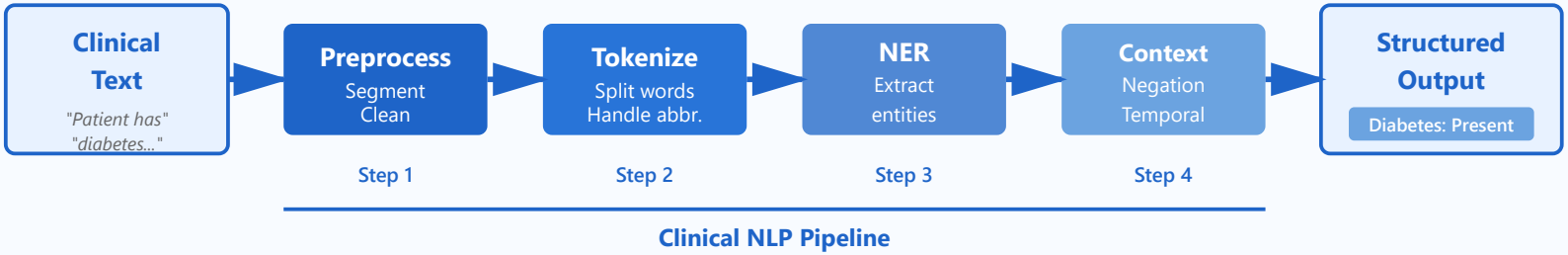
Clinical NLP Basics



Text Preprocessing

- Sentence segmentation
- Lowercasing, punctuation removal
- Handling abbreviations
- PHI removal

Tokenization

- Word-level tokens
- Subword tokenization (BPE)
- Clinical-specific tokenizers
- Handling medical jargon

Named Entity Recognition

- Diseases, symptoms

Negation & Section Detection

- NegEx, ConText algorithms

- Medications, dosages
- Anatomical sites
- Procedures

- Identifying negated findings
- Section headers (HPI, ROS, A&P)
- Temporal expressions

1. Text Preprocessing

Text preprocessing is the crucial first step in clinical NLP that transforms raw, unstructured clinical text into a clean, standardized format suitable for further analysis. This step handles the unique challenges of medical documentation including complex formatting, abbreviations, and protected health information.

Example: Raw Clinical Text Transformation

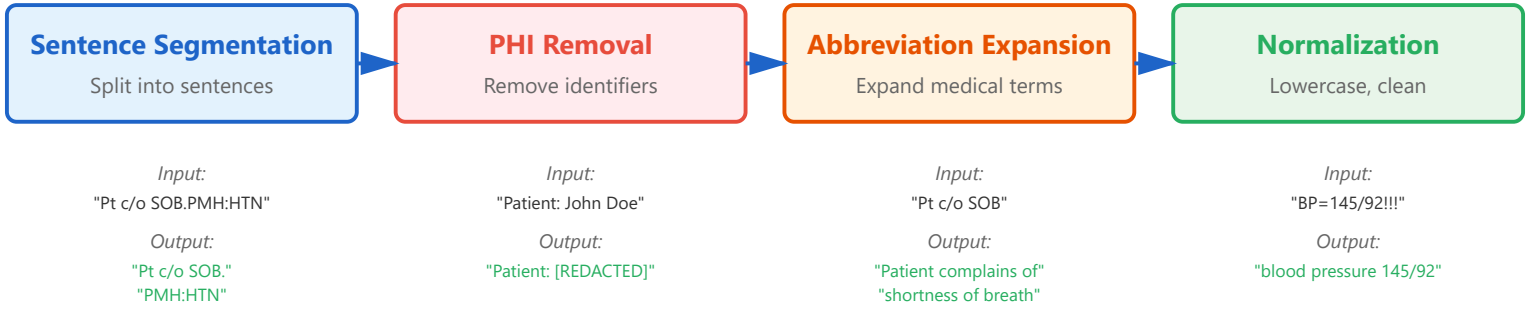
BEFORE (Raw Text)

```
PATIENT: John Doe (DOB: 01/15/1975)
CHIEF COMPLAINT: c/o SOB & CP x3days.
HPI: Pt. presents w/ acute SOB... PMH
includes HTN,DM2, h/o MI in 2019. MEDS:
Metformin 500mg BID;Lisinopril 10mg QD
VITALS: BP=145/92 HR=88 T=98.6F
```

