Automatic Key Term Extraction from Scientific Articles in GROBID

Patrice Lopez Laurent Romary

INRIA & Humboldt University

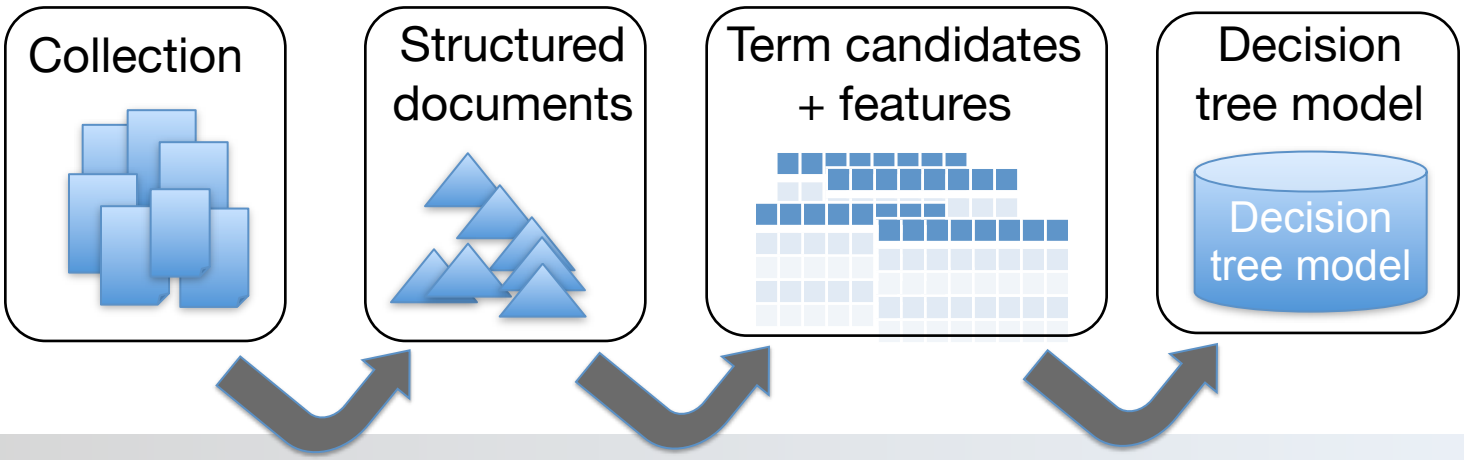
Berlin, Germany

Background

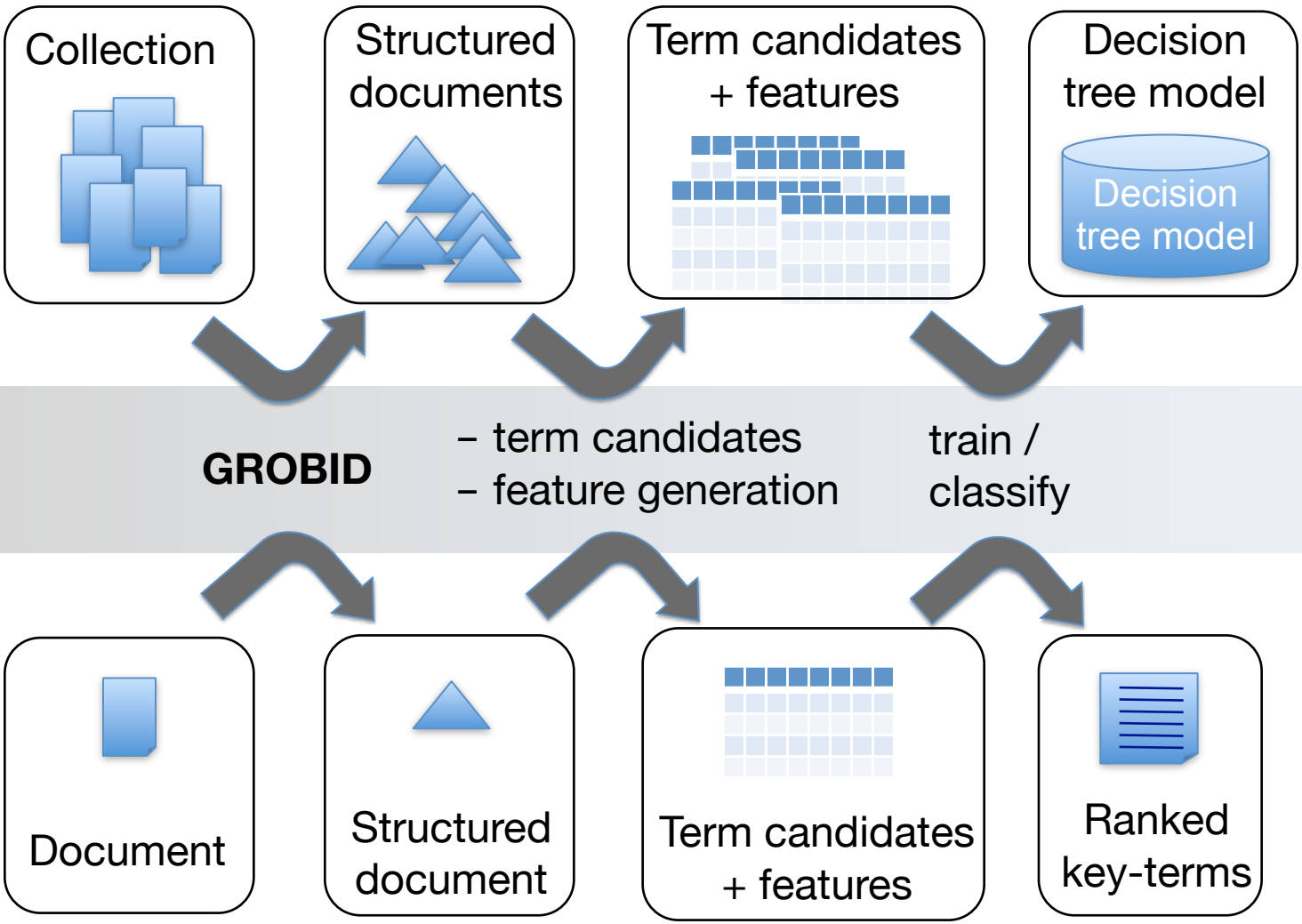
- Semeval Task 5: “Automatic Keyphrase Extraction from Scientific Articles”
- Keyphrases/Keywords:
  - Determined by authors and readers
  - Provide important/discriminant topical information about the content
- The keyword determination process is subjective
- How can we characterize this subjectivity when considering automatic term extraction mechanisms ?
- We view this work as a subtask of extracting technical terms from texts: key-term extraction

