Lemmatizer

Topic-wise study notes | Natural Language Processing

Topics: stemming versus lemmatization, dictionary-based lemmatization, and morphological analysis for lemma generation. Includes definitions, characteristics, architectures, worked examples, and applications.

1. Definition of a Lemmatizer

A **lemma** is the conventional dictionary or citation form of a lexical item. A **lemmatizer** maps an inflected word form to its lemma, often using its part of speech, morphological features, and sentence context.

Word form and use	Lemma
cars (plural noun)	car
running (verb)	run
went (past-tense verb)	go
mice (plural noun)	mouse
better (comparative adjective)	good

Outputs depend on language and lexical conventions. For example, better as a verb (“to better a record”) normally keeps the lemma better rather than becoming good.

Characteristics

• Usually returns a lexical base form rather than a truncated stem.

• Handles regular inflections and, with suitable resources, irregular forms.

• Can use part-of-speech information and context to resolve ambiguity.

• May combine a dictionary, morphological rules, and learned models.

• Requires language-specific knowledge or training.

• Does not normally collapse every derived word into its root.

Input and output

Input: a token and its context. Output: a selected lemma, optionally accompanied by part-of-speech and morphological features. The architectures in the following sections show how the main approaches produce this output.

2. Stemming versus Lemmatization

Topic definitions

Stemming: reduces words to stems through word-form operations such as suffix stripping. A stem need not be a dictionary word.
Lemmatization: maps inflected forms to conventional lexical lemmas, potentially considering grammatical information.

Comparison of characteristics

Aspect	Stemming	Lemmatization
Main goal	Group forms under a stem.	Recover a lexical citation form.
Typical approach	Ordered suffix rules.	Lexicon, morphology, context, or learned prediction.
Output	May be a nonword.	Usually a dictionary-form lexical item.
POS dependence	Often no POS analysis.	POS can determine the correct lemma.
Irregular forms	Often remain unrelated.	Can map went to go and mice to mouse.
Resources	Simple rules may suffice.	May need lexical or trained resources.
Cost	A simple rule stemmer is often inexpensive.	Full contextual pipelines may cost more.

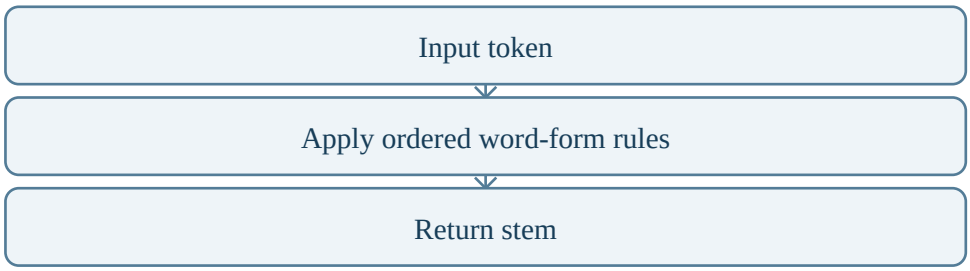
Examples: original Porter versus lemmas

Input	Porter stem	Lemma in stated use
studies (noun/verb)	studi	study
running (verb)	run	run
went (verb)	went	go
mice (noun)	mice	mouse
better (adjective)	better	good
universities (noun)	univers	university

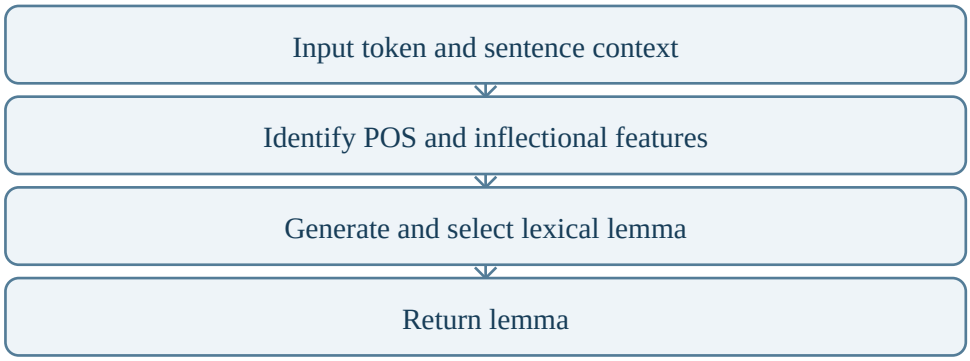
Matching outputs do not imply identical methods: both approaches produce run for running, but stemming does not establish the word’s grammatical analysis.

