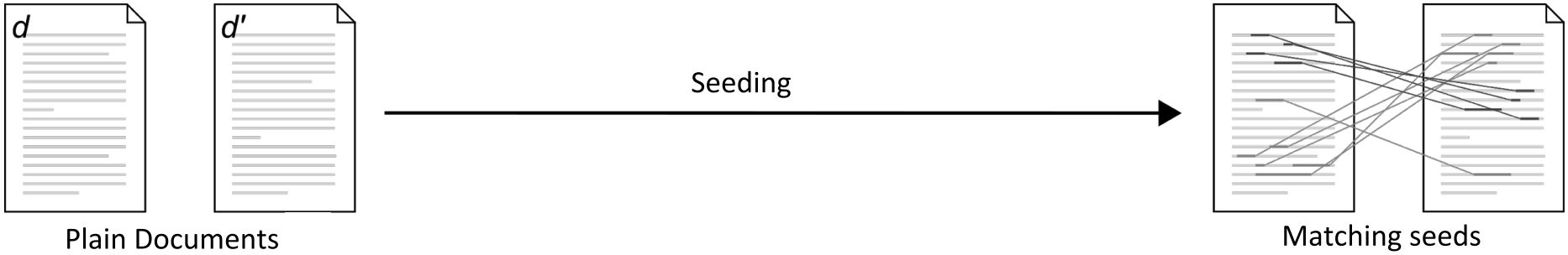


Contributions

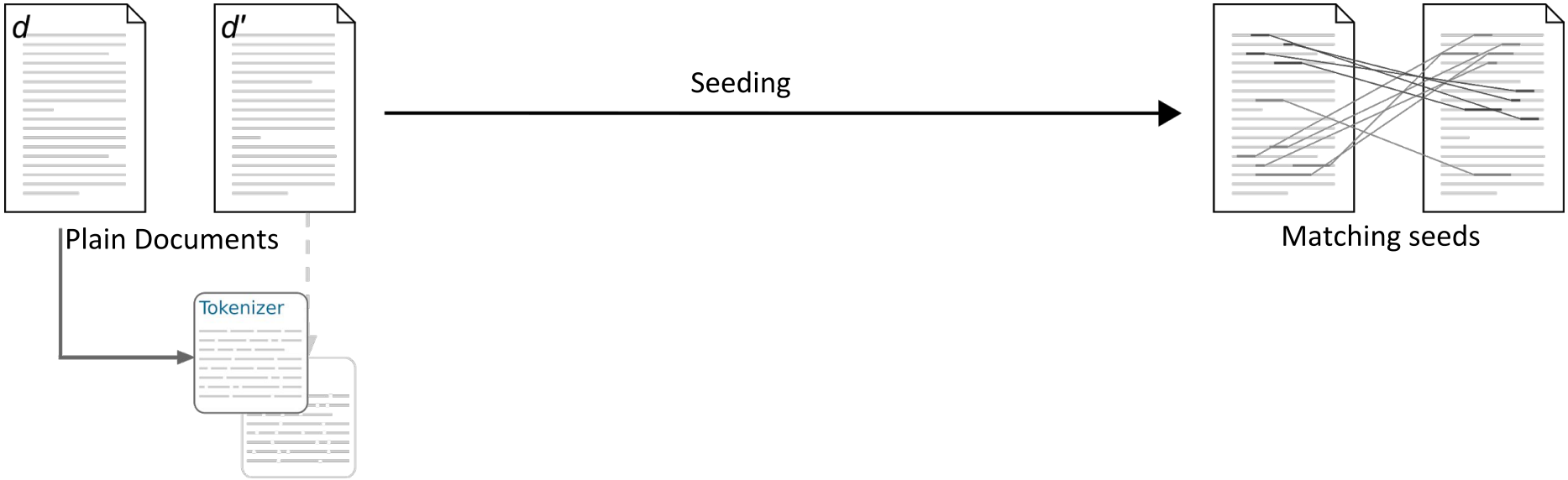


- Model of seeding
- Seeder combination
- Relaxation
- Model of extension
- Parameter estimation

Model of Seeding

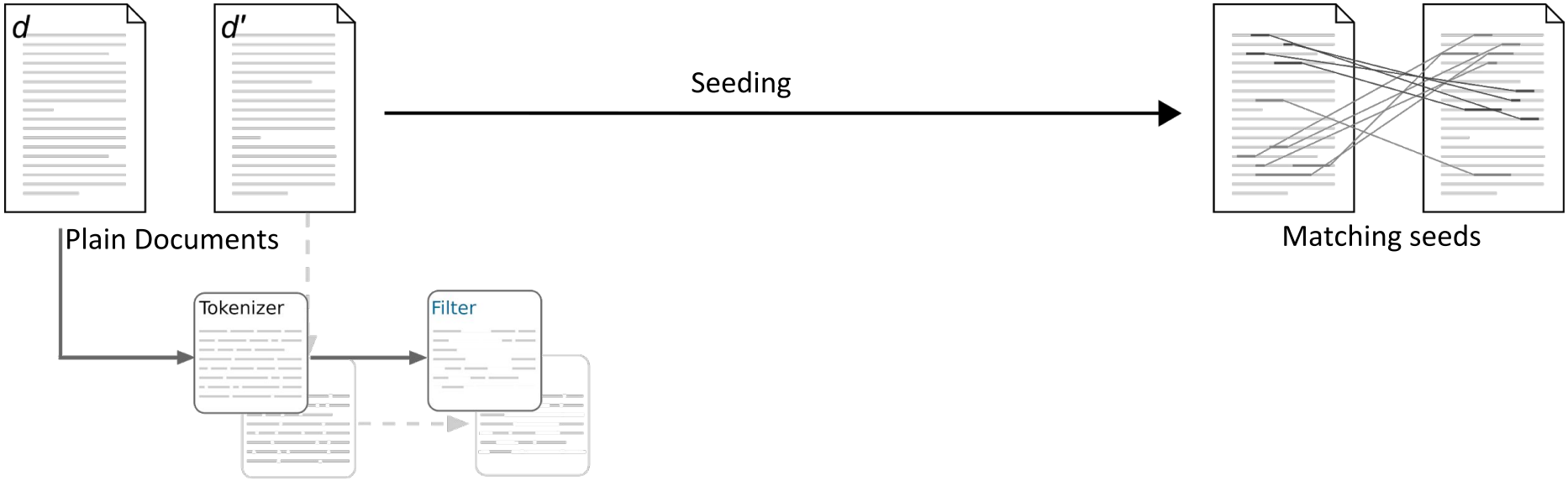


Model of Seeding



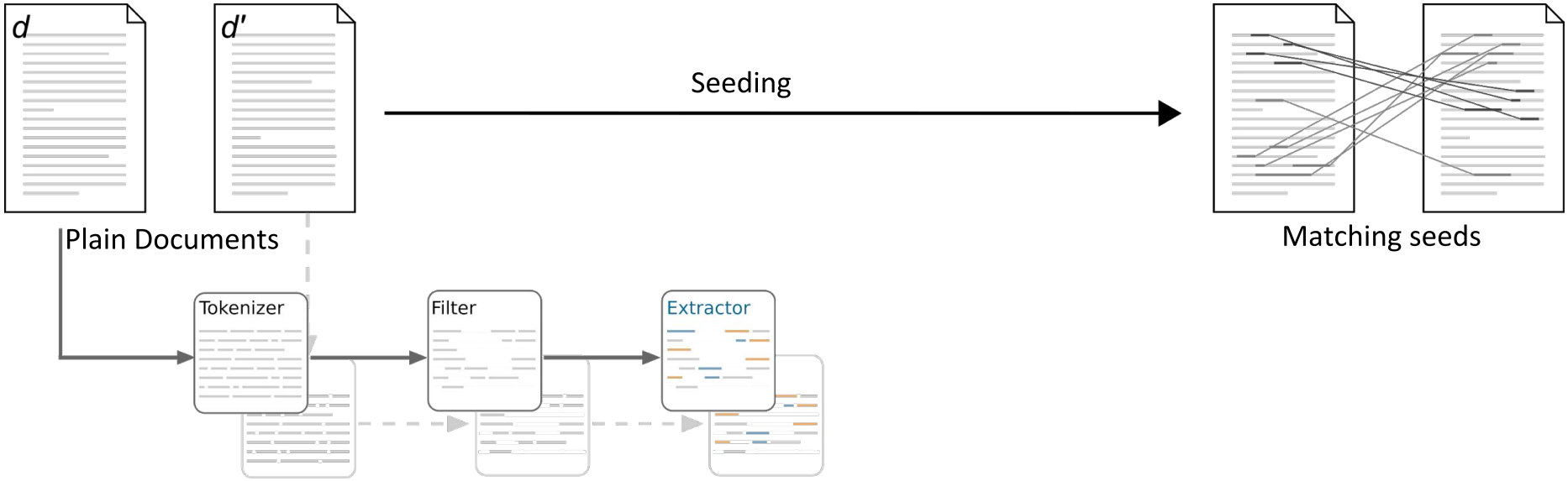
- i.e. whitespace tokenizer, sentence or paragraph splitter

