MALLA REDDY COLLEGE OF ENGINEERING

NATURAL LANGAUGE PROCESSING

UNIT – I

Finding the Structure of Words: Words and Their Components, Issues and Challenges, Morphological Models

Finding the Structure of Documents: Introduction, Methods, Complexity of the Approaches, Performances of the Approaches, Features

NATURAL LANGAUGE PROCESSING

Finding the Structure of Words: Words and Their Components, Issues and Challenges, Morphological Models

Words and Their Components

Words are defined in most languages as the smallest linguistic units that can form a complete utterance by themselves. The minimal parts of words that deliver aspects of meaning to them are called **morphemes**. Depending on the means of communication, morphemes are spelled out via **graphemes**—symbols of writing such as letters or characters—or are realized through **phonemes**, the distinctive units of sound in spoken language. It is not always easy to decide and agree on the precise boundaries discriminating words from morphemes and from phrases

NATURAL LANGAUGE PROCESSING

Tokens

Suppose, for a moment, that words in English are delimited only by whitespace and punctuation

[3], and consider Example 1–1:

Example 1–1: Will you read the newspaper? Will you read it? I won’t read it.

If we confront our assumption with insights from etymology and syntax, we notice two words here: *newspaper* and *won’t*.

Being a compound word, *newspaper* has an interesting derivational structure.

We might wish to describe it in more detail, once there is a lexicon or some other linguistic evidence on which to build the possible hypotheses about the origins of the word.

In writing, ***newspaper*** **and the associated** concept is distinguished from the isolated *news* and *paper*. In speech, however, the distinction is far from clear, and identification of words becomes an issue of its own.

this kind of **tokenization** and **normalization** may apply to just a limited set of cases, but in other languages,

these phenomena have to be treated in a less trivial manner.

NATURAL LANGAUGE PROCESSING

In the writing systems of Chinese, Japanese [5], and Thai, whitespace is not used to separate words. The units that are delimited graphically in some way are sentences or clauses. In Korean, character strings are called ***eojeol*** ‘word segment’ and roughly correspond to speech or cognitive units, which are usually larger than words and smaller than clauses [6],

as shown in Example 1–2:

Example 1–2: 학생들에게만 주셨는데

*hak.sayng.tul.ey.key.man cwu.syess.nun.te*2

haksayng-tul-eykey-man cwu-si-ess-nunte

student+*plural*+*dative*+only give+*honorific*+*past*+while

while (he/she) gave (it) only to the students

NATURAL LANGAUGE PROCESSING

Lexemes

By the term word, we often denote not just the one linguistic form in the given context but also the concept behind the form and the set of alternative forms that can express it. Such sets are called lexemes or lexical items, and they constitute the lexicon of a language.

Lexemes can be divided by their behavior into the lexical categories of **verbs, nouns, adjectives, conjunctions, particles, or other parts of speech**. The citation form of a lexeme, by which it is commonly identified, is also called its lemma.

When we convert a word into its other forms, such as turning the **singular *mouse*** into the **plural *mice*** or *mouses*, we say we **inflect the lexeme**. When we transform a lexeme into another one that is morphologically related, regardless of its lexical category,

we say we derive the lexeme: for instance, the nouns *receiver* and *reception* are derived from the verb *to receive*.

Example 1–3: Did you see him? I didn’t see him. I didn’t see anyone.

Example 1–3 presents the problem of tokenization of *didn’t* and the investigation of the internal structure of *anyone*. In the paraphrase *I saw no one*, the lexeme *to see* would be inflected into the form *saw* to reflect its grammatical function of expressing positive past tense.

Likewise, *him* is the oblique case form of *he* or even of a more abstract lexeme representing all personal pronouns. In the paraphrase, *no one* can be perceived as the minimal word synonymous with *nobody*.

The difficulty with the definition of what counts as a word need not pose a problem for the syntactic description if we understand *no one* as two closely connected tokens treated as one fixed element.

NATURAL LANGAUGE PROCESSING

Morphemes

Morphological theories differ on whether and how to associate the properties of **word forms** with their **structural components**. These components are usually called **segments** or **morphs**. The morphs that by themselves represent some aspect of the meaning of a word are called **morphemes** of some function.

Human languages employ a variety of devices by which morphs and morphemes are combined into word forms. The simplest morphological process concatenates morphs one by one, as in *dis-agree-ment-s*, where *agree* is a free lexical morpheme and the other elements are bound grammatical morphemes contributing some partial meaning to the whole word.

In a more complex scheme, morphs can interact with each other, and their forms may become subject to additional phonological and orthographic changes denoted as morphophonemic. The alternative forms of a morpheme are termed **allomorphs**.

NATURAL LANGAUGE PROCESSING

Morphemes

Examples of morphological alternation and phonologically dependent choice of the form of a **morpheme are abundant in the Korean language**. In Korean, many morphemes change their forms systematically with the phonological context.