AFTER (Preprocessed)

```
PATIENT: [REDACTED] (DOB: [REDACTED])
CHIEF COMPLAINT: complains of shortness
of breath and chest pain for 3 days HPI:
Patient presents with acute shortness of
breath PMH includes hypertension
diabetes mellitus type 2 history of
myocardial infarction in 2019 MEDS:
Metformin 500 milligrams twice daily
Lisinopril 10 milligrams once daily
VITALS: blood pressure 145/92 heart rate
88 temperature 98.6 fahrenheit
```

Preprocessing Pipeline Visualization



Common Medical Abbreviations

c/o → complains of	SOB → shortness of breath	HTN → hypertension	DM2 → diabetes mellitus type 2	h/o → history of
BID → twice daily	QD → once daily	MI → myocardial infarction	CP → chest pain	PMH → past medical history

Key Points:

- **Sentence Segmentation:** Splits text into individual sentences, handling clinical-specific punctuation patterns (e.g., abbreviations with periods)
- **PHI Removal:** Removes or redacts Protected Health Information (names, dates, IDs) to comply with HIPAA regulations
- **Abbreviation Expansion:** Converts medical abbreviations to full terms, handling context-dependent meanings
- **Normalization:** Standardizes text format including lowercasing, removing extra whitespace, and handling special characters

2. Tokenization

Tokenization breaks preprocessed text into meaningful units (tokens) that can be processed by NLP models. Clinical text presents unique challenges including complex medical terms, compound words, and specialized notation that require domain-specific tokenization strategies.

Comparison: Word-level vs. Subword Tokenization

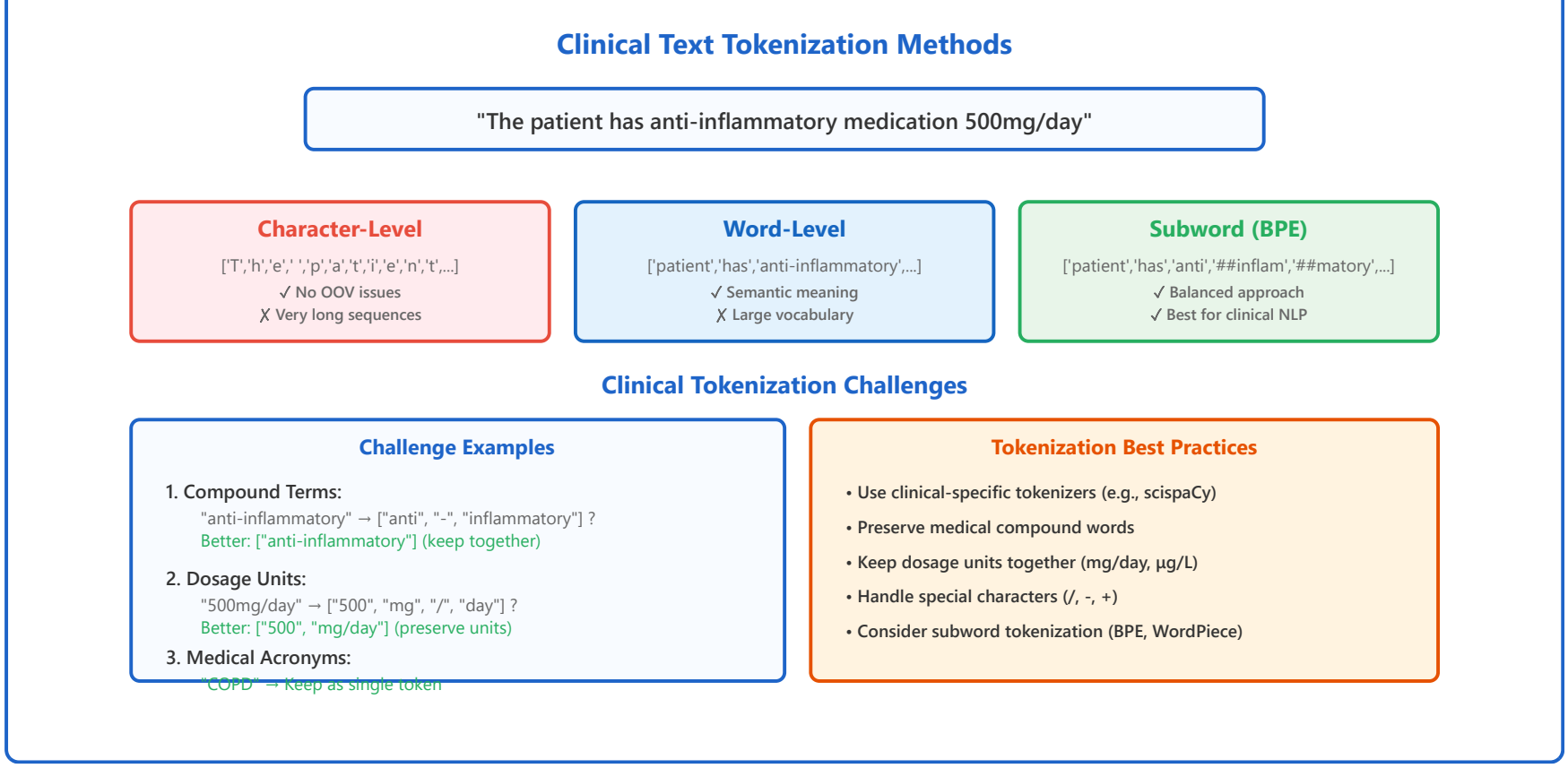
WORD-LEVEL TOKENIZATION

```
Input: "Patient diagnosed with hypothyroidism" Tokens: ["Patient", "diagnosed", "with", "hypothyroidism"] Problem: - Large vocabulary size - Unknown words (OOV) - Can't handle misspellings
```