Key-term extraction overview

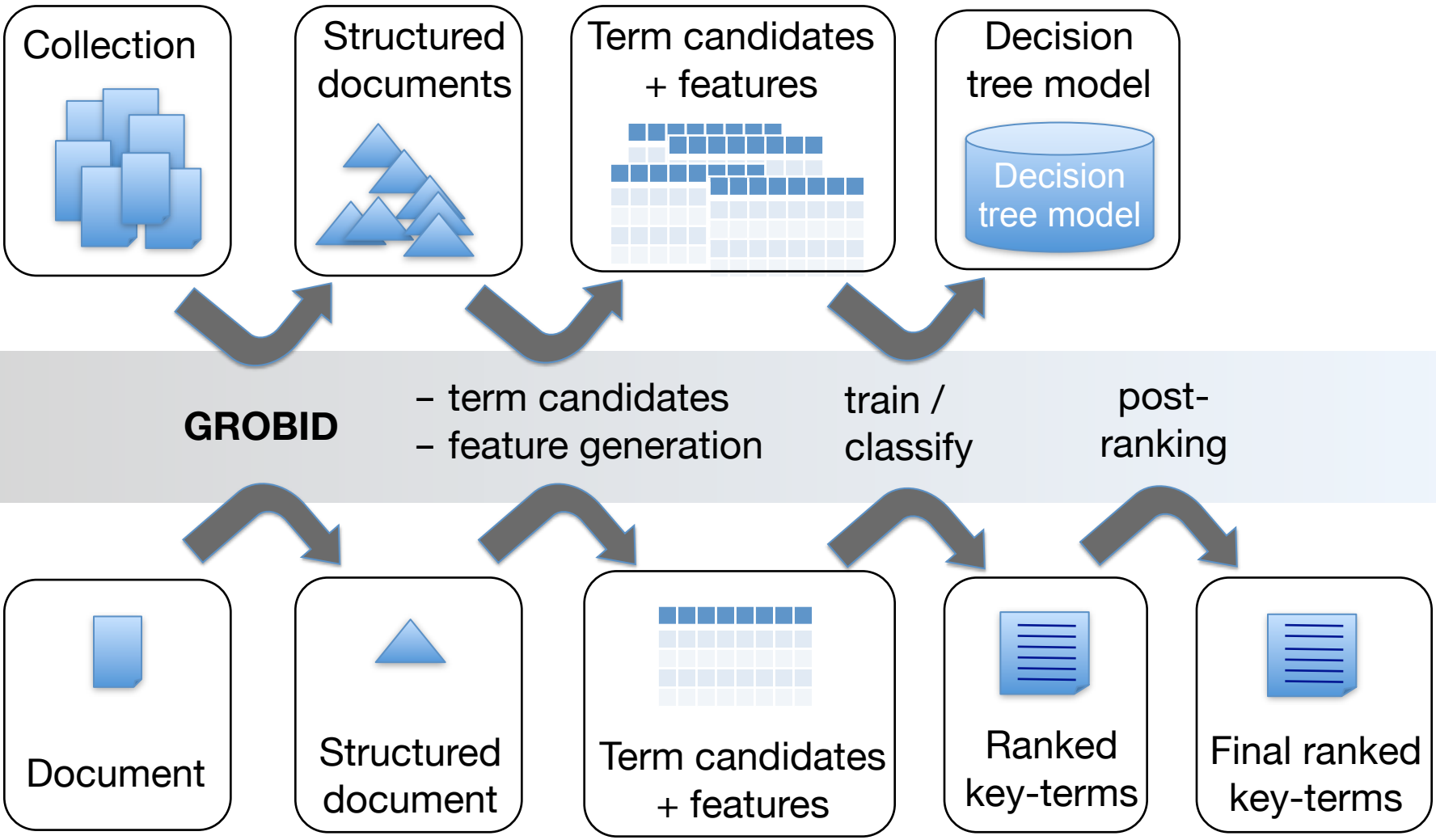
Key-term extraction overview



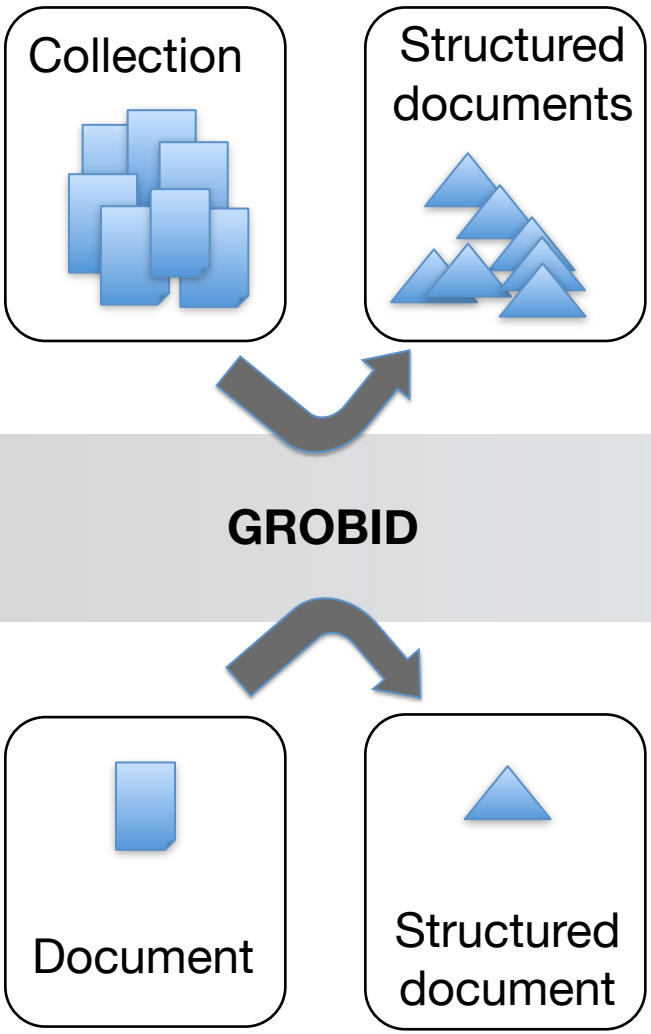
Key-term extraction overview



Key-term extraction overview



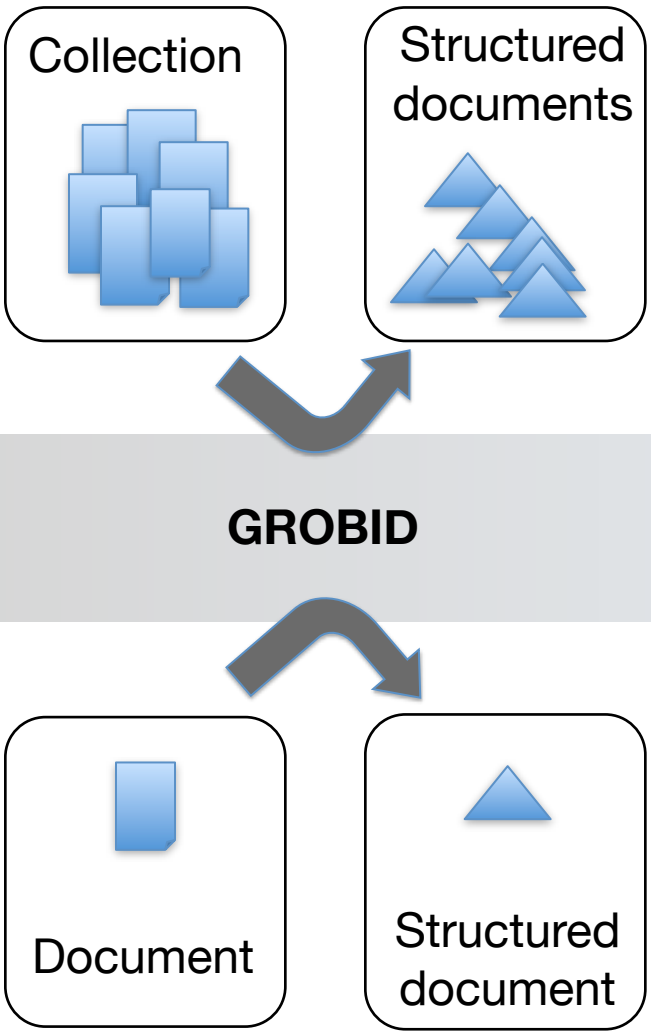
Key-term extraction overview



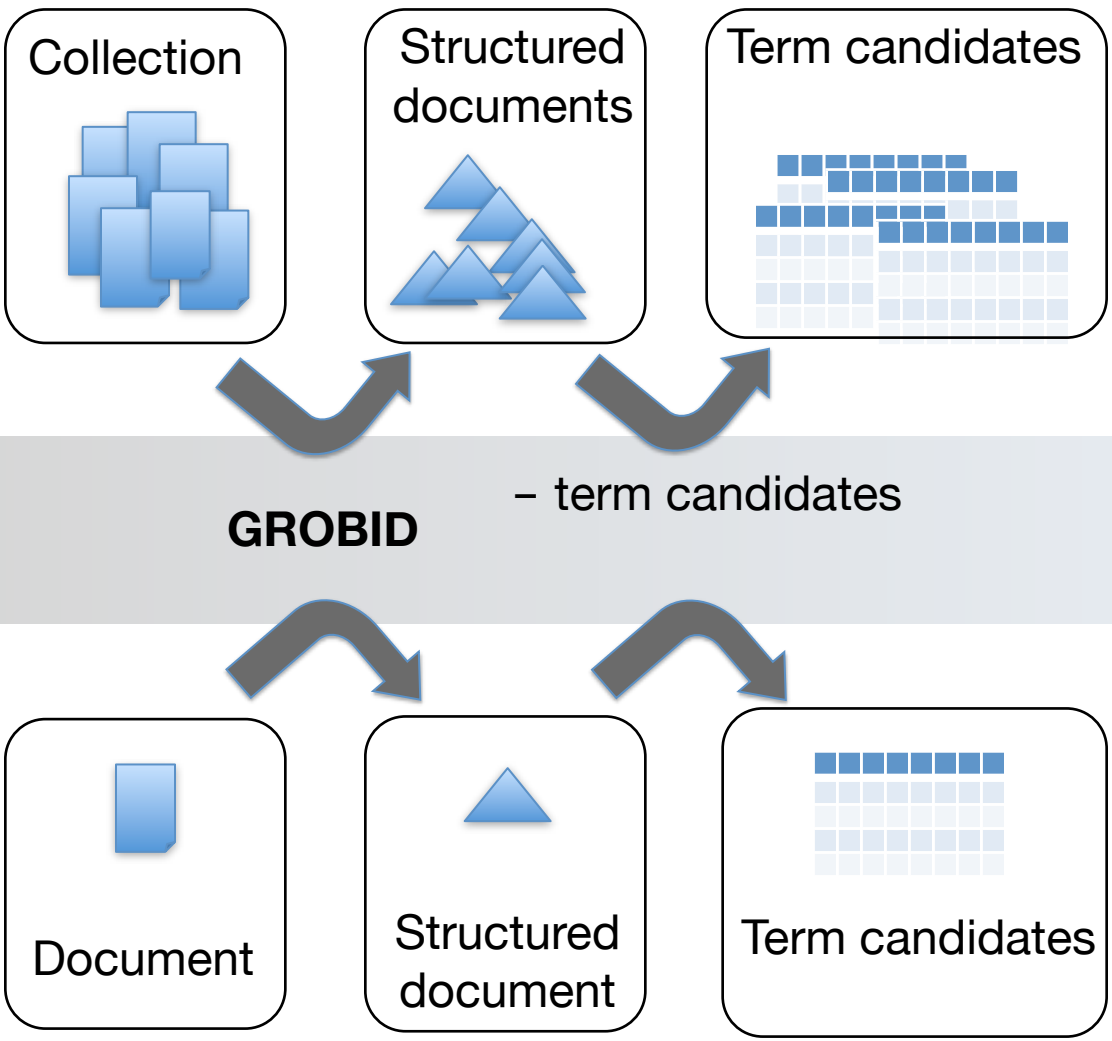
Structure analysis: GROBID

- Conversion of each text into a TEI conformant document
- Conditional Random Fields (Peng & McCallum, 2004) based on a very large training corpus
  - recognition of header metadata and references
  - recognition of the different sections
- Specific training on a few ACM documents
- Between 98% (section titles, reference titles) and 99% (title, abstract) accuracy for the Semeval collection