Example 1–5 lists the allomorphs *-ess-*, *-ass-*, *-yess-* of the temporal marker indicating past tense. The first two alter according to the phonological condition of the preceding verb stem; the last one is used especially for

the verb *ha-* ‘do’. The appropriate allomorph is merely concatenated after the stem, or it can be further contracted with it, as was *-si-ess-* into *-syess-* in

Example 1–2. During morphological parsing,

normalization of allomorphs into some canonical form of the morpheme is desirable,

especially because the contraction of morphs interferes with simple segmentation:

Example 1–5: concatenated	contracted
(a) 보았- *po-ass-*	봤- *pwass-* ‘have seen’
(b) 가지었- *ka.ci-ess-*	가졌- *ka.cyess-* ‘have taken’
(c) 하였- *ha-yess-*	했- *hayss-* ‘have done’
(d) 되었- *toy-ess-*	됐- *twayss-* ‘have become’
(e) 놓았- *noh-ass-*	놨- *nwass-* ‘have put’

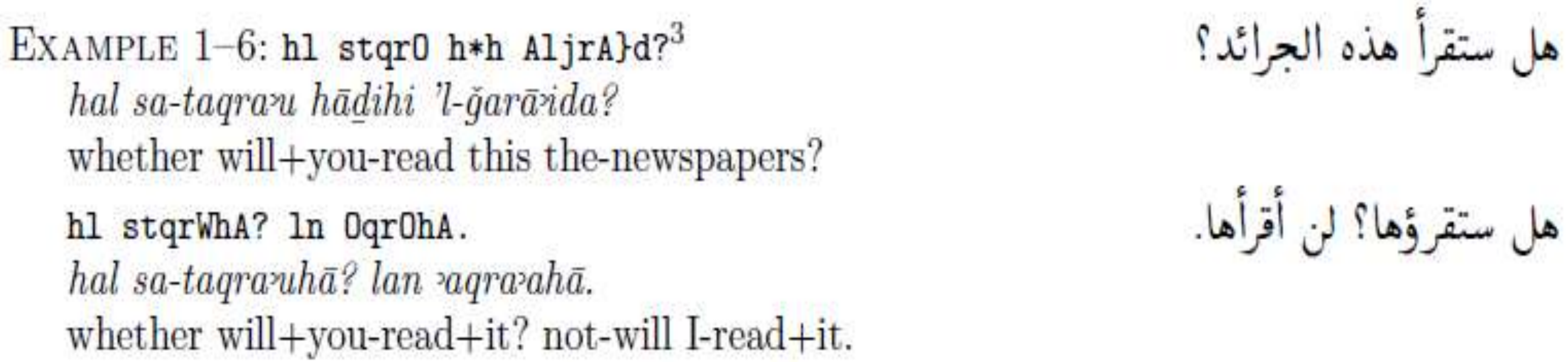
NATURAL LANGAUGE PROCESSING

Morphemes

The meaning of Example 1–6 is similar to that of Example 1–1, only the phrase *h ¯ad ¯ihi ’l-ˇgar ¯aida*

refers to ‘these newspapers’. While *sa-taqrau* ‘you will read’ combines the future marker *sa-* with the imperfective second-person masculine singular verb *taqrau* in the indicative mood and active voice, *sa-taqrauh ¯a* ‘you will read it’ also adds the cliticized feminine singular personal pronoun in the accusative case.4

The citation form of the lexeme to which ***taqrau* ‘you-masc-sg read**’ belongs is ***qara*,** roughly ‘**to read**’. This form is classified by linguists as the basic verbal form represented by the template *faal* merged with the consonantal root *q r* , where the *f l* symbols of the template are substituted by the respective root consonants. Inflections of this lexeme can modify the pattern *faal* of the stem of the lemma into *fal* and concatenate it, under rules of morphophonemic changes, with further prefixes and suffixes. The structure of *taqrau* is thus parsed into the template *ta-fal-u* and the invariant root.



EXAMPLE 1–6: hl stqrO h*h AljrA}d?[3]

هل ستقرأ هذه الجرائد؟

hal sa-taqraʾu hāḏihi ’l-ǧarāʾida?

whether will+you-read this the-newspapers?

hl stqrWhA? ln OqrOhA.

هل ستقرؤها؟ لن أقرأها.

hal sa-taqraʾuhā? lan ʾaqraʾahā.

whether will+you-read+it? not-will I-read+it.

NATURAL LANGAUGE PROCESSING

Morphology is the domain of linguistics that analyses the internal structure of words.

Morphological analysis–exploring the structure of words

Words are built up of minimal meaningful elements called morphemes:

played=play-ed

cats = cat-s

unfriendly = un-friend-ly

Two types of morphemes:

I Stems: play, cat, friend

ii Affixes: -ed, -s, un-, -ly

Two main types of affixes:

I Prefixes precede the stem: un-

Ii Suffixes follow the stem:-ed,-s,un-,-ly

Stemming = find the stem by stripping off affixes

play =play

replayed = re-play-ed

computerized = comput-er-ize-d

NATURAL LANGAUGE PROCESSING

Typology

Morphological typology divides languages into groups by characterizing the prevalent morphological

Phenomena, in those languages. It can consider various criteria, and during the history of linguistics, different classifications have been proposed [13, 14]. **Let us outline the typology that is based on quantitative relations between words, their morphemes,**

and their features:

Isolating, or **analytic**, languages include no or relatively few words that would comprise more than one morpheme (typical members are Chinese, Vietnamese, and Thai; analytic tendencies are also found in English).

Synthetic languages can combine more morphemes in one word and are further divided into agglutinative and fusional languages.

Agglutinative languages have morphemes associated with only a single function at a time (as in Korean, Japanese, Finnish, and Tamil, etc.).

NATURAL LANGAUGE PROCESSING

Typology

Fusional languages are defined by their feature-per-morpheme ratio higher than one (as in Arabic, Czech, Latin, Sanskrit, German, etc.).

In accordance with the notions about word formation processes mentioned earlier,

We can also discern:

Concatenative languages linking morphs and morphemes one after another.

Nonlinear languages allowing structural components to merge non sequentially to apply tonal morphemes or change the consonantal or vocalic templates of words.

While some morphological phenomena, such as orthographic collapsing, phonological contraction, or complex inflection and derivation, are more dominant in some languages than in others, in principle, we can find, and should be able to deal with, instances of these phenomena across different language families and typological classes.