Model of Seeding



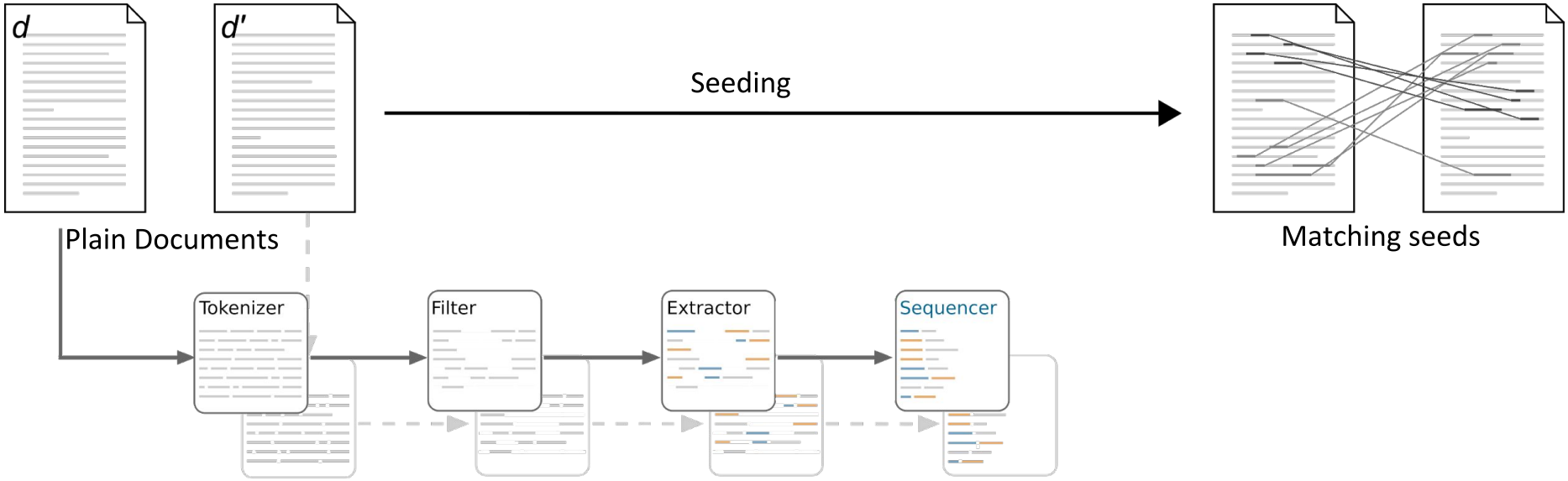
- i.e. wordlist filter, POS-tag filter

Model of Seeding



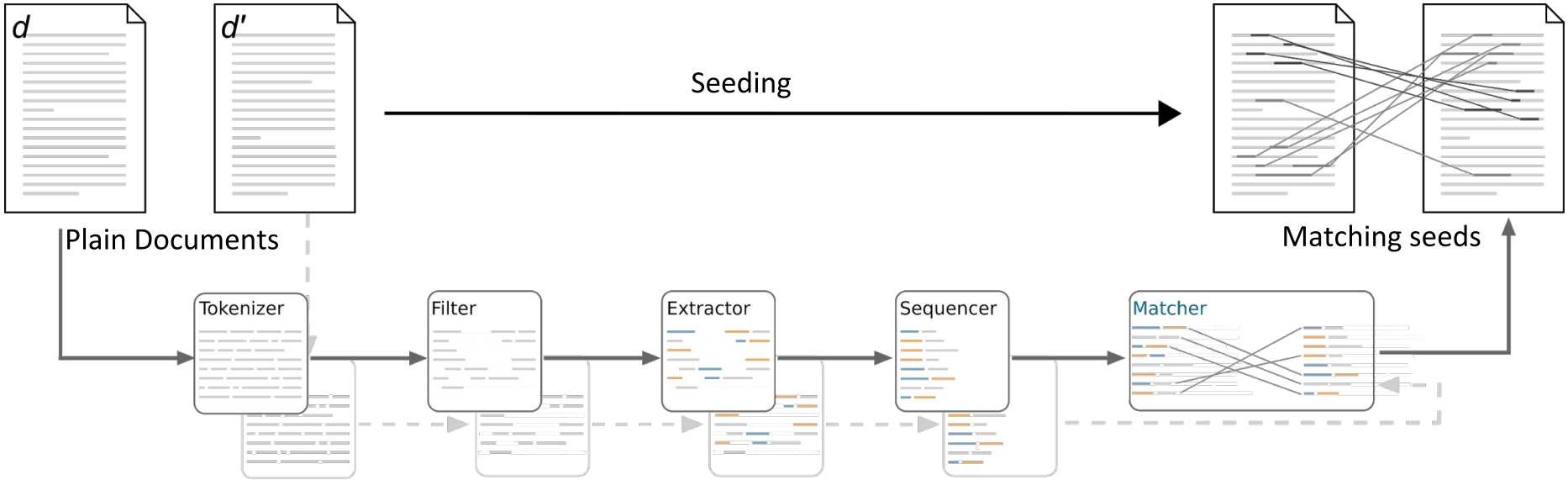
- i.e. plaintext, hypernyms, frequency or word vectors