SUBWORD TOKENIZATION (BPE)

```
Input: "Patient diagnosed with hypothyroidism" Tokens: ["Patient", "diagnosed", "with", "hypo", "##thyroid", "##ism"] Advantages: - Smaller vocabulary - Handles rare words - Better generalization
```

Tokenization Strategies



Real Clinical Text Tokenization Example

```
Input Text: "Patient presents with type-2 diabetes mellitus, prescribed metformin 500mg BID"
Standard Tokenization: ['Patient', 'presents', 'with', 'type', '-', '2', 'diabetes',
'mellitus', ',', 'prescribed', 'metformin', '500', 'mg', 'BID'] Clinical-Aware Tokenization
(Better): ['Patient', 'presents', 'with', 'type-2_diabetes_mellitus', 'prescribed',
'metformin', '500mg', 'BID'] Subword (BPE) Tokenization: ['Patient', 'presents', 'with',
'type', '-', '2', 'diabet', '##es', 'mell', '##itus', 'prescribed', 'met', '##form', '##in',
'500', 'mg', 'BID']
```

Key Points:

- **Word-Level:** Simple but struggles with rare medical terms and large vocabulary requirements
- **Subword (BPE/WordPiece):** Most effective for clinical NLP, balances vocabulary size and semantic meaning
- **Clinical-Specific:** Custom tokenizers (scispaCy, ClinicalBERT tokenizer) better handle medical terminology
- **Special Considerations:** Preserve compound terms, dosage units, and medical abbreviations as single tokens when appropriate

3. Named Entity Recognition (NER)

Named Entity Recognition identifies and classifies key medical entities in clinical text. This is crucial for extracting structured information about diseases, medications, procedures, and anatomical sites from unstructured clinical notes.

Clinical NER Example with Entity Types

Clinical Text:

"Patient diagnosed with type 2 diabetes and hypertension. Started on metformin 500mg twice daily and lisinopril 10mg once daily. Patient reports pain in left knee. Scheduled for MRI scan of knee joint next week."