NATURAL LANGAUGE PROCESSING

Issues and challenges

Morphological parsing tries to eliminate or alleviate the variability of word forms to provide higher-level linguistic units whose lexical and morphological properties are explicit and well defined. It attempts to remove unnecessary irregularity and give limits to ambiguity, both of which are present inherently in human language.

By irregularity, we mean existence of such forms and structures that are not described appropriately by a prototypical linguistic model. Some irregularities can be understood by redesigning the model and improving its rules, but other lexically dependent irregularities often cannot be generalized.

Ambiguity is indeterminacy in interpretation of expressions of language. Next to accidental ambiguity and ambiguity due to lexemes having multiple senses, we note the issue of **syncretism**, or systematic ambiguity.

Morphological modeling also faces the problem of productivity and creativity in language, by which unconventional but perfectly meaningful new words or new senses are coined.

Usually, though, words that are not licensed in some way by the lexicon of a morphological system will remain completely unparsed. This **unknown word** problem is particularly severe in speech or writing that gets out of the expected domain of the linguistic model, such as when special terms or foreign names are involved in the discourse or when multiple languages or dialects are mixed together.

NATURAL LANGAUGE PROCESSING

Issues and challenges

- ✓ Irregularity
- ✓ Ambiguity
- ✓ Productivity

irregularity

Morphological parsing is motivated by the quest for generalization and abstraction in the world of words. Immediate descriptions of given linguistic data may not be the ultimate ones, due to either their inadequate accuracy or inappropriate complexity, and better formulations may be needed. The design principles of the morphological model are therefore very important.

In Arabic, the deeper study of the morphological processes that are in effect during inflection and derivation, even for the so-called irregular words, is essential for mastering the whole morphological and phonological system. With the proper abstractions made, irregular morphology can be seen as merely enforcing some extended rules, the nature of which is phonological, over the underlying or prototypical regular word forms

NATURAL LANGAUGE PROCESSING

Table 1–1 illustrates differences between a naive model of word structure in Arabic and the model proposed in Smrˇz [12] and Smrˇz and Bielick´y [17]

where morphophonemic merge rules and templates are involved.

Morphophonemic templates capture morphological processes by just organizing stem patterns and generic affixes without any context-dependent variation of the affixes or ad hoc modification of the stems.

The merge rules, indeed very terse, then ensure that such structured representations can be converted into exactly the surface forms, both orthographic and phonological, used in the natural language.

Applying the merge rules is independent of and irrespective of any grammatical parameters or information

other than that contained in a template.

Most morphological irregularities are thus successfully removed.