Model of Seeding



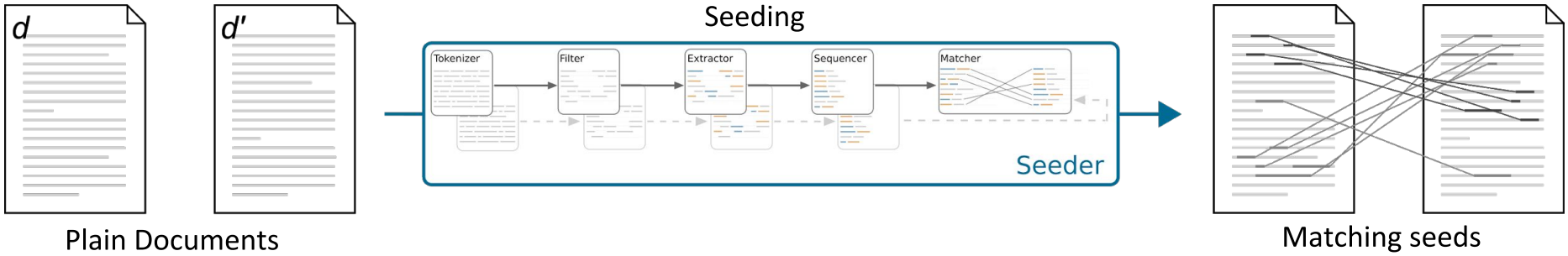
- i.e. n-grams or skip-grams

Model of Seeding

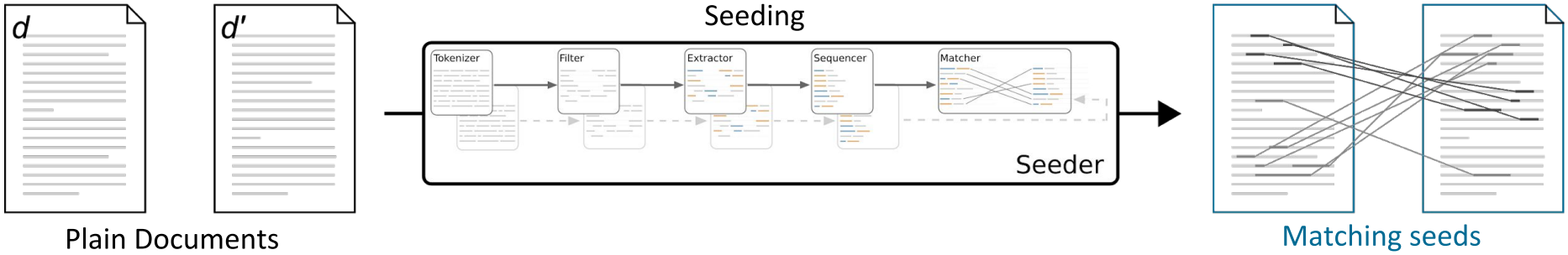


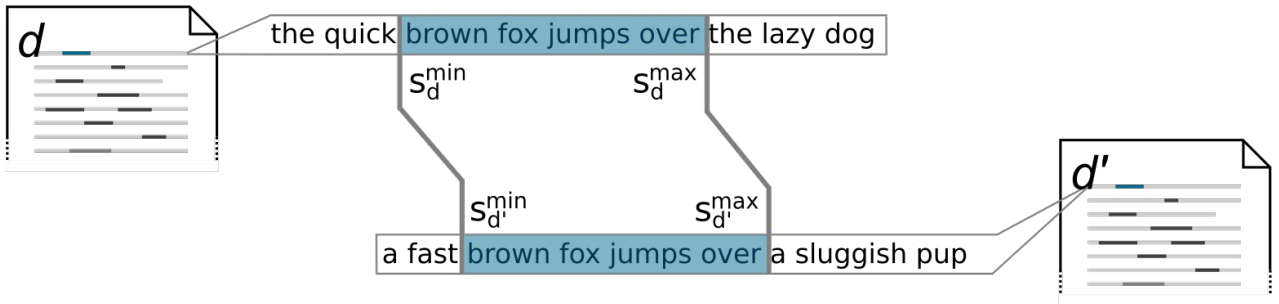
- i.e. exact or set match, Jaccard or cosine similarity

Model of Seeding

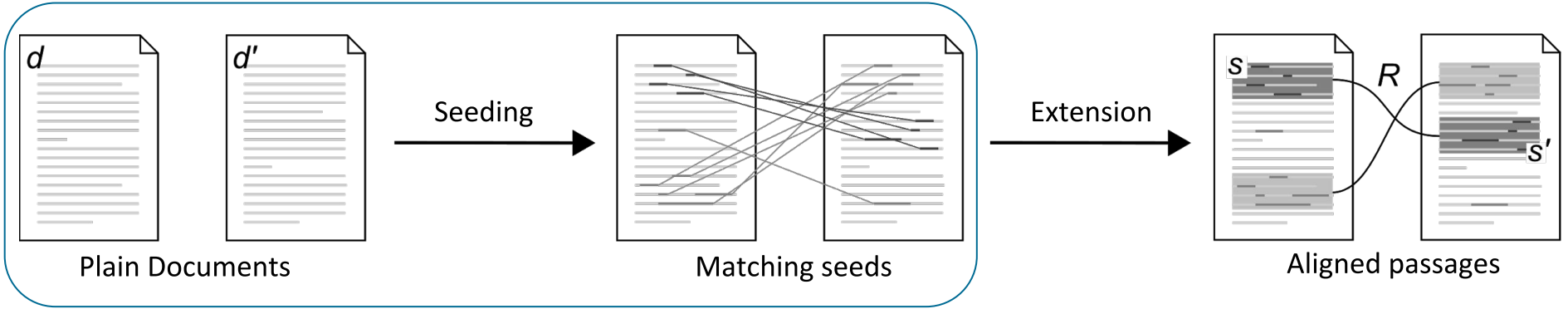


Model of Seeding



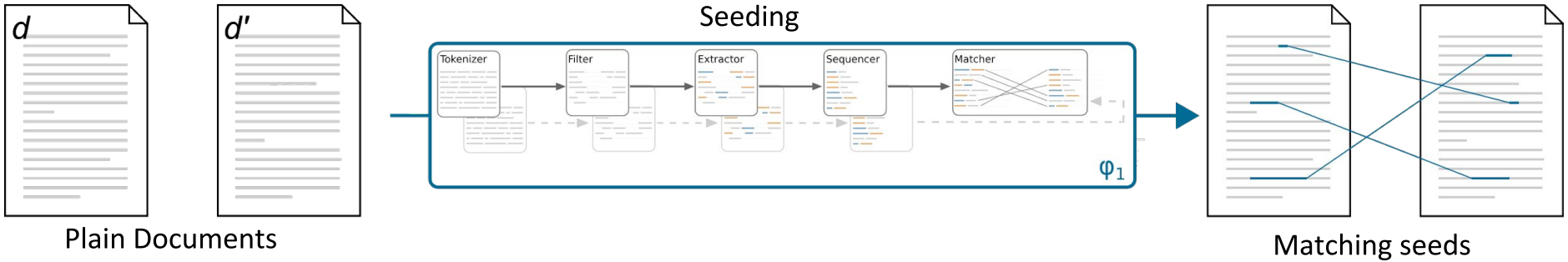


Contributions

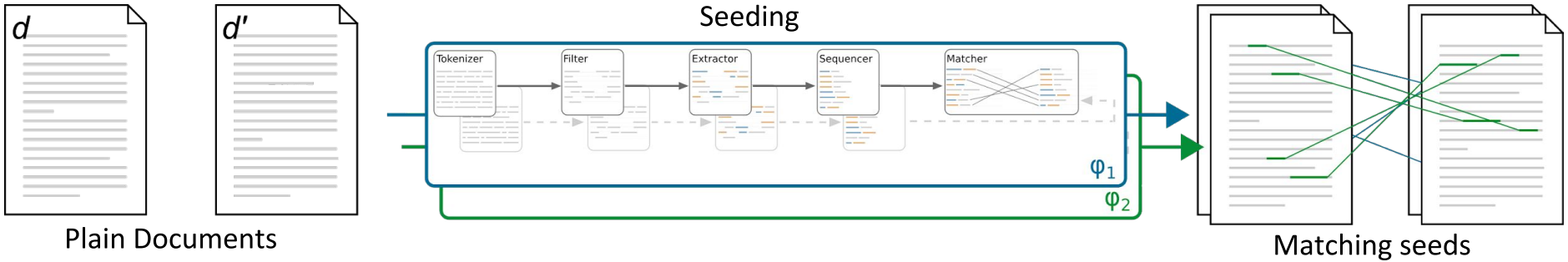


- Model of seeding
- Seeder combination
- Relaxation
- Model of extension
- Parameter estimation

