Entity Linking Techniques

String Matching

- Exact matching
- Fuzzy matching
- Levenshtein distance
- Soundex, Metaphone

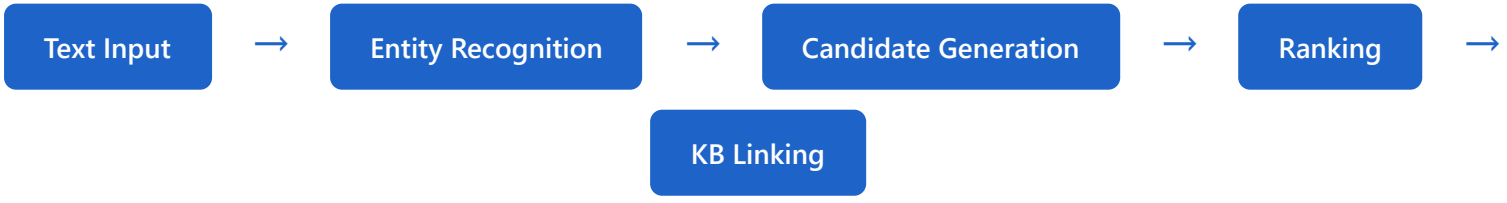
Semantic Similarity

- Word embeddings
- BERT embeddings
- Cosine similarity
- Semantic distance

Context-based Linking (Ensemble)

Combining rule-based + string matching + semantic similarity
Considering surrounding words and context
Can achieve 90%+ accuracy

📊 Entity Linking Process



🔍 Detailed Technique Descriptions

1. String Matching

Exact Matching: Perfect match search
Example: "Apple" → "Apple Inc."

Fuzzy Matching: Allows typos
Example: "Microsft" → "Microsoft"

Levenshtein Distance: Edit distance calculation
Measures insert/delete/replace operations

2. Semantic Similarity

Word Embeddings: Word vectorization
Utilizing Word2Vec, GloVe

BERT Embeddings: Context-based embeddings
Can distinguish homonyms

Cosine Similarity: Vector similarity
Range -1 to 1, closer to 1 means more similar

3. Rule-based

Naming Rules: Proper noun patterns
Capital letter start, special formats

Domain Knowledge: Field-specific rules
Medical, legal, technical terms

Context Rules: Surrounding word patterns
"CEO of", "located in", etc.

4. Ensemble Methods

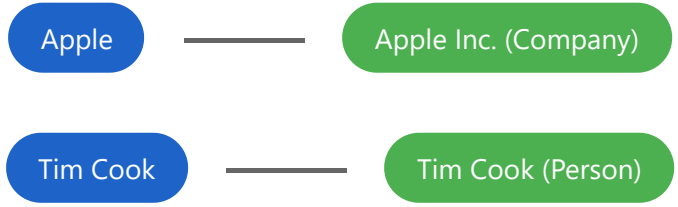
Weighted Combination: Summing scores from each technique
$\alpha \cdot \text{string} + \beta \cdot \text{semantic} + \gamma \cdot \text{rule}$

Voting Approach: Majority decision
Combining predictions from multiple models

Sequential Application: Stepwise filtering
High confidence → low confidence order

💡 Real-world Application Example

Sentence: "Apple's CEO Tim Cook announced a new iPhone"



Step-by-Step Processing

1. **Entity Extraction**
   Identify entities from text through NER → Person names, organization names, place names, etc.

2. **Candidate Generation**
   Select similar entities from knowledge base as candidates → Maximum 10-20 candidates

3. **Context Analysis**
   Analyze surrounding words and sentence structure to understand meaning

4. **Ranking**
   Calculate confidence score for each candidate by combining multiple techniques

5. **Final Linking**
   Select the candidate with the highest score and link to KB entity

Technique Comparison

Technique	Advantages	Disadvantages	Use Cases
String Matching	Fast processing speed Simple implementation	Cannot distinguish homonyms No context consideration	Proper noun search Initial filtering

Technique	Advantages	Disadvantages	Use Cases
Semantic Similarity	Context understanding Excellent synonym handling	High computational cost Requires embedding training	Natural language understanding Semantic search
Rule-based	Domain specialization High interpretability	Difficult maintenance Limited scalability	Specialized domains Structured data
Ensemble	High accuracy Excellent robustness	Increased complexity Requires tuning	Production systems High accuracy requirements

💡 **Practical Tips:**

- Initial Prototype: Start with string matching
- Accuracy Improvement: Add semantic similarity
- Optimization: Apply domain-specific rules
- Production: Integrate with ensemble methods
- Continuous monitoring and feedback incorporation essential

🎯 Key Summary

Entity Linking is the process of connecting entities in text with structured entities in a knowledge base. It uses various techniques from simple string matching to deep learning-based semantic analysis, and in real systems, multiple methods are ensembled to achieve accuracy rates over 90%.