Exceptional Inflection comparative and superlative adjectives

Ex:	big	bigger	biggest
	Dark	darker	darkest
	Good	better	best

NATURAL LANGAUGE PROCESSING

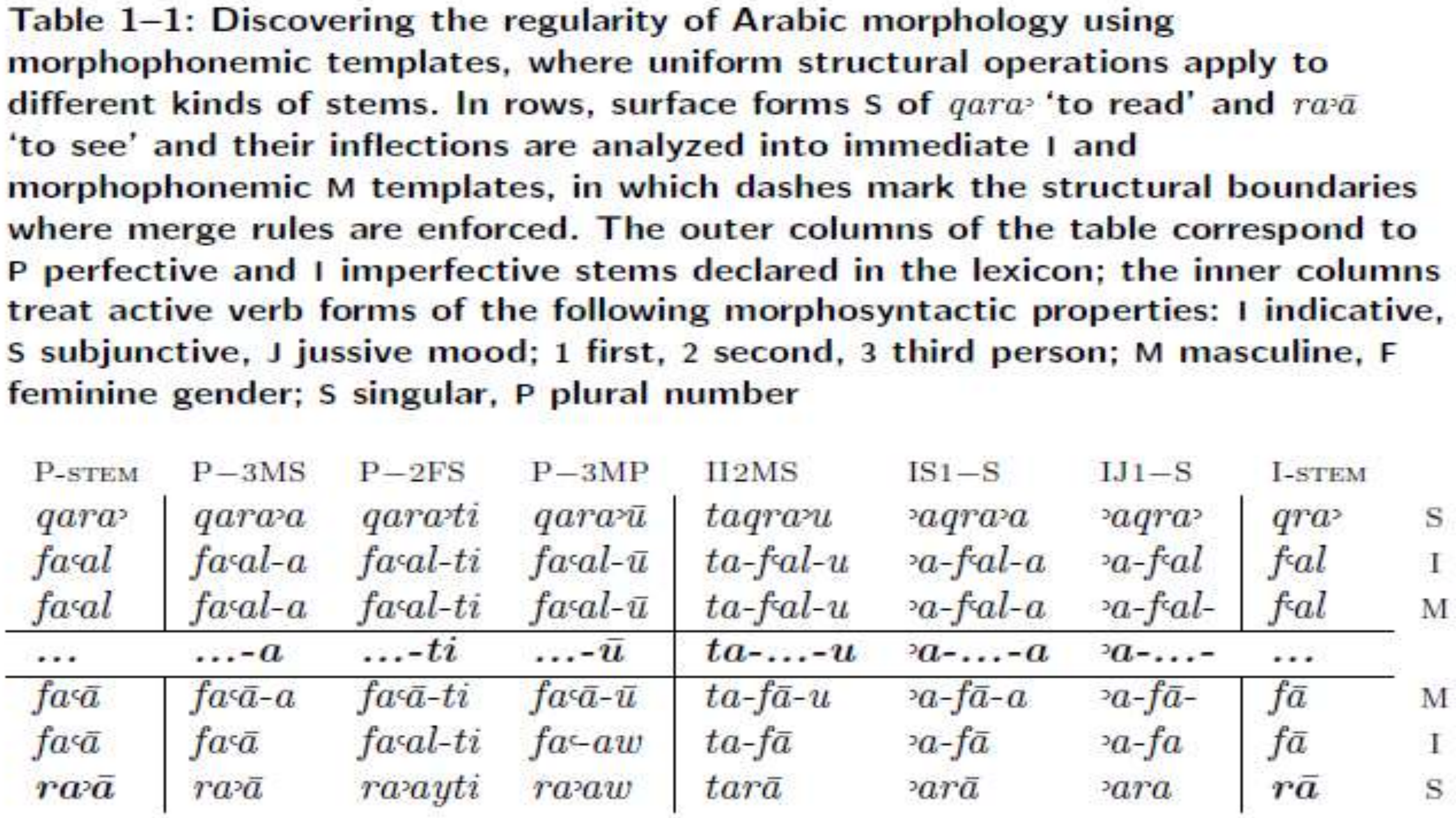


Table 1–1: Discovering the regularity of Arabic morphology using morphophonemic templates, where uniform structural operations apply to different kinds of stems. In rows, surface forms S of *qaraʾ* ‘to read’ and *raʾā* ‘to see’ and their inflections are analyzed into immediate I and morphophonemic M templates, in which dashes mark the structural boundaries where merge rules are enforced. The outer columns of the table correspond to P perfective and I imperfective stems declared in the lexicon; the inner columns treat active verb forms of the following morphosyntactic properties: I indicative, S subjunctive, J jussive mood; 1 first, 2 second, 3 third person; M masculine, F feminine gender; S singular, P plural number

P-STEM	P−3MS	P−2FS	P−3MP	II2MS	IS1−S	IJ1−S	I-STEM	
qaraʾ	*qaraʾa*	*qaraʾti*	*qaraʾū*	*taqraʾu*	*ʾaqraʾa*	*ʾaqraʾ*	*qraʾ*	S
faʿal	*faʿal-a*	*faʿal-ti*	*faʿal-ū*	*ta-fʿal-u*	*ʾa-fʿal-a*	*ʾa-fʿal*	*fʿal*	I
faʿal	*faʿal-a*	*faʿal-ti*	*faʿal-ū*	*ta-fʿal-u*	*ʾa-fʿal-a*	*ʾa-fʿal-*	*fʿal*	M
. . .	***. . .-a***	***. . .-ti***	***. . .-ū***	***ta-. . .-u***	***ʾa-. . .-a***	***ʾa-. . .-***	**. . .**	
faʿā	*faʿā-a*	*faʿā-ti*	*faʿā-ū*	*ta-fā-u*	*ʾa-fā-a*	*ʾa-fā-*	*fā*	M
faʿā	*faʿā*	*faʿal-ti*	*faʿ–aw*	*ta-fā*	*ʾa-fā*	*ʾa-fa*	*fā*	I
raʾā	*raʾā*	*raʾayti*	*raʾaw*	*tarā*	*ʾarā*	*ʾara*	***rā***	S

NATURAL LANGAUGE PROCESSING

Ambiguity

Morphological ambiguity is the possibility that word forms be understood in multiple ways out of the context of their discourse. Words forms that look the same but have distinct functions or meaning are called **homonyms.**

Ambiguity is present in all aspects of morphological processing and language processing at large.

Morphological parsing is not concerned with complete disambiguation of words in heir context, however; it can effectively restrict the set of valid interpretations of a given word form.

In Korean, homonyms are one of the most problematic objects in morphological analysis because they prevail all around frequent lexical items.

Table 1–3 arranges homonyms on the basis of their behavior with different endings. Example 1–8 is an example of homonyms through nouns and verbs.

Word sense ambiguity : meaning depending on the context in which they are used Ex: bank