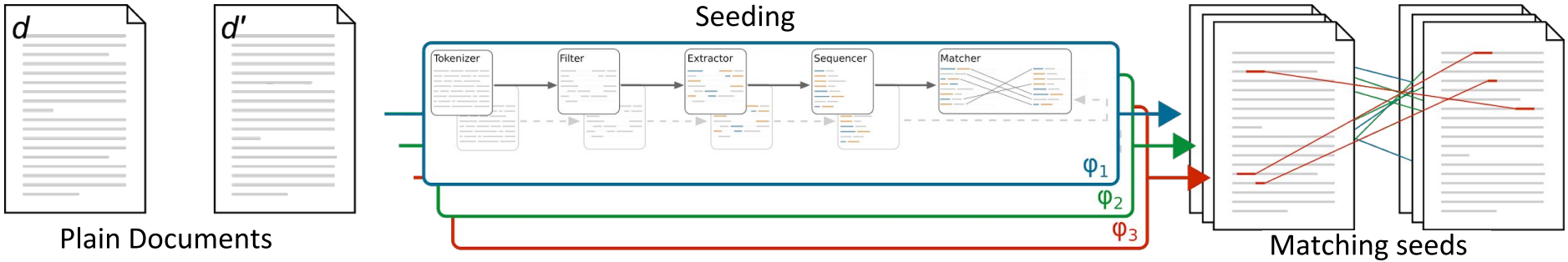
Seeder combination



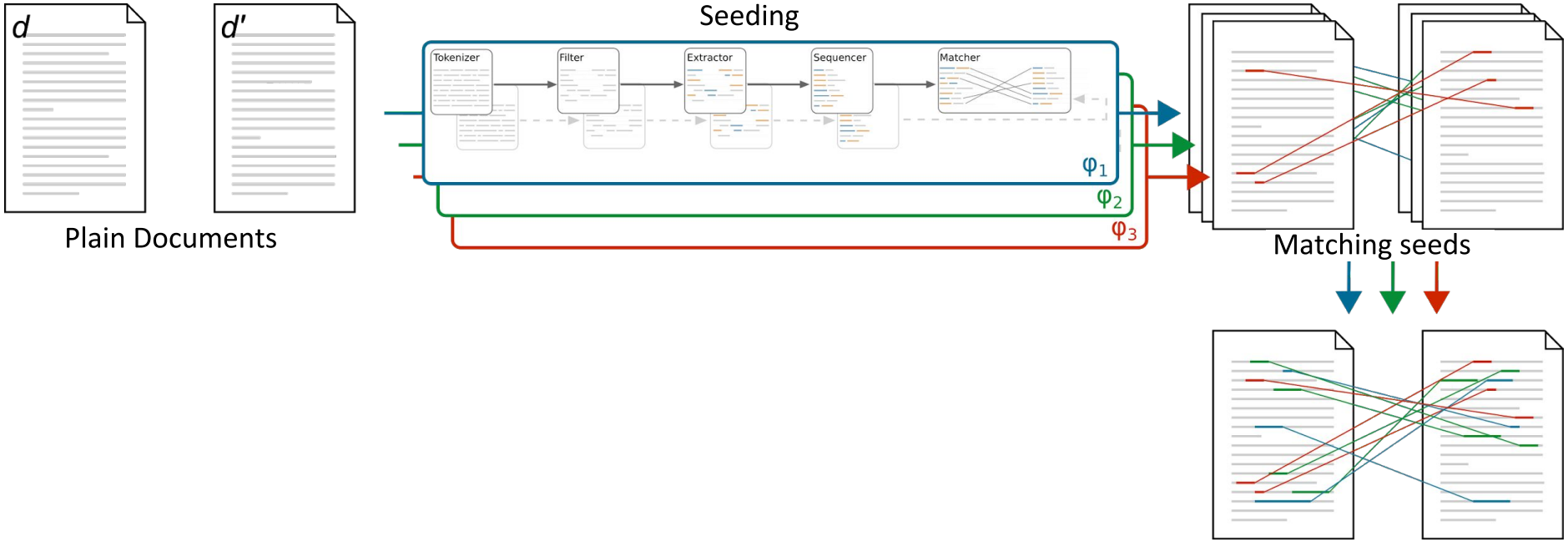
Seeder combination



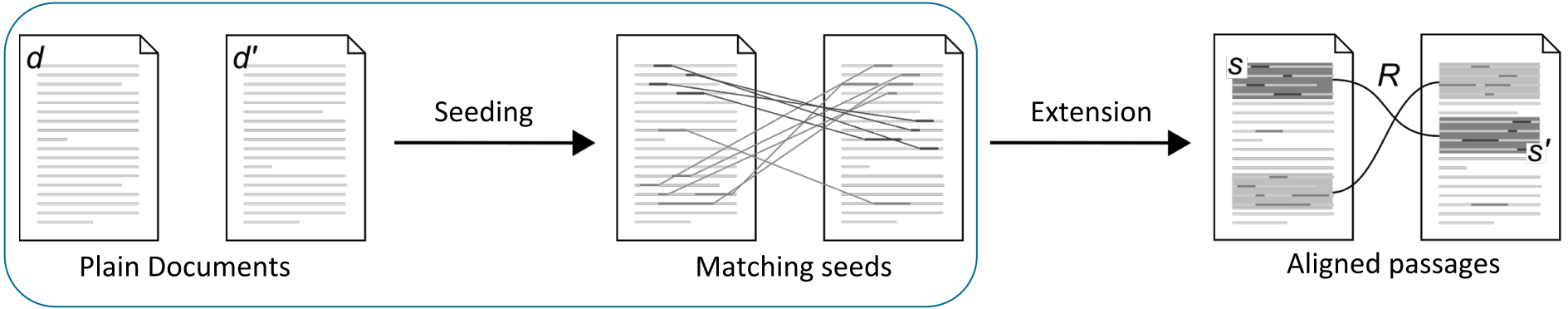
Seeder combination



Seeder combination

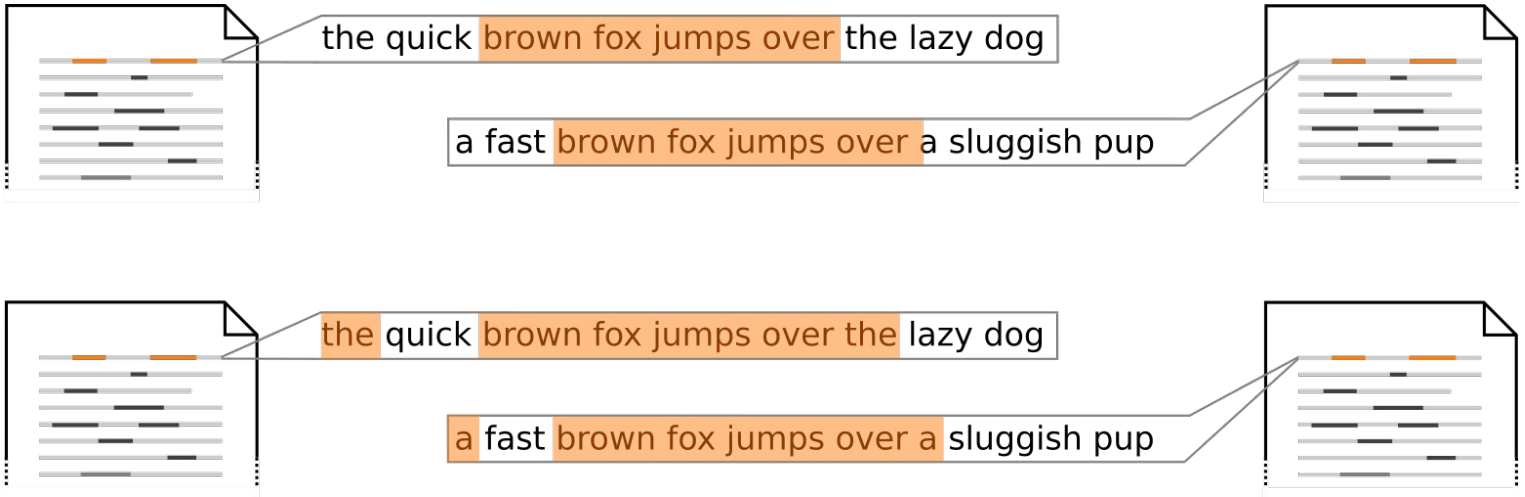


Contributions

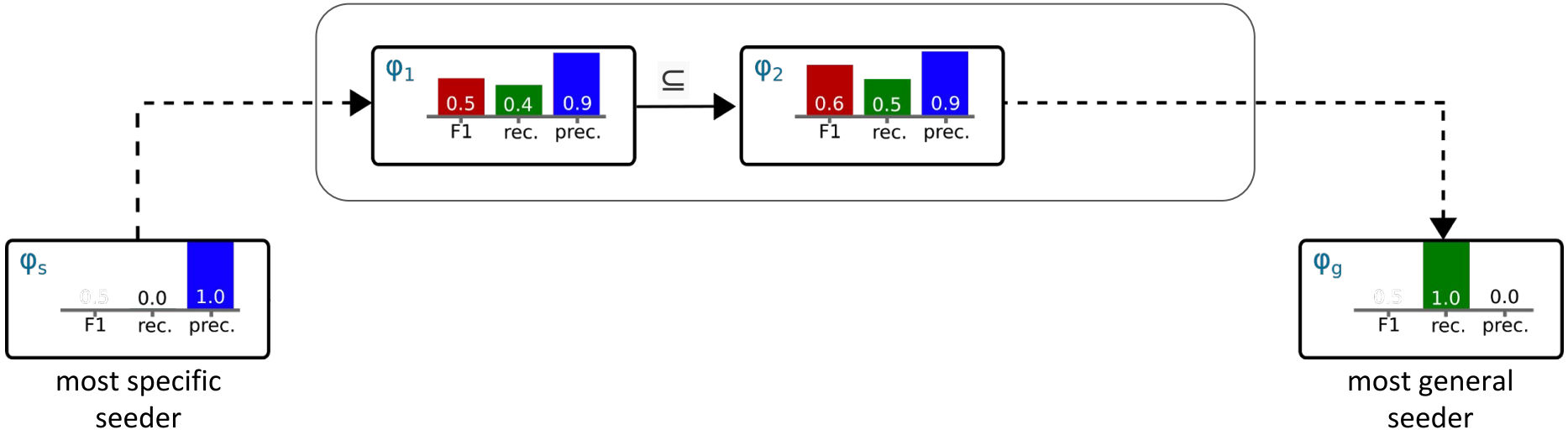


- Model of seeding
- Seeder combination
- Relaxation
- Model of extension
- Parameter estimation