Key-term extraction overview



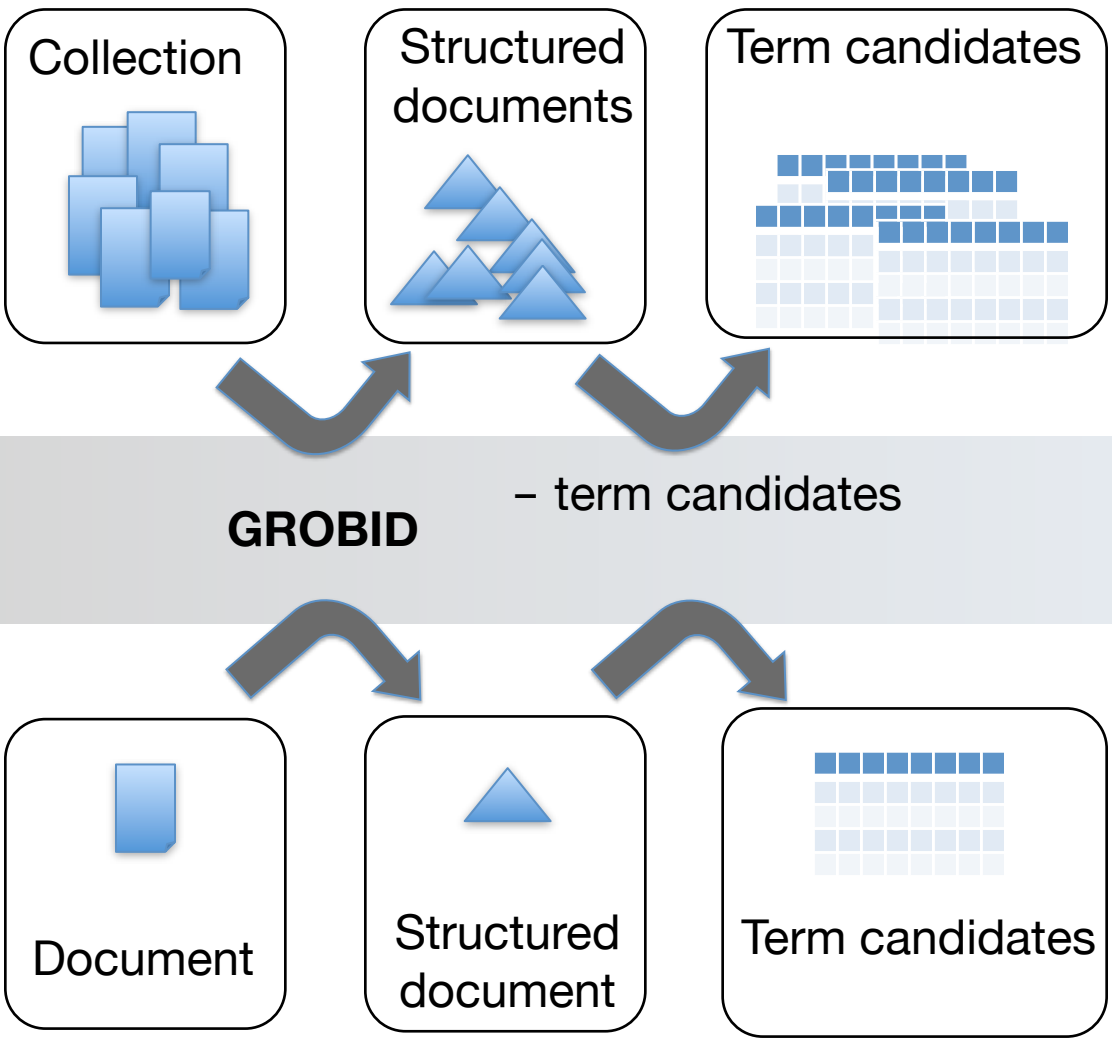
Key-term extraction overview



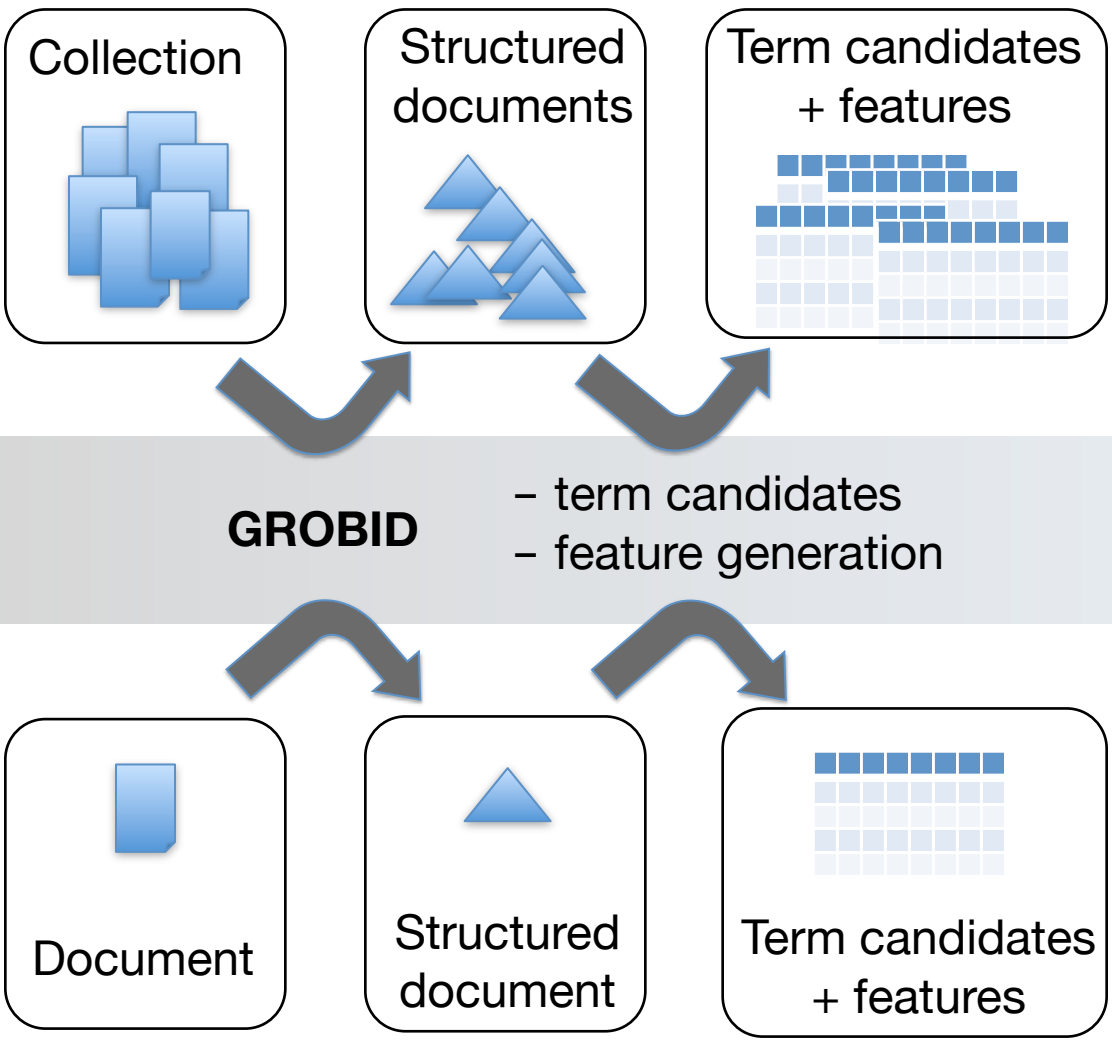
Candidate phrase selection

- Extraction of all n-grams up to 5 tokens (word)
- Remove sequences starting or ending with a stop word
- Normalize candidate sequences: lowercasing, Porter stemming

Key-term extraction overview



Key-term extraction overview



Feature generation

- Three types of features:
  - structural features
  - content features
  - lexical/semantic features

14

Structural features

- Determination of reliable features depending on the logical composition of a document
- Fields where the main concept of a paper is introduced: header (title, abstract), text body (introduction, section titles, conclusion), reference list (article, journal or book title)
    - Proceedings of the 378th Conference on ***Semantic Web***
- Similarity with writing/reading behaviors
- Features
  - Presence in: *title*, *abstract*, *introduction*, *section title, conclusion*, cited *article/book/conference titles*
  - Position of the first occurrence of the word in the document

Content features

- Objective:
  - capture distributional properties of a term in comparison to the document or the collection
- Three distributional measures:
  - Phraseness
  - Informativeness
  - Keywordness

Phraseness

- Degree to which a ngram can be considered as a phrase
- Based on lexical cohesion of a sequence of words within a document
  - Measure: Generalized Dice Coefficients (Park et al, 2002), applicable to arbitrary n-grams (n ≥ 2)

$$GDC(T) = \frac{\mid T \mid log_{10}(freq(T))freq(T)}{\sum_{w_i \in T} freq(w_i)}$$

- With:
  - T: term
  - |T|: number of words in T
  - freq(T): number of occurrences (frequency) of T
  - freq(w_i): number of occurrences of w_i

Informativeness

- Degree to which a term is representative of a document given a collection of documents
  - measure: TF-IDF (Witten et al., 1999)

$$\text{TF-IDF}(T, D) = \frac{freq(T, D)}{\mid D \mid} \times -log_2 \frac{count(T)}{N}$$

  - With:
    - freq_D(T): frequence of T in D
    - |D|: number of words in D
    - count(T): number of occurrences of T in the corpus
    - N: number of documents in the corpus

Keywordness

- Degree to which a term is selected as a keyword (Witten et al., 1999).
- Based on observation of existing manually indexed sources
- Frequency of the keyword in the global corpus