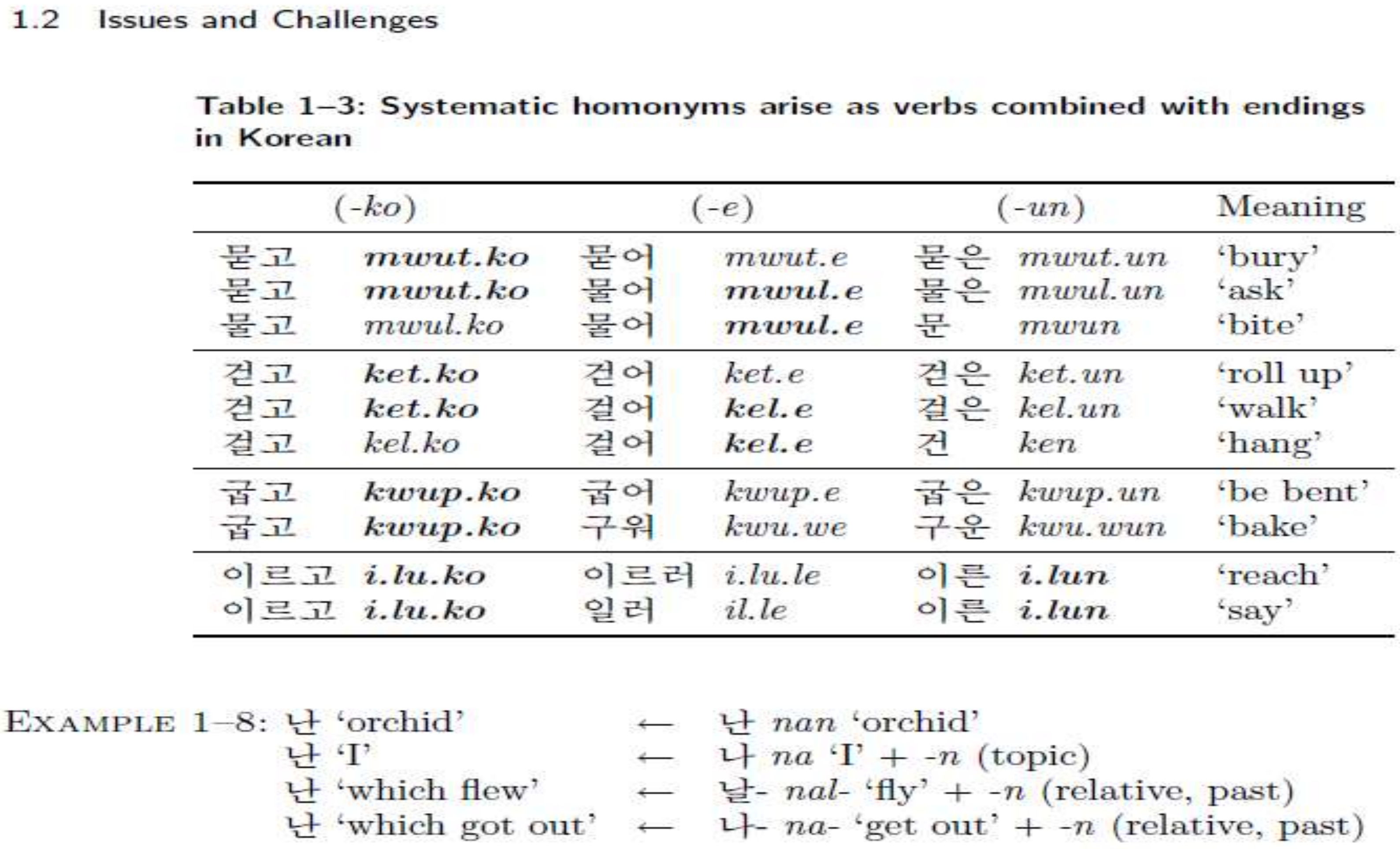
Parts of speech ambiguity : different parts of speech on their usage

Ex: I run (verb) , he went for run (noun)

Structural ambiguity

NATURAL LANGAUGE PROCESSING

Ambiguity



1.2 Issues and Challenges

Table 1–3: Systematic homonyms arise as verbs combined with endings in Korean

(*-ko*)		(*-e*)		(*-un*)		Meaning
묻고	***mwut.ko***	묻어	*mwut.e*	묻은	*mwut.un*	‘bury’
묻고	***mwut.ko***	물어	***mwul.e***	물은	*mwul.un*	‘ask’
물고	*mwul.ko*	물어	***mwul.e***	문	*mwun*	‘bite’
걷고	***ket.ko***	걷어	*ket.e*	걷은	*ket.un*	‘roll up’
걷고	***ket.ko***	걸어	***kel.e***	걸은	*kel.un*	‘walk’
걸고	*kel.ko*	걸어	***kel.e***	건	*ken*	‘hang’
굽고	***kwup.ko***	굽어	*kwup.e*	굽은	*kwup.un*	‘be bent’
굽고	***kwup.ko***	구워	*kwu.we*	구운	*kwu.wun*	‘bake’
이르고	***i.lu.ko***	이르러	*i.lu.le*	이른	***i.lun***	‘reach’
이르고	***i.lu.ko***	일러	*il.le*	이른	***i.lun***	‘say’

Example 1–8: 난 ‘orchid’ ← 난 *nan* ‘orchid’
난 ‘I’ ← 나 *na* ‘I’ + *-n* (topic)
난 ‘which flew’ ← 날- *nal-* ‘fly’ + *-n* (relative, past)
난 ‘which got out’ ← 나- *na-* ‘get out’ + *-n* (relative, past)

NATURAL LANGAUGE PROCESSING

Productivity

Is the inventory of words in a language finite, or is it unlimited? This question leads directly to discerning two fundamental approaches to language, summarized in the distinction between ***langue* and *parole*** by Ferdinand de Saussure, or in the competence versus performance duality by Noam Chomsky.

In one view, language can be seen as simply a collection of utterances (parole) actually pronounced or written (performance). This ideal data set can in practice be approximated by linguistic corpora, which are finite collections of linguistic data that are studied with empirical methods and can be used for comparison when linguistic models are developed.

if we consider language as a system (langue), we discover in it structural devices like recursion, iteration, or compounding that allow to produce (competence) an infinite set of concrete linguistic utterances. This general potential holds for morphological processes as well and is called morphological productivity

We denote the set of word forms found in a corpus of a language as its vocabulary

The distribution of words [33] or other elements of language follows the “80/20 rule,”

also known as the law of the vital few. It says that most of the word tokens in a given corpus can be identified with just a couple of word types in its vocabulary, and words from the rest of the vocabulary occur much less commonly if not rarely in the corpus.

NATURAL LANGAUGE PROCESSING

Productivity

negation is a productive morphological operation. Verbs, nouns, adjectives, and adverbs can be prefixed with *ne*- to define the complementary lexical concept.

In Example

1–9, *budeˇs* ‘you will be’ is the second-person singular of *b´yt* ‘to be’, and *nebudu* ‘I will not

be’ is the first-person singular of *neb´yt*, the negated *b´yt*. We could easily have *ˇc´ıst* ‘to read’ and *neˇc´ıst* ‘not to read’, or we could create an adverbial phrase like *noviny nenoviny* that would express ‘indifference to newspapers’

Let us give an example where creativity, productivity, and the issue of unknown words meet nicely. According to Wikipedia, the word *googol* is a made-up word denoting the number “one followed by one hundred zeros,” and the name of the company Google is an

inadvertent misspelling thereof. Nonetheless, both of these words successfully entered the lexicon of English where morphological productivity started working, and we now know the verb *to google* and nouns like *googling* or even *googlish* or *googleology* [34].