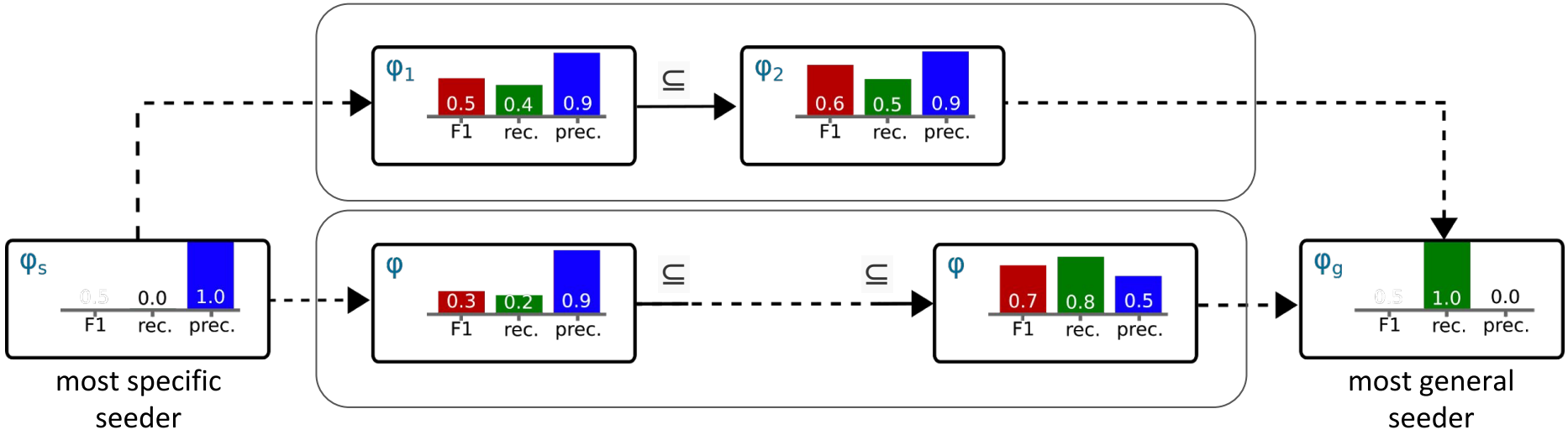
Relaxation


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

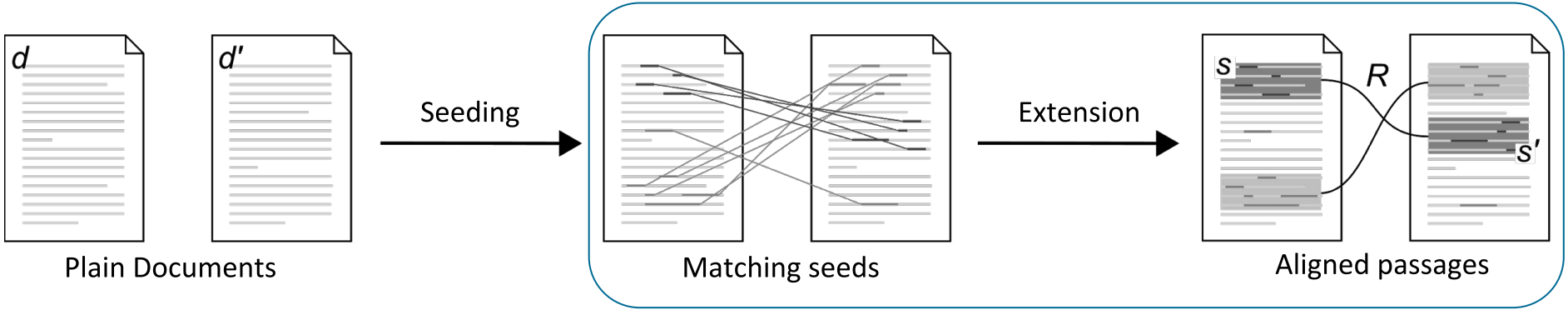
Relaxation


φ1
0.5
0.4
0.9
F1
rec.
prec.
⊆
φ2
0.6
0.5
0.9
F1
rec.
prec.
φs
0.5
0.0
1.0
F1
rec.
prec.
most specific seeder
φg
0.5
1.0
0.0
F1
rec.
prec.
most general seeder

Relaxation


φ1
0.5
0.4
0.9
F1
rec.
prec.
⊆
φ2
0.6
0.5
0.9
F1
rec.
prec.
φs
0.5
0.0
1.0
F1
rec.
prec.
most specific seeder
φ
0.3
0.2
0.9
F1
rec.
prec.
⊆
⊆
φ
0.7
0.8
0.5
F1
rec.
prec.
φg
0.5
1.0
0.0
F1
rec.
prec.
most general seeder

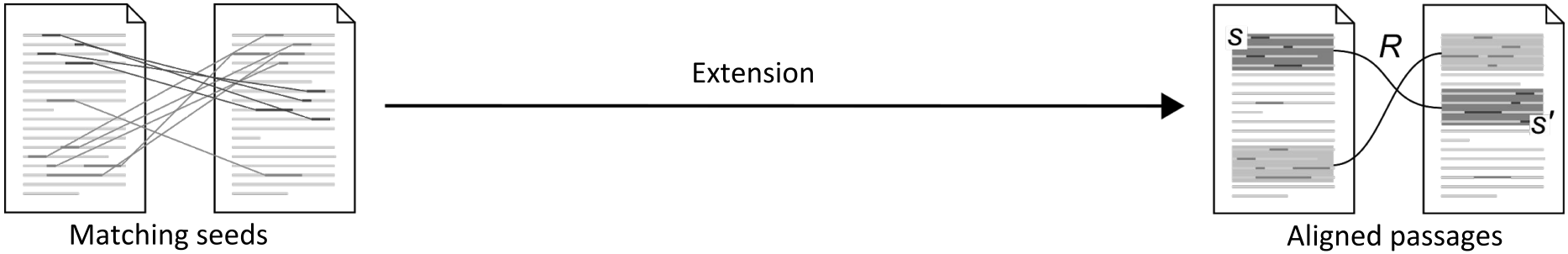
Contributions



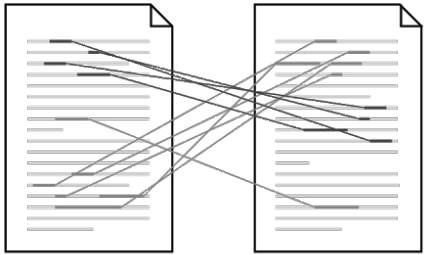
- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation

Model of Extension



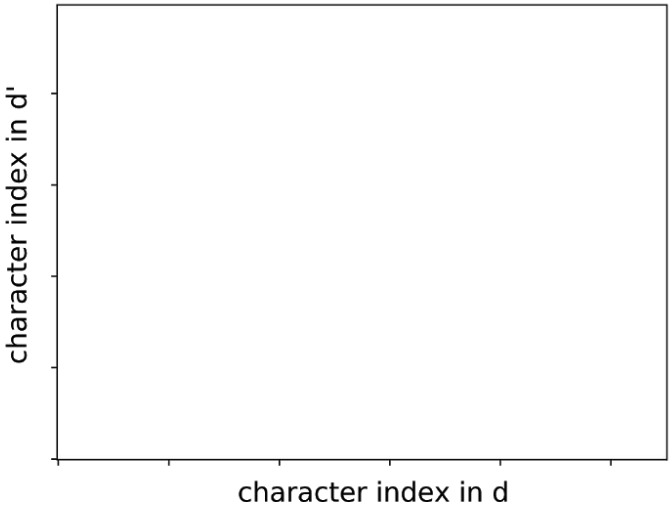
Model of Extension



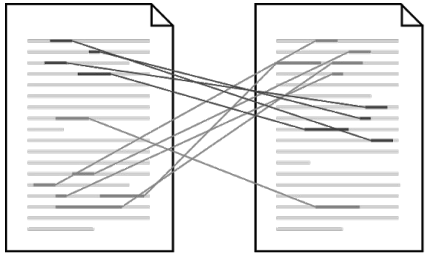
Matching seeds

Extension

Aligned passages

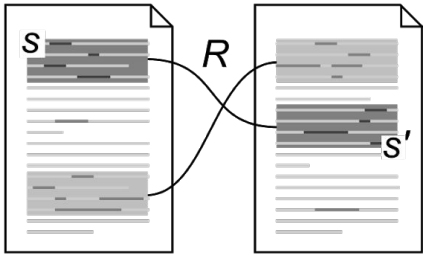


Model of Extension

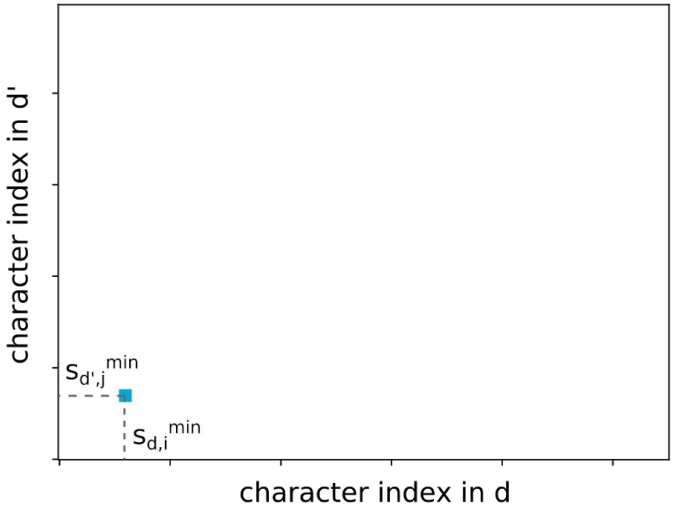


Matching seeds

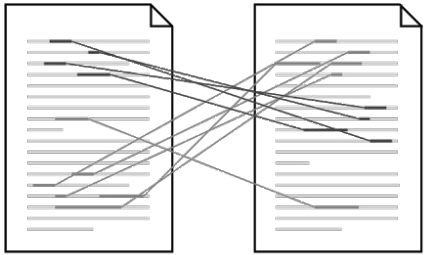
Extension



Aligned passages



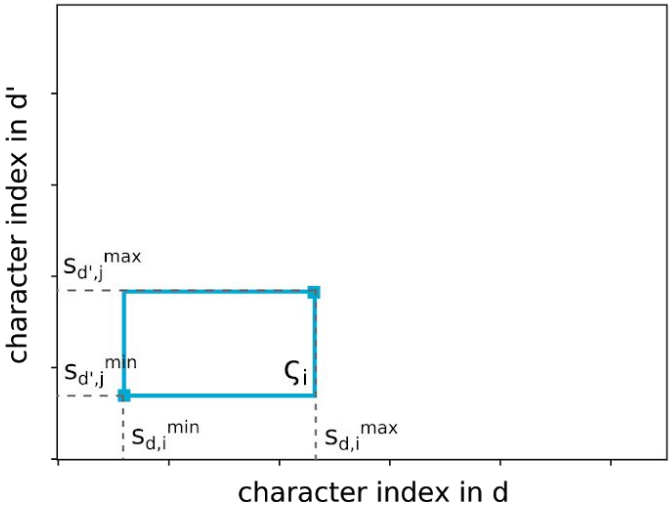
Model of Extension



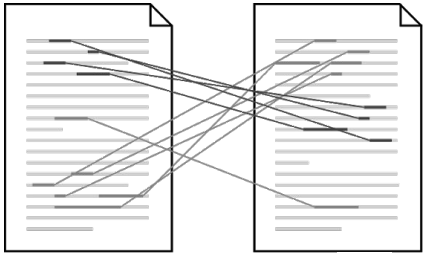
Matching seeds

Extension

Aligned passages

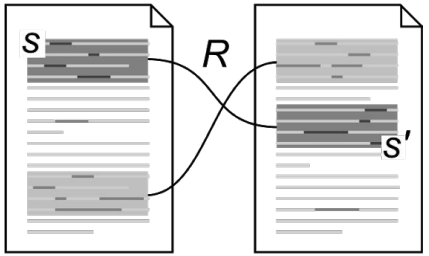


Model of Extension

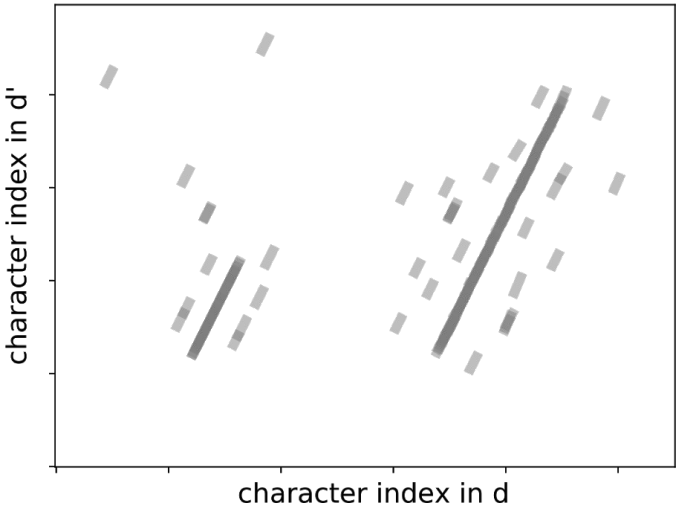


Matching seeds

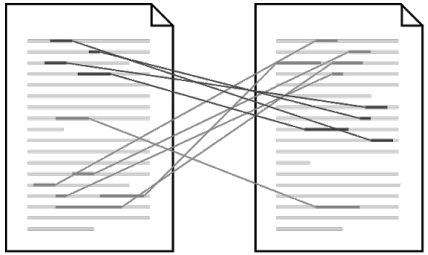
Extension



Aligned passages

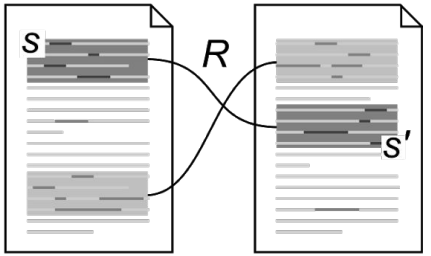


Model of Extension

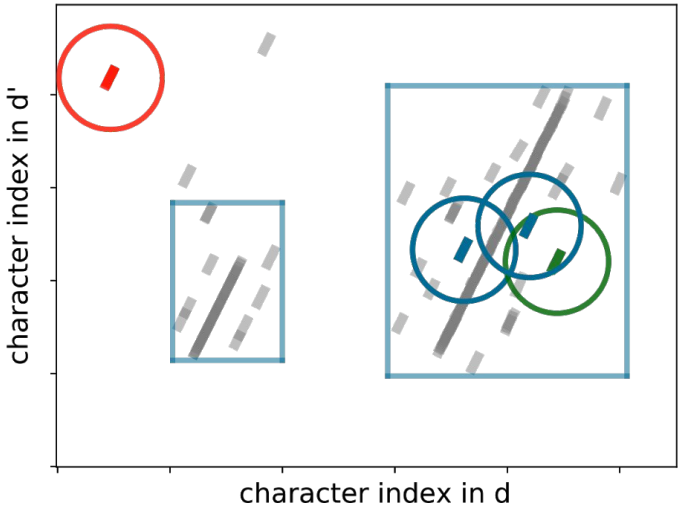


Matching seeds

Extension



Aligned passages

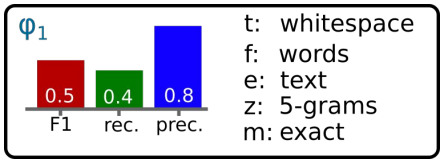


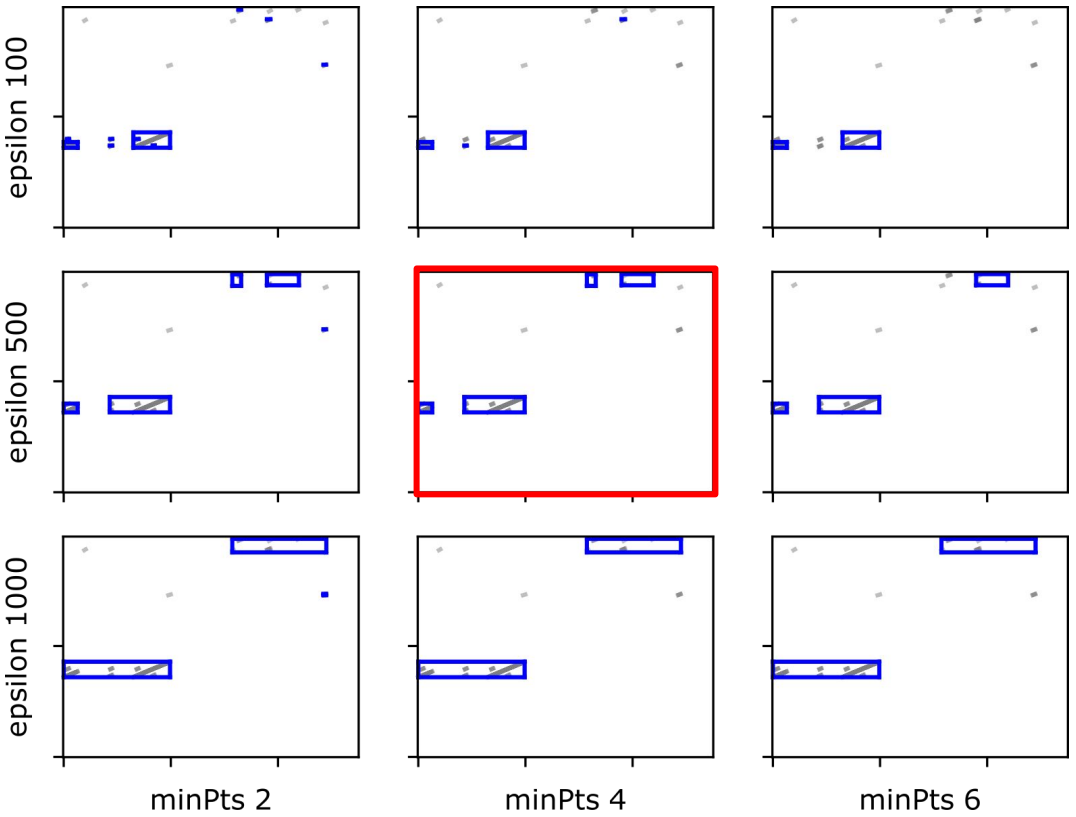