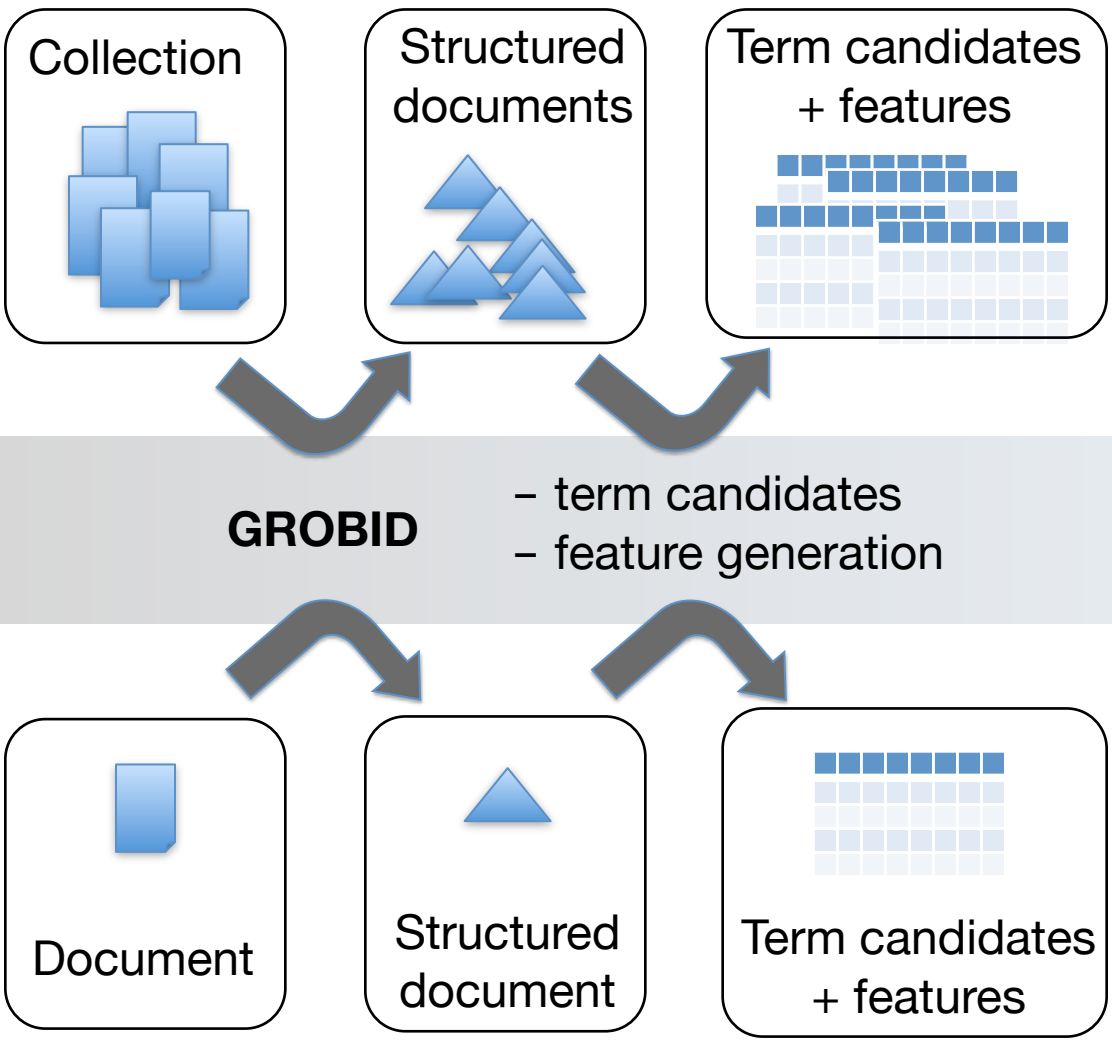
Lexical/Semantic features

- Comparison of term candidates with controlled terms in reference lexical sources
  - GRISP, a large scale terminological database for technical and scientific domains based on machine learning driven fusion of resources (cf. Lopez et Romary, 2010):
    - Terminological resources (MeSH, Gene Ontology, ChEBI)
    - Linguistic resources (part of Wordnet)
    - Wikipedia
  - Wikipedia keyphraseness (cf. Medelyan, 2009)
    - Keyphraseness= ~probability of the term to be an anchor
- Additional feature: length of the candidate T