The original names have been adopted by other languages, too, and their own morphological processes have been triggered. In Czech, one says *googlovat*, *googlit* ‘to google’ or *vygooglovat*, *vygooglit* ‘to google out’, *googlov´an´ı* ‘googling’, and so on.

In Arabic, the names are transcribed as *ˇg ¯uˇg ¯ul* ‘googol’ and *ˇg ¯uˇgil* ‘Google’. The latter one got transformed to the verb *ˇgawˇgal* ‘to google’ through internal inflection, as if there were a genuine root *ˇg w ˇg l*, and the corresponding noun *ˇgawˇgalah* ‘googling’ exists as well.

NATURAL LANGAUGE PROCESSING

Morphological Models

There are many possible approaches to designing and implementing morphological models. Over time, computational linguistics has witnessed the development of a number of formalisms and frameworks, in particular grammars of different kinds and expressive power, with which to address whole classes of problems in processing natural as well as formal languages.

Various domain-specific programming languages have been created that allow us to implement the theoretical problem using hopefully intuitive and minimal programming

effort. These special-purpose languages usually introduce idiosyncratic notations of programs and are interpreted using some restricted model of computation. The motivation for such approaches may partly lie in the fact that, historically, computational resources were too limited compared to the requirements and complexity of the tasks being solved. Other motivations are theoretical given that finding a simple but accurate and yet generalizing model is the point of scientific abstraction.

There are also many approaches that do not resort to domain-specific programming.

They, however, have to take care of the runtime performance and efficiency of the computational model themselves. It is up to the choice of the programming methods and the design style whether such models turn out to be pure, intuitive, adequate, complete, reusable, elegant, or not.

Let us now look at the most prominent types of computational approaches to morphology.

Needless to say, this typology is not strictly exclusive in the sense that comprehensive morphological models and their applications can combine various distinct implementational aspects, discussed next.

NATURAL LANGAUGE PROCESSING

Morphological Models

1.3.1 Dictionary Lookup

1.3.2 Finite-State Morphology

1.3.3 Unification-Based Morphology

1.3.4 Functional Morphology

1.3.5 Morphology Induction

Dictionary Lookup

- Morphological parsing is a process by which word forms of a language are associated with corresponding linguistic descriptions.

- Morphological systems that specify these associations by merely enumerating (is **the act or process of making or stating a list of things one after another)**them case by case do not offer any generalization means.

- Likewise for systems in which analyzing a word form is reduced to looking it up verbatimin word lists, dictionaries, or databases, unless they are constructed by and kept in sync with more sophisticated models of the language.

.

NATURAL LANGAUGE PROCESSING

Morphological Models

1.3.1 Dictionary Lookup

□In this context, a dictionary is understood as a data structure that directly enables obtaining some precomputed results, in our case word analyses.

□The data structure can be optimized for efficient lookup, and the results can be shared. Look up operations are relatively simple and usually quick.

□Dictionaries can be implemented, for instance, as lists, binary search trees, tries, hash tables, and soon.

□Because the set of associations between word forms and their desired descriptions is declared by plain enumeration, the coverage of the model is finite and the generative potential of the language is not exploited.

□Despite all that, an enumerative model is often sufficient for the given purpose, deals easily with exceptions, and can implement even complex morphology.

□For instance, dictionary-based *approaches* to Korean depend on a large dictionary of all possible combinations of allomorphs and morphological alternations.

□These approaches do not allow development of reusable morphological rules, though

.

NATURAL LANGAUGE PROCESSING

1.3.3 Finite-State Morphology

- By finite –state morphological models, we mean those in which the specifications written by human programmers are directly compiled into finite-state transducers.
- The two most popular tools supporting this approach, XFST(Xerox Finite- State Tool)and LexTools.
- Finite-state transducers are computational devices extending the power of finite-state automata.
- They consist of a finite set of nodes connected by directed edges labelled with pairs of input and output symbols.
- In such a network or graph, nodes are also called **states,** while edges are called **arcs.**
- Traversing the network from the set of initial states to the set of final states along the arcs is equivalent to reading the sequences of encountered input symbols and writing the sequences of corresponding output symbols.
- The set of possible sequences accepted by the transducer defines the input language; the set of possible sequences emitted by the transducer defines the output language.

Two types of morphemes:

I Stems: play, cat, friend ii Affixes: -ed, -s, un-, -ly

NATURAL LANGAUGE PROCESSING

1.3.3 **Finite-State Morphology**

Based on finite automata and formal language theory , Used for generation and recognition tasks

Basic formal language theory :

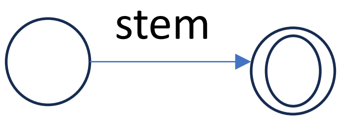
Input Output States Transitions (->)