Extracted Entities:

Diseases/Conditions:
- type 2 diabetes
- hypertension

Medications:
- metformin
- lisinopril

Dosages:
- 500mg twice daily

Anatomical Sites:
- left knee

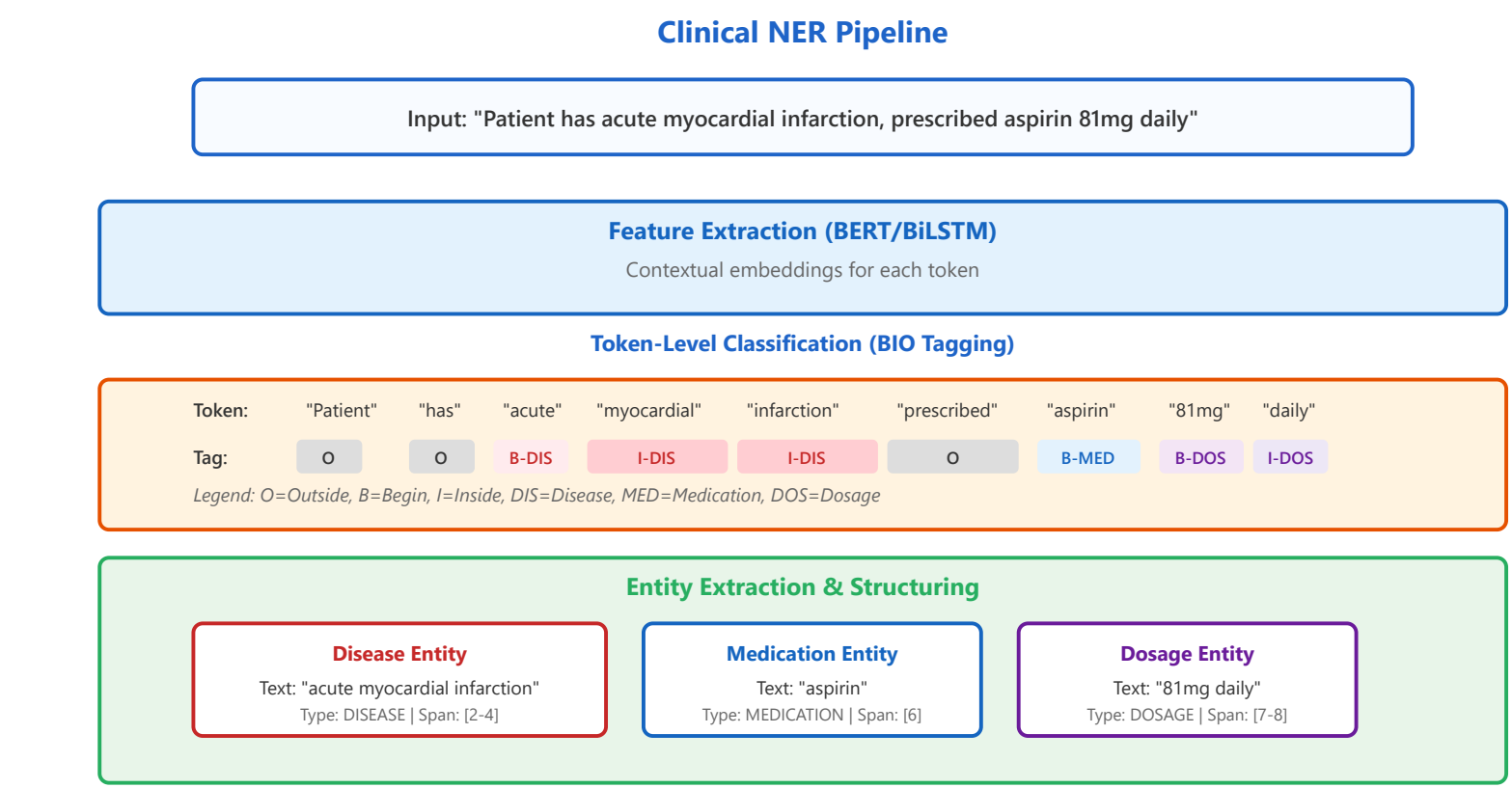
• 10mg once daily

Procedures:

• MRI scan

• knee joint

NER Architecture and Process



Common Clinical Entity Types

Diseases & Conditions

Medications

- Diabetes mellitus
- Myocardial infarction
- Hypertension
- Pneumonia
- Chronic kidney disease

- Metformin
- Lisinopril
- Aspirin
- Insulin
- Warfarin

Anatomical Sites

- Left ventricle
- Right lung
- Anterior chest wall
- Lumbar spine
- Femoral artery

Procedures

- CT scan
- Cardiac catheterization
- Blood transfusion
- Appendectomy
- Colonoscopy

Key Points:

- **BIO Tagging:** Standard approach using Begin-Inside-Outside tags to identify entity boundaries
- **Multiple Entity Types:** Clinical NER must recognize diseases, medications, dosages, anatomy, procedures, and more
- **Context Matters:** Same term can be different entity types based on context (e.g., "Aspirin therapy" vs "Aspirin allergy")
- **Popular Tools:** spaCy with scispaCy models, ClinicalBERT, BioBERT, and custom-trained transformers

4. Negation & Context Detection

Identifying whether a medical finding is affirmed, negated, hypothetical, or historical is critical in clinical NLP. This step prevents false positives by distinguishing between what a patient has versus what they don't have, might have, or had in the past.

Negation Detection Examples

AFFIRMED (Positive)

✓ "Patient has diabetes" → Diabetes: PRESENT ✓ "Confirmed diagnosis of pneumonia" → Pneumonia: PRESENT ✓ "Patient presents with chest pain" → Chest pain: PRESENT

NEGATED (Absent)

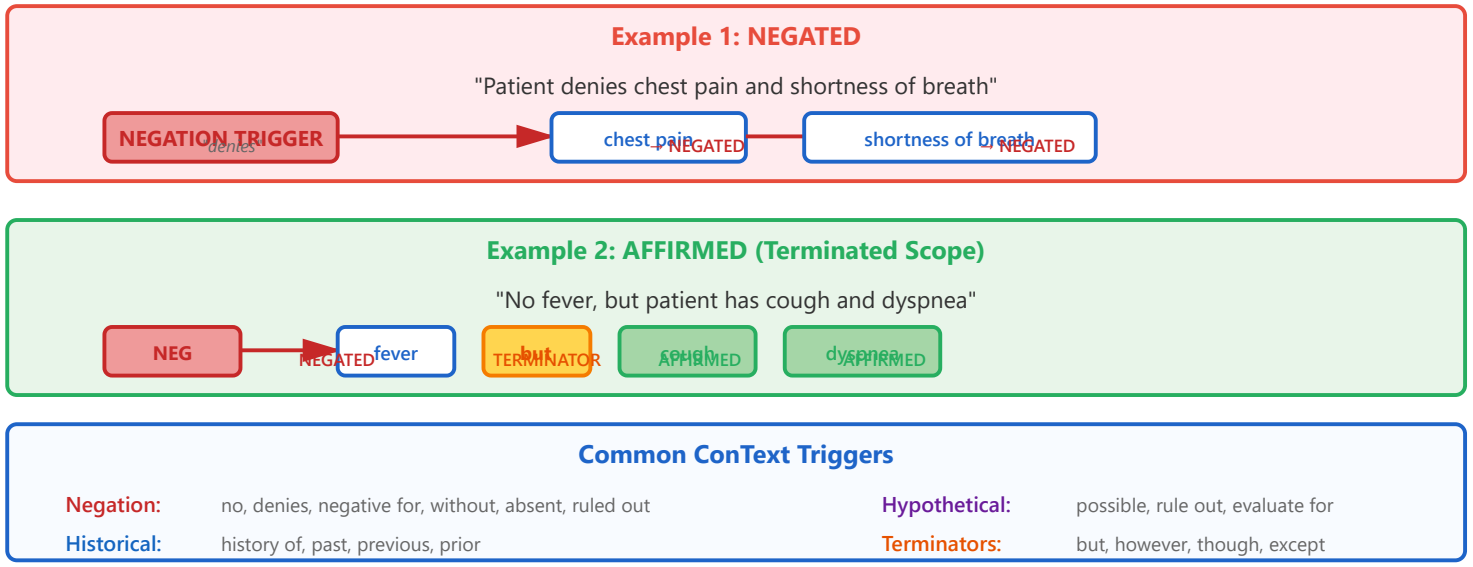
✗ "No evidence of diabetes" → Diabetes: ABSENT ✗ "Patient denies chest pain" → Chest pain: ABSENT ✗ "Without signs of infection" → Infection: ABSENT