2.1 Architecture and Worked Comparison

Stemming architecture



Lemmatization architecture



Step-by-step example

Sentence: “The children were running.”

Step	Stemming	Lemmatization
1. Tokens	children, were, running	children, were, running
2. Analysis	Apply original Porter rules to each form.	children: plural noun; were: past be; running: participial verb.
3. Transformation	children remains children; were remains were; running becomes run.	children → child; were → be; running → run.
4. Output	children, were, run	child, be, run

Applications and selection

• **Stemming:** traditional search indexes, keyword grouping, and bag-of-words features.

• **Lemmatization:** readable lexical normalization, grammar-aware analysis, terminology processing, and corpus annotation.

• **Both:** can support retrieval or classification; measure task performance rather than assume one is always better.

Normalized token sequences are analytical representations. They are not necessarily grammatical replacement sentences for display.

3. Dictionary-Based Lemmatization

Definition

Dictionary-based lemmatization uses a lexical resource to map a surface word form to one or more possible lemmas. Entries may also store part of speech and morphological information.

A lexical resource can contain explicit form-to-lemma mappings, lemma entries plus inflection rules, or both. A practical lemmatizer often uses dictionary lookup together with morphological analysis.

Characteristics

• Uses known lexical entries for reliable mappings.

• Supports exceptions such as went → go.

• Can return multiple candidates for ambiguous forms.

• POS information helps distinguish noun and verb uses.

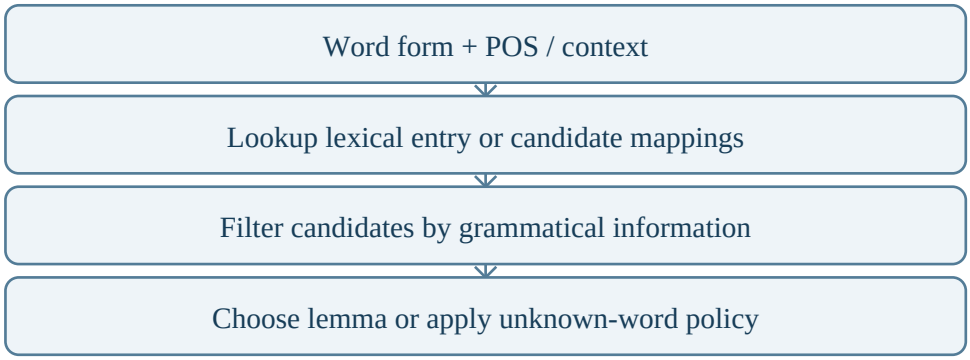
• Coverage depends on dictionary size, language, and domain.

• Unknown words require a fallback or an unresolved result.

Illustrative dictionary entries

Form	POS / features	Lemma
mice	Noun, plural	mouse
went	Verb, past	go
saw	Verb, past	see
saw	Noun, singular	saw
better	Adjective, comparative	good
better	Verb, base form	better

Architecture



3.1 Dictionary Lookup: Worked Examples

Example A: irregular plural

Sentence: “The mice ate cheese.”
Target: mice.

• **Step 1:** Tokenize and locate mice.

• **Step 2:** Identify mice as a plural noun in context.

• **Step 3:** Look up the form mice in the noun lexicon.

• **Step 4:** Retrieve the irregular mapping mice → mouse.

• **Step 5:** Return mouse while preserving plural information separately if needed.

Example B: ambiguous form saw

Context	Analysis	Lemma
She saw a bird.	Saw is the past-tense verb see.	see
The saw is sharp.	Saw names a cutting tool; noun.	saw
They saw wood daily.	Saw is a verb meaning cut with a saw.	saw

Step 1: Retrieve possible entries for saw.
Step 2: Use POS to distinguish the noun use from verb uses.
Step 3: Use context and morphological analysis to distinguish past see from the verb saw.
Step 4: Select see or saw accordingly.

Important: POS alone cannot resolve every ambiguity. Both see and saw can be verb lemmas for the surface form saw.

Applications

• **Lexical search:** match an inflected term to its dictionary entry.