Core Point: if at least **minPts** other points within **epsilon** radius

Border Point: within **epsilon** of a core point

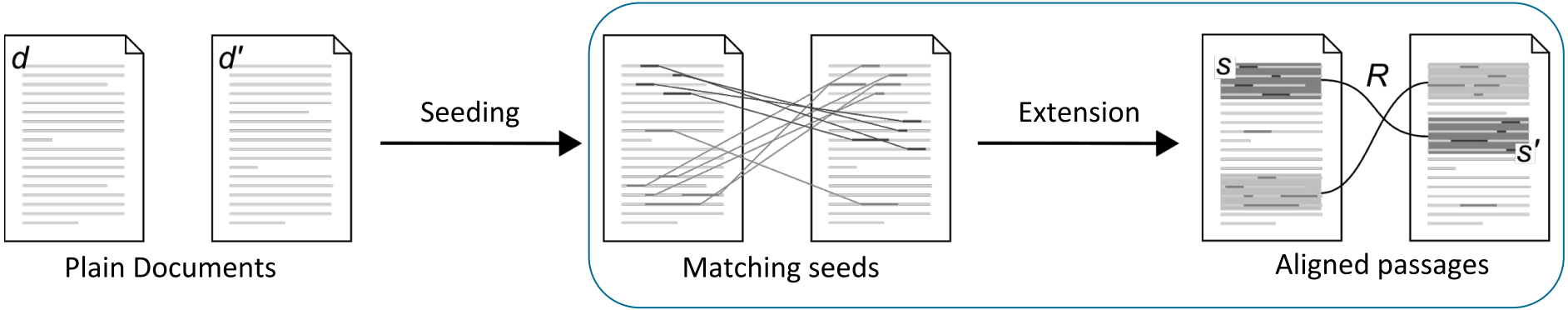
Noise Point: if neither core nor border point

Model of Extension





Contributions



- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation

Hyperparameter estimation

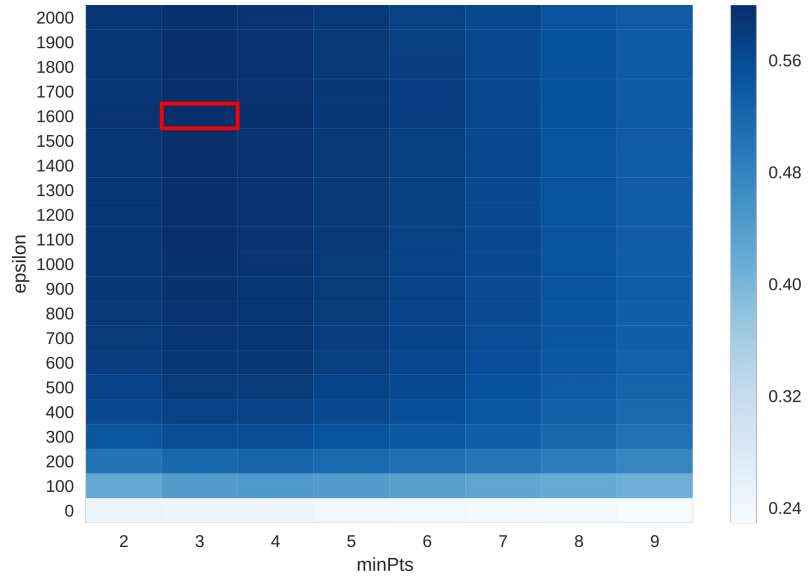
- Collection estimation
  - Given the seeder, find the best parameters for all documents
  - Can be determined once with the ground truth

	None	collection estimate
F1	0.54	0.62

Hyperparameter estimation

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

	None	collection estimate
F1	0.54	0.62

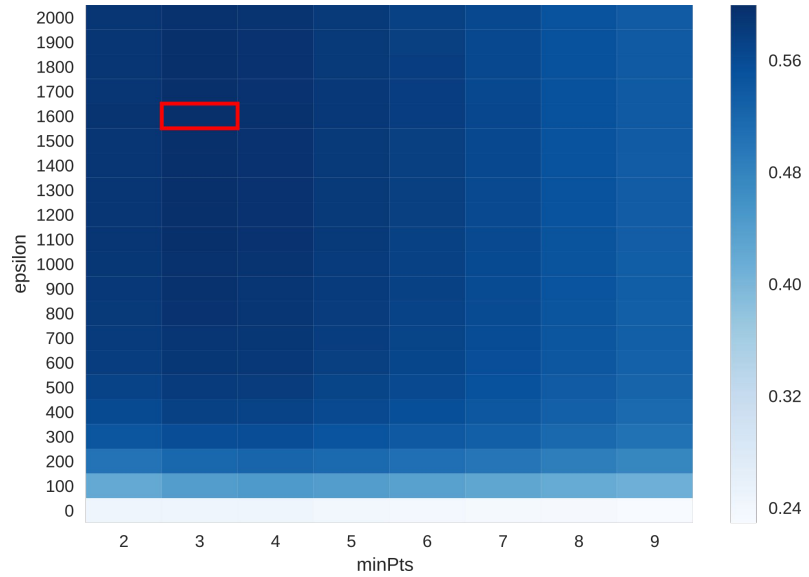


Hyperparameter estimation

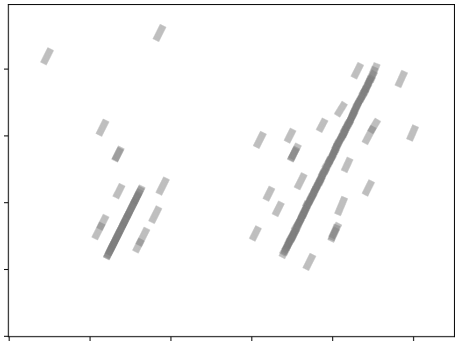
- Collection estimation
  - Given the seeder, find the best parameters for all documents
  - Can be determined once with the ground truth