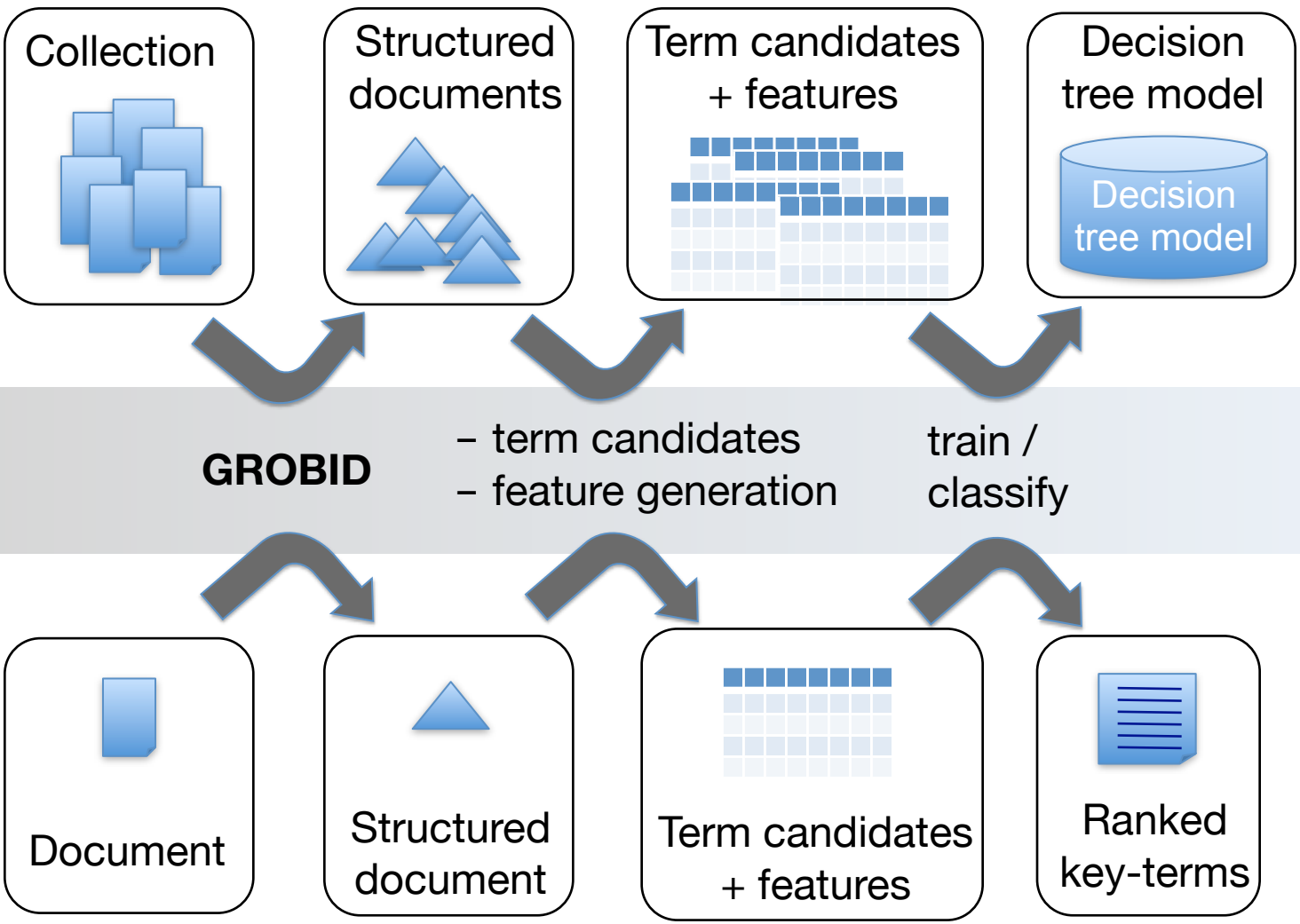
Contribution of features

Features	F-score top 15	#
All	27.5	
without structure positions	26.6	-3.3%
without relative position in doc.	25.4	-7.7%
without phraseness (Generalized Dice coefficients)	25.9	-5.8%
without informativeness (TF-IDF)	22.0	-19.8%
without keywordness	26.5	-3.4%
without GRISP matching	26.9	-2.1%
without Wikipedia keyphraseness	27.0	-1.9%
without key-term length	26.4	-3.8%

Key-term extraction overview



Key-term extraction overview



Training Corpus

- Corpus Semeval Task 5
  - Articles from ACM (Association of Computational Machinery)
    - Four sub-domains: C.2.4 *Distributed Systems*, H.3.3 *Information Search and Retrieval*, I.2.6 *Learning* and J.4 *Social and Behavioral Sciences*
- Training data
  - Semeval training data
    - 144 articles from ACM ; from the selected domains
  - National University of Singapore (NUS) corpus
    - 156 annotated ACM articles; all ACM domains

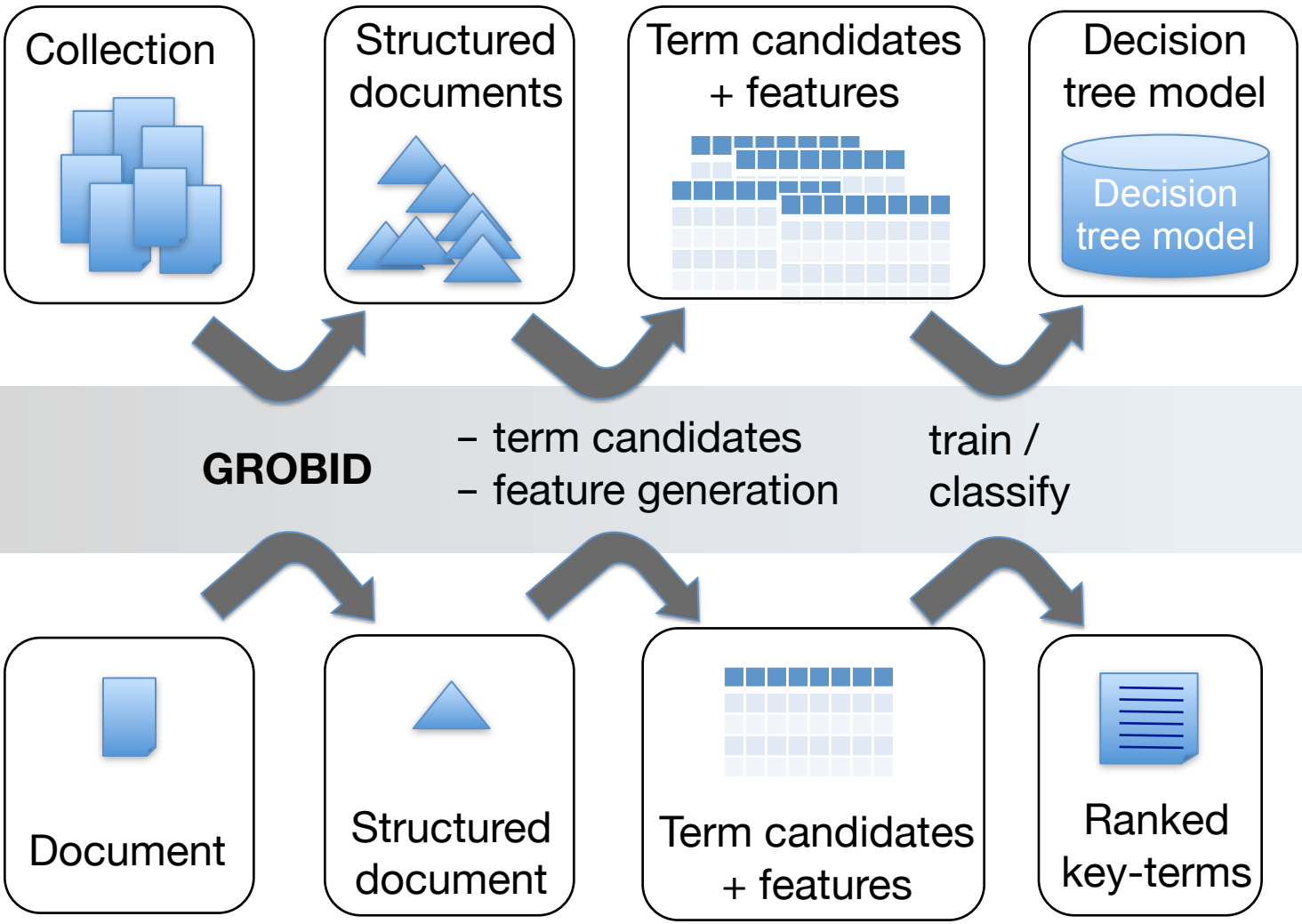
Training Corpus

- Impact of the additional training data on the results:

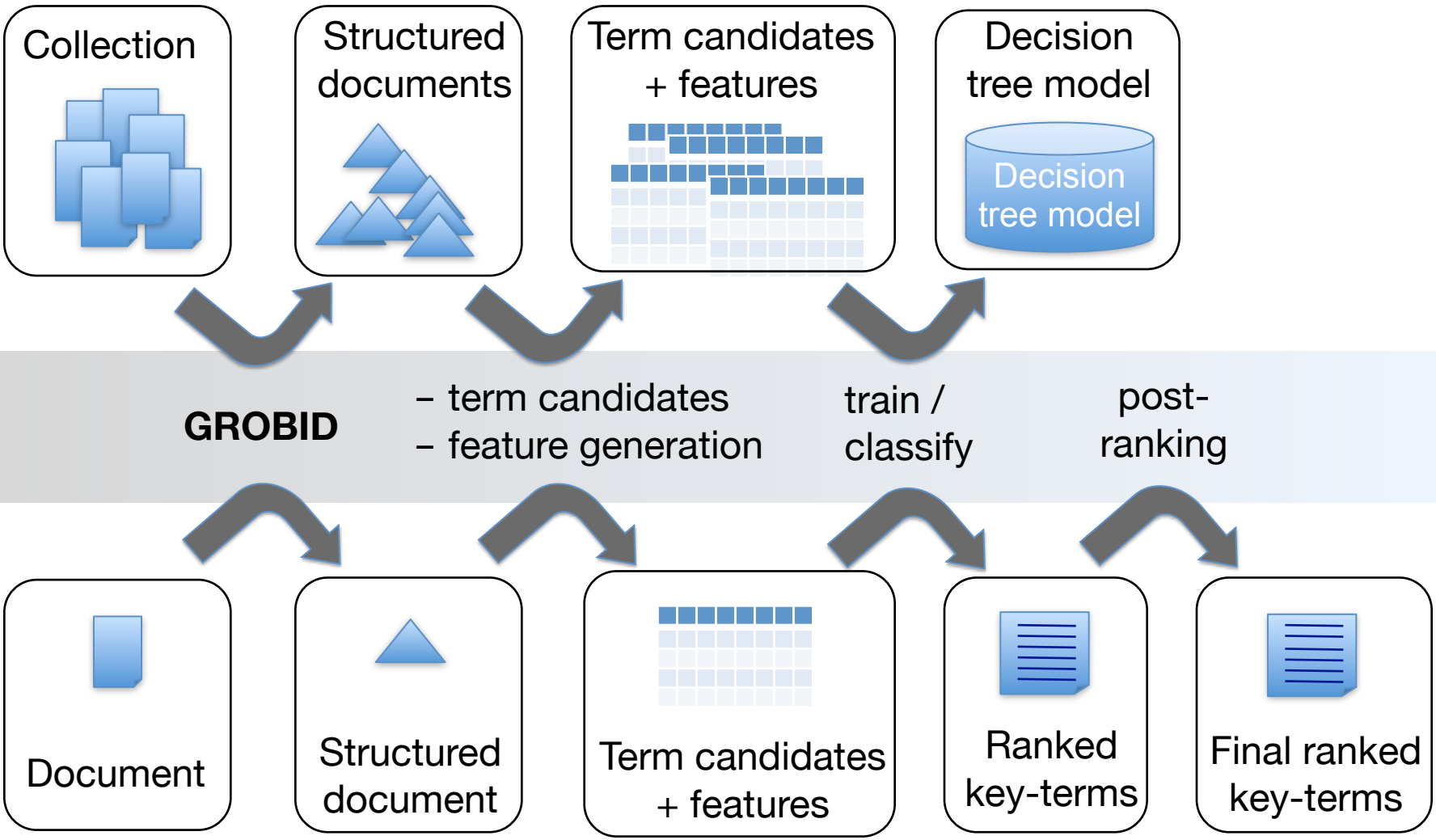
Training corpus	F-score top 15	#
Semeval + NUS	27.5	
Semeval	25.6	-6.91%

25

Key-term extraction overview



Key-term extraction overview



Final re-ranking

- Exploits general relationships between the candidate keywords:

$$s'(T_i) = s(T_i) + \alpha^{-1} \sum_{j \neq i} P(T_j|T_i) s(T_j)$$

- **$s(T_i)$**: score for term candidate i;
- **α**: re-ranking factor
- **P(Tj|Ti)**: Usage of keyword co-frequencies in HAL research article archive
  - French institutional Open Archive repository for research publications
  - 139,000 full texts articles
  - Restriction to English and computer science domain
  - 16,412 different author-selected keywords

Final re-ranking

- Impact of re-ranking on results:

Processing steps	F-score top 15	#
with post-ranking	27.5	
without post-ranking	26.6	-3.2%

What did not work

- **Term variants conflation**, based on Fastr (Jacquemin, 2001): –11.5% for the F-score at top 15
- **Noun phrase filtering**, based on the TreeTagger (Schmid, 1994): –7.6% for F-score at top 15
- **Language Model deviation** as informativeness measure, based on Lingpipe (Alias-i, 2008): –3.7% for the F-score at top 15
- **Global keywordness**, based on HAL repository: no result change
- **Wikipedia term Relatedness**, based on Wikepedia Miner (Milne & Witten, 2009): no result change

Current & future work

- Adaptation of the model for patent documents:
  - participation at CLEF IP 2010
  - key-term extraction on 2,6 million patent documents
  - in average less than 1 second processing per document
  - topic description based on key-terms for language model-based information retrieval
- Keyword suggestions for author self-archiving of articles in HAL

Thank you !

General principles

- Machine Learning based approach
  1. Analysis of the document structure
  2. Selection of candidate phrases
  3. Computation of associated features
  4. Independent evaluation of each candidate phrase through an ML model
  5. Final re-ranking to capture relationships between candidates
- Step 1-3 are applied on the training set to create the ML models