• **Corpus annotation:** label tokens with lexical forms.

• **Domain terminology:** preserve specialist words using a domain lexicon.

• **Language learning:** connect irregular forms to their base entries.

Unknown-word policy

For an unseen technical word or name, a system may preserve the input, apply validated inflection rules, use a learned fallback, or mark the item unresolved. It should not invent a dictionary lemma without support.

4. Morphological Analysis

Definition

Morphological analysis identifies a word's lexical base and grammatical features, including number, tense, person, case, or degree. An analyzer can produce one or more analyses; context then helps choose among them.

Example: cats → lemma cat + POS noun + number plural. The suffix -s expresses plural number.

Key terms

Term	Definition and example
Morpheme	A minimal meaningful or grammatical unit: cat + -s.
Root	Core lexical material of a word.
Stem	Form to which inflection attaches; not necessarily identical to the root.
Lemma	Conventional citation form used to identify the lexical item.
Inflection	Grammatical variation: walk → walked; cat → cats.
Derivation	Formation of another lexical item: happy → happiness.
Allomorph / spelling variant	A variant form associated with a morpheme or base in a particular environment.

Characteristics

• Connects word forms with grammatical features.

• Handles spelling alternations such as studies → study.

• Uses irregular mappings where simple segmentation fails.

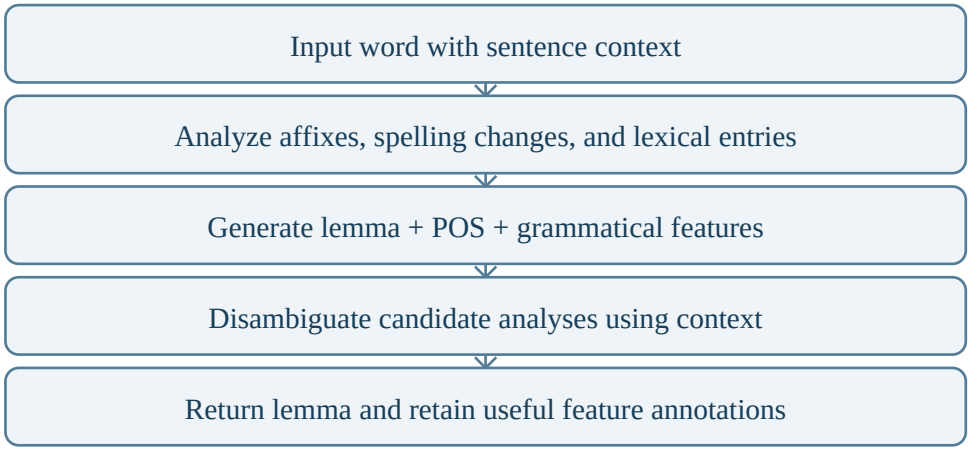
• Can represent richer information than a lemma alone.

• May be rule-based, finite-state, statistical, or neural.

Inflection is not derivation

Lemmatization normally maps happiness to happiness and unhappy to unhappy, rather than removing derivational material to reach happy. A lemmatizer recovers a lexical citation form; it is not automatically a root extractor.

4.1 Morphology-to-Lemma Architecture



Rule-based examples

Observed form	Illustrative analysis
cats	cat + noun + plural; regular -s inflection.
studies	study + noun plural, or study + verb third-person singular.
running	run + verb participial form; reverse consonant doubling.
baked	bake + verb past / participle; restore lexical final e.
went	go + verb past; irregular lexical mapping.

Why dictionary validation is useful

Suffix removal can produce several possible bases. For baked, blindly removing ed gives bak. A morphological rule and lexical knowledge recover bake. Conversely, sing must not become s merely because it ends in ing.

Finite-state view

A finite-state morphological analyzer can encode mappings between lexical forms with features and surface forms. Analysis reads a surface word and returns possible lexical analyses. Generation performs the reverse mapping, for example walk + past → walked. Rules can encode spelling changes, while lexical entries cover exceptions.

Applications

Morphological analysis supports grammar-aware search, corpus annotation, translation, language-learning tools, and analysis of languages with many inflected forms. Lemmatization is one output of this wider analysis.

4.2 Morphology: Step-by-Step Examples

Example A: studies

Sentence: “She studies every evening.”