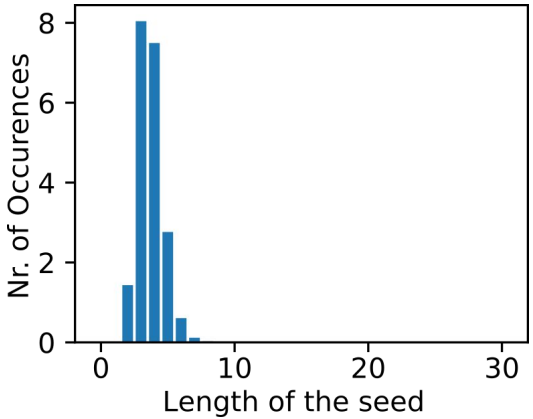
- Document estimation
  - Given the seeder and a pair of documents, find the best parameters for that particular pair
  - This can be learned

	None	collection estimate	per document optima
F1	0.54	0.62	0.72



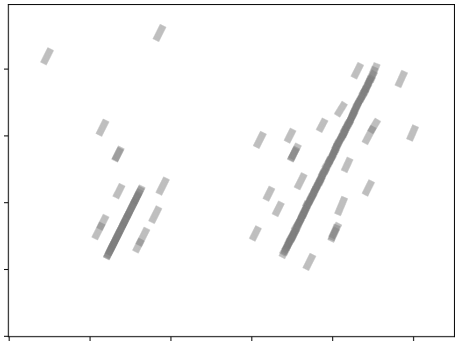
Hyperparameter estimation

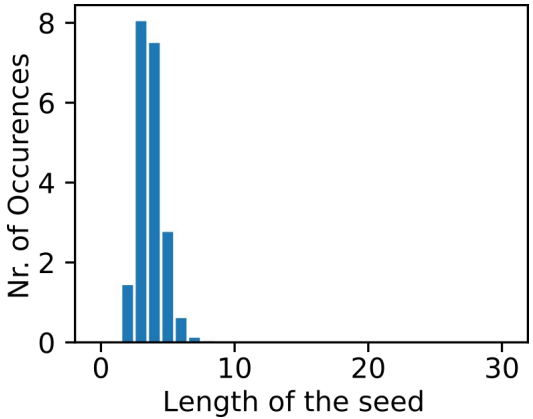




- Determine the truth (best parameters) for a pair of documents via gridsearch
- use length-frequency-histogram as features
- get 10.000 examples per seeder from the PAN corpora

Hyperparameter estimation

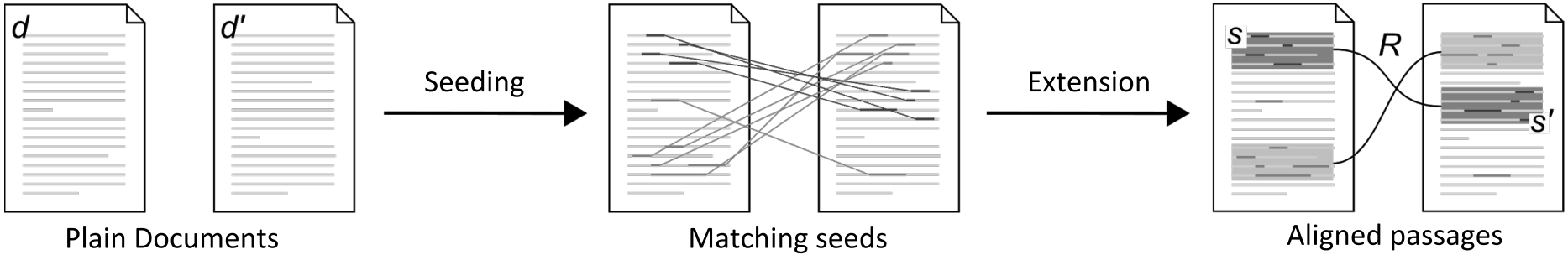




- Determine the truth (best parameters) for a pair of documents via gridsearch
- use length-frequency-histogram as features
- get 10.000 examples per seeder from the PAN corpora

	prediction	collection estimate	per document optima
$\varphi_1, \varphi_2 \rightarrow \varphi_1 \cup \varphi_2$	0.70	0.69	0.86
$\varphi_1, \varphi_3 \rightarrow \varphi_1 \cup \varphi_2 \cup \varphi_3$	0.65	0.69	0.85

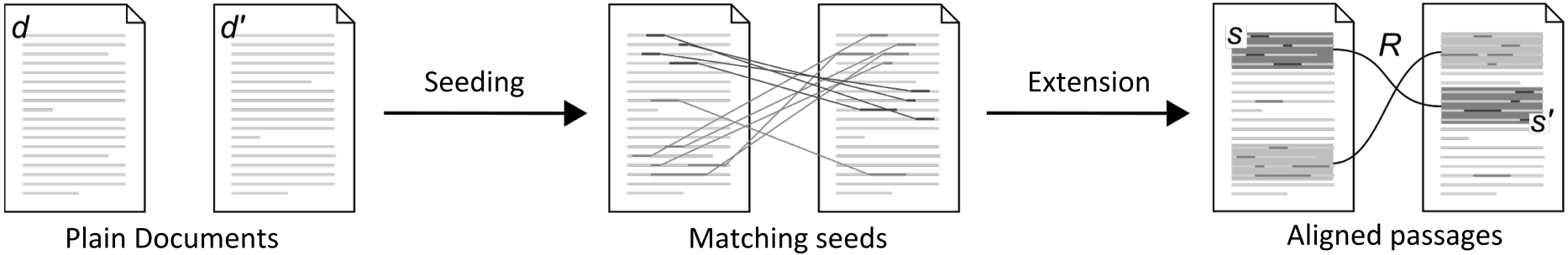
Contributions



- Model of seeding
- Seeder combination
- Relaxation

- Model of extension
- Parameter estimation

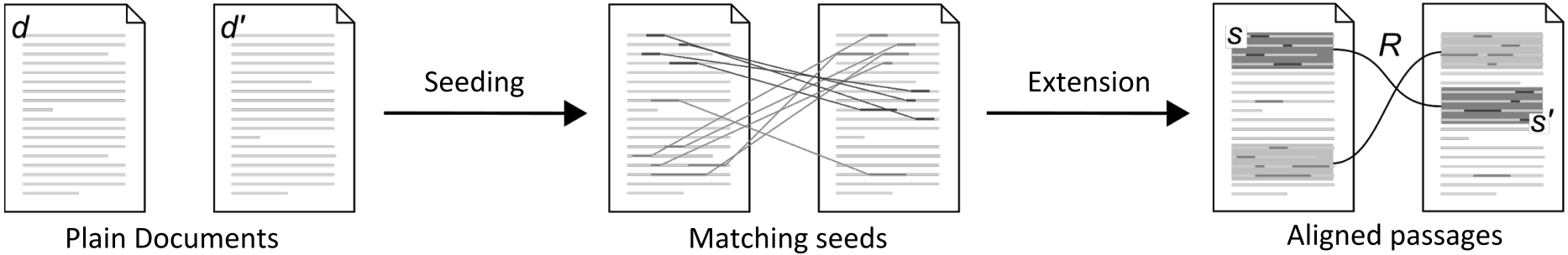
Future Work



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

- Model of extension
- Parameter estimation
- Improved parameter learning

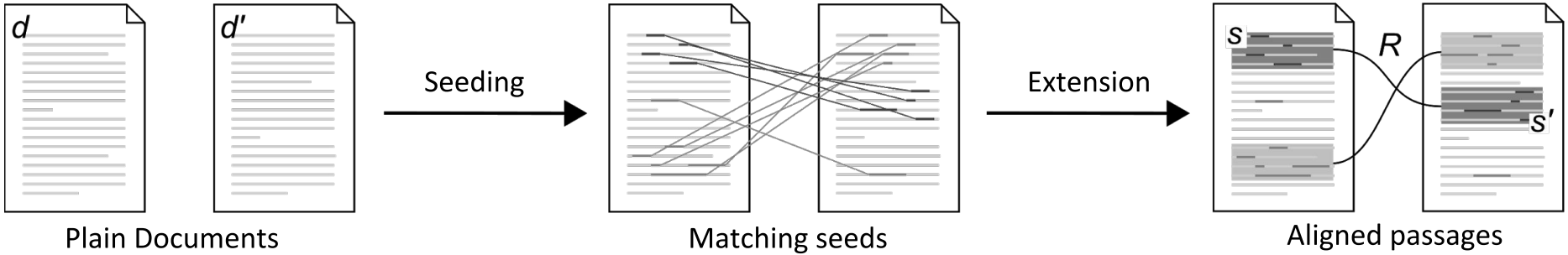
Future Work



- 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