Finite state transduces : This process is represented by FST's

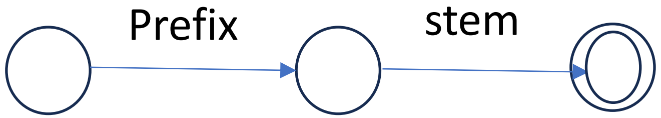
Ex: Word : Classify

FSTs

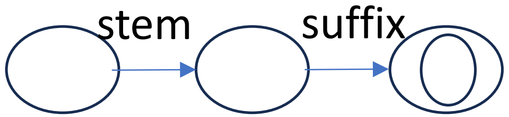
Classify



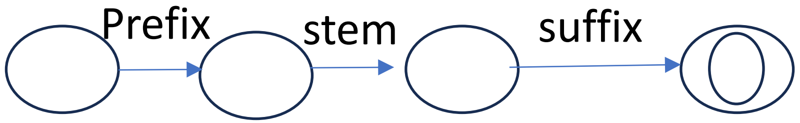
Declassify



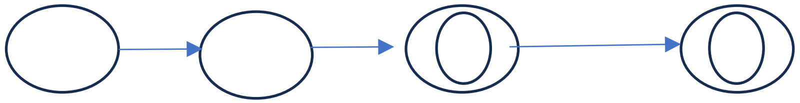
Classification



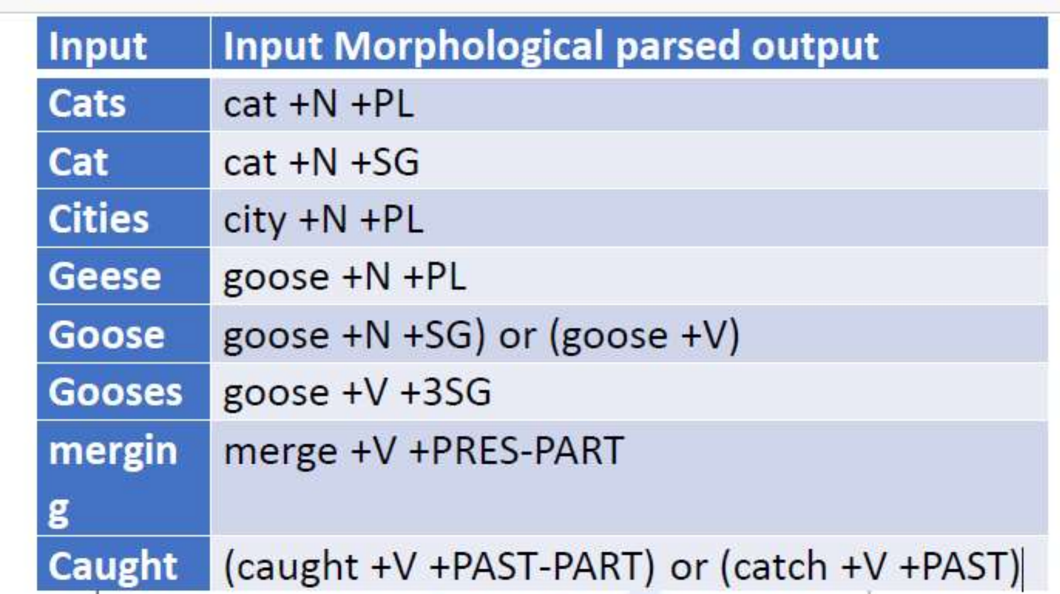
Declassification



.final state Transduces :



NATURAL LANGAUGE PROCESSING



Input	Input Morphological parsed output
Cats	cat +N +PL
Cat	cat +N +SG
Cities	city +N +PL
Geese	goose +N +PL
Goose	goose +N +SG) or (goose +V)
Gooses	goose +V +3SG
merging	merge +V +PRES-PART
Caught	(caught +V +PAST-PART) or (catch +V +PAST)

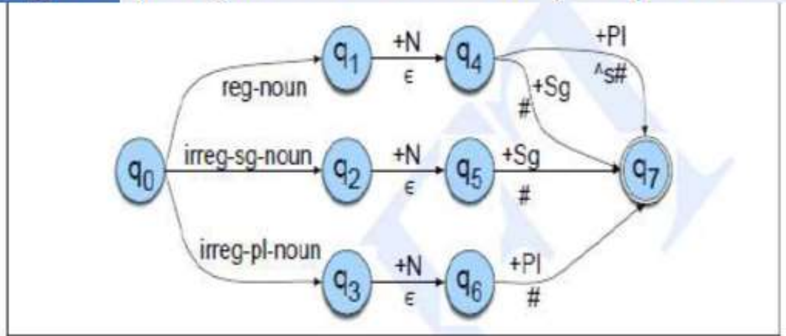


Figure 3.13 A schematic transducer for English nominal number inflection T_{num}. The symbols above each arc represent elements of the morphological parse in the lexical tape; the symbols below each arc represent the surface tape (or the intermediate tape, to be described later), using the morpheme-boundary symbol ^ and word-boundary marker #. The labels on the arcs leaving q_0 are schematic, and need to be expanded by individual words in the lexicon.

NATURAL LANGAUGE PROCESSING

For example, a finite-state transducer could translate the infinite regular language consisting of the words *vnuk*, *pravnuk*, *prapravnuk*,...to the matching words in the infinite regular language defined by *grandson*, *great-grandson*, *great-great-grandson*.

□In finite-state computational morphology, it is common to refer to the input word forms as **surface strings** and to the output descriptions as **lexical strings**, if the transducer is used for morphological analysis, or vice versa, if it is used for morphological generation.

□Relations on languages can also be viewed as functions. Let us have a relation *R*, and let us denote by [Σ] these to fall sequences over some set of symbols Σ,so that the domain and the range of *R* are subsets of [Σ].

□We can then consider *R* as a function mapping an input string in to a set of output strings, formally denoted by this type signature, where [Σ] equals *String*:

□A theoretical limitation of finite-state models of morphology is the problem of capturing **reduplication** of words or their elements (e.g.,to express plurality) found in several human languages.

□Finite-state technology can be applied to the morphological modelling of isolating and agglutinative languages in a quite straight forward manner.

Korean finite- state models are discussed by Kim, Lee and Rim, and Han, to mention a few.