⚠ Ambiguous Cases

- "Patient may have diabetes" → **POSSIBLE**
- "History of diabetes" → **HISTORICAL**
- "Family history of diabetes" → **FAMILY_HISTORY**
- "Rule out diabetes" → **HYPOTHETICAL**

ConText Algorithm Visualization

ConText Algorithm: Negation Detection



Section Detection Example

```
CHIEF COMPLAINT: Chest pain HISTORY OF PRESENT ILLNESS: Patient is a 65-year-old male presenting with acute chest pain... PAST MEDICAL HISTORY: - Hypertension (diagnosed 2018) - Type 2 diabetes (diagnosed 2020) - Prior myocardial infarction (2019) CURRENT MEDICATIONS: - Metformin 500mg BID - Lisinopril 10mg daily - Aspirin 81mg daily ASSESSMENT AND PLAN: 1. Acute coronary syndrome - admit for observation 2. Continue current medications 3. Order troponin levels q6h
```

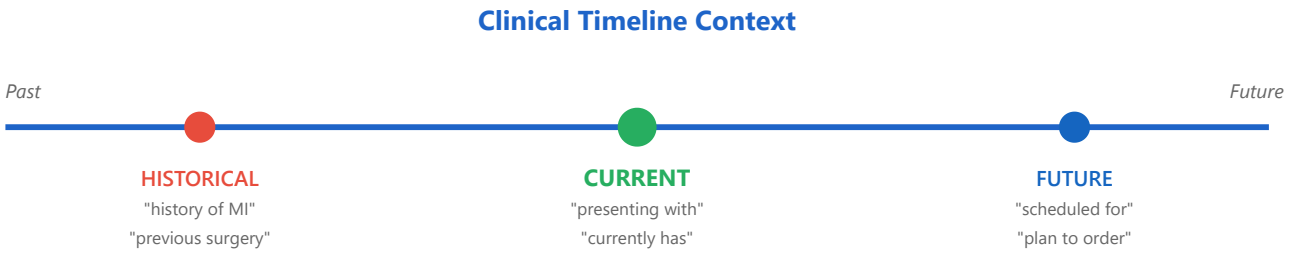
Extracted Section Information:

HPI Section: Acute chest pain → CURRENT, AFFIRMED

PMH Section: MI in 2019 → HISTORICAL, AFFIRMED

Assessment: Acute coronary syndrome → CURRENT, POSSIBLE

Temporal Context Detection



Key Points:

- **NegEx Algorithm:** Rule-based system that identifies negation triggers and their scope within sentences
- **ConText Algorithm:** Extended version handling negation, temporality, experiencer, and hypothetical contexts
- **Section Detection:** Identifies clinical note sections (HPI, PMH, Assessment) to provide contextual information
- **Temporal Expressions:** Distinguishes between current, historical, and future medical conditions
- **Scope Terminators:** Words like "but" and "however" that limit the scope of negation or context triggers

Clinical NLP Pipeline Summary

These four fundamental steps work together to transform unstructured clinical text into structured, actionable data. Each component addresses unique challenges in medical language processing, from handling abbreviations and medical jargon to accurately

identifying negated findings and temporal contexts. Modern clinical NLP systems combine these traditional approaches with deep learning models (BERT, transformers) for improved accuracy and robustness.