• **Step 1:** Identify studies as the target word.

• **Step 2:** Determine its contextual POS: verb.

• **Step 3:** Analyze the grammatical form: present tense, third-person singular.

• **Step 4:** Apply the relevant spelling correspondence: study + inflection → studies.

• **Step 5:** Validate study as the lexical base.

• **Output:** lemma = study; POS = verb; tense = present; person = third; number = singular.

Example B: running

Sentence: “They are running.”

• **Step 1:** Determine that running is a verb participial form after are.

• **Step 2:** Recognize the -ing form; stripping it temporarily gives runn.

• **Step 3:** Reverse the applicable consonant-doubling pattern to recover run.

• **Step 4:** Validate run in the lexicon or model.

• **Output:** lemma = run. The full phrase are running expresses progressive aspect; aspect is not determined from the token alone.

Example C: went

Sentence: “He went home.”

• **Step 1:** Identify went as a past-tense verb.

• **Step 2:** Check the irregular-form lexicon or learned analysis.

• **Step 3:** Retrieve the lemma go.

• **Step 4:** Output go and preserve past tense separately.

Went cannot be reliably converted to go through ordinary suffix stripping. The relationship is irregular and must be represented in lexical knowledge or learned parameters.

5. Complete Sentence Analysis

Input: “The children were running and carried boxes.”

Token	Contextual analysis	Lemma
The	Determiner	the
children	Plural noun; irregular form	child
were	Past-tense form of be	be
running	Participial verb after were	run
and	Coordinating conjunction	and
carried	Past-tense verb; spelling change	carry
boxes	Plural noun; -es inflection	box

Step-by-step workflow

• **1. Tokenize:** split into tokens, keeping sentence information.

• **2. Tag POS:** assign contextual grammatical categories.

• **3. Analyze morphology:** identify plural and verb forms.

• **4. Generate candidates:** use regular rules and irregular lexical mappings.

• **5. Select:** validate candidates using grammar and context.

• **6. Output:** the, child, be, run, and, carry, box.

Teaching pseudocode

```
LEMMATIZE(sentence):
    tokens = TOKENIZE(sentence)
    tags = CONTEXTUAL_POS_AND_MORPHOLOGY(tokens)
    results = []
    for token, tag in paired(tokens, tags):
        candidates = LEXICON_AND_RULES(token, tag)
        lemma = SELECT(candidates, sentence, tag)
        if no supported lemma is available:
            lemma = UNKNOWN_WORD_POLICY(token)
        append (token, lemma, tag) to results
    return results
```

This illustrates a hybrid system. A neural lemmatizer may predict the lemma directly rather than perform an explicit dictionary lookup.

6. Applications, Limitations, and Revision

Application	Example use
Information retrieval	Match went with go in a lexical search.
Text classification	Group study, studies, and studied into study features.
Corpus linguistics	Count occurrences by lexical item rather than inflected form.
Translation	Provide lemma and grammatical features to downstream processing.
Language learning	Show irregular forms and their dictionary entries.
Terminology analysis	Normalize domain words with an appropriate lexical resource.

Limitations

• Wrong POS tags can produce the wrong lemma.

• POS alone does not resolve every lexical ambiguity.

• Unknown names, specialist terms, and spelling errors challenge coverage.

• Lexical conventions and language-specific morphology affect output.

• Normalization can discard useful distinctions unless features are retained.

Practice questions and answer hints

1. Define a lemma. A conventional dictionary or citation form of a lexical item.
2. Why does went need lexical knowledge? It is an irregular past form of go, without a regular removable suffix.
3. What is the lemma of studies? Usually study; morphology distinguishes plural noun from third-person singular verb.
4. Why is happiness not normally lemmatized to happy? Happiness is a separate derived lexical item, not merely an inflection of happy.
5. Does a lemma sequence form a grammatical sentence? Not necessarily; it is a normalized representation.

Key takeaway

Stemming produces rule-derived stems. Lemmatization recovers lexical citation forms. Dictionary lookup supplies lexical knowledge, while morphological analysis explains how surface forms express grammatical information. Context selects among ambiguous analyses.

Further reading

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Chapter 2 discusses stemming and lemmatization.
Beesley, K. R., and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications.
Examples and architectures in these notes are illustrative; exact feature labels vary across tools.