•In English, a finite-state transducer could analyze the surface string children into the lexical

String child[+plural], for instance, or generate women from woman[+plural].

NATURAL LANGAUGE PROCESSING

Unification based morphology

The concepts and methods of these formalisms are often closely connected to those of logic programming.

- Infinite state morphological models, both surface and lexical forms are by themselves unstructured strings of atomic symbols.
- In higher-level approaches, linguistic information is expressed by more appropriated at a structures that can include complex values or can be recursively nested if needed.
- Morphological parsing *P* thus associates linear forms φ with alternatives of structured content ψ, cf.

$$\mathcal{P} :: \phi \rightarrow \{\psi\} \qquad \mathcal{P} :: form \rightarrow \{content\} \qquad (1.2)$$

Erjavec argues that for morphological modelling, word forms are best captured by regular expressions, while the linguistic content is best described through **typed feature structures**.

- Feature structures can be viewed as directed acyclicgraphs.
- A node in a feature structure comprises a set of attributes whose values can be

NATURAL LANGAUGE PROCESSING

Unification based morphology

Nodes are associated with types, and atomic values are attribute less nodes distinguished by their type.

- Instead of unique instances of values every where, references can be used to establish value instance identity.
- Feature structures are usually displayed as attribute-value matrices or as nested symbolic expressions.
- Unification is the key operation by which feature structures can be merged into a more informative feature structure.
- Unification of feature structures can also fail, which means that the information in the mis mutually incompatible.
- Morphological models of this kind are typically formulated as logic programs, and unification is used to solve the system of constraints imposed by the model.
- Advantages of this approach include better abstraction possibilities for developing a morphological grammar as well as elimination of redundant information from it.
- Unification-based models have been implemented for Russian, Czech, Slovene, Persian, Hebrew, Arabic, and other languages.

Functional morphology

Morphology Induction

NATURAL LANGAUGE PROCESSING

Functional morphology

Functional morphology defines its models using principles of functional programming and type theory.

It treats morphological operations and processes as pure mathematical functions and organizes the linguistic as well as abstract elements of a model into distinct types of values and type classes.

Though functional morphology is not limited to modelling particular types of morphologies in human languages, it is especially useful for fusional morphologies.

Linguistic notions like paradigms, rules and exceptions, grammatical categories and parameters, lexemes, morphemes, and morphs can be represented intuitively (with out conscious reasoning; instinctively) and succinctly (in a brief and clearly expressed manner)in this approach.

Functional morphology implementations are intended to be reused as programming libraries capable of handling the complete morphology of a language and to be incorporated into various kinds of applications

Morphology Induction

NATURAL LANGAUGE PROCESSING

Morphological parsing is just one usage of the system, the others being morphological generation, lexicon browsing, and soon.

□we can describe inflection *I*, derivation *D*, and look up *L* as functions of these generic type

□Many functional morphology implementations are embedded in a general-purpose programming language, which gives programmers more freedom with advanced programming techniques and allows them to develop full-featured, real-world applications for their models.

□The Zen tool kit for Sanskrit morphology is written in OCaml.

□It influenced the functional morphology frame work in Haskell, with which morphologies of Latin, Swedish, Spanish, Urdu, and other languages have been implemented.

□In Haskell, in particular, developers can take advantage of its syntactic flexibility and design their own notation for the functional constructs that model the given problem.

The notation then constitutes a so-called domain-specific embedded language, which makes programming even more fun.

□Evenwithouttheoptionsprovidedbygeneral-purposeprogramminglanguages,functionalmorphologymodelsachievehighlevelsofabstraction.

□MorphologicalgrammarsinGrammaticalFrameworkcanbeextendedwithdescriptionsofthesyntaxandsemanticsofalanguage.

□GrammaticalFrameworkitselfsupportsmultilinguality,andmodelsofmorethanadozenlanguagesareavailableinitasopen-sourcesoftware

NATURAL LANGAUGE PROCESSING

2.1.1 Sentence Boundary Detection

- **Sentence boundary detection** (Sentence segmentation) deals with automatically segmenting a sequence of word tokens into sentence units.

- In written text in English and some other languages, the beginning of a sentence is usually marked with an uppercase letter, and the end of a sentence is explicitly marked with a period(.), a question mark(?), an exclamation mark(!),or an other type of punctuation.

- In addition to their role as sentence boundary markers, capitalized initial letters are used distinguish proper nouns, periods are used in abbreviations, and numbers and punctuation marks are used inside proper names.

- The period at the end of an abbreviation can mark a sentence boundary at the same time.

- **Example: I spoke with Dr. Smith. and My house is on Mountain Dr.**

- In the first sentence, the abbreviation Dr. does not end a sentence, and in the second it does.

- Especially **quoted sentences** are always problematic, as the speakers may have uttered multiple sentences, and sentence boundaries inside the quotes are also marked with punctuation marks.

- An automatic method that outputs word boundaries a sending sentences according to the presence of such punctuation marks would result in cutting some sentences incorrectly.

.

NATURAL LANGAUGE PROCESSING

2.1.1 Sentence Boundary Detection

- Ambiguous abbreviations and capitalizations are not only problem of sentence segmentation in written text.

- Spontaneously written texts, such as short message service (SMS) texts or instant messaging (IM) texts, tend to be non grammatical and have poorly used or missing punctuation, which makes sentence segmentation even more challenging.

- Similarly, if the text input to be segmented into sentences comes from an **automatic system**, such as optical character recognition (OCR) or ASR, that aims to translate images of handwritten, typewritten, or printed text or spoken utterances into machine editable text, the finding of sentences boundaries must deal with the errors of those systems as well.

- On the other hand, for conversational speech or text or multiparty meetings with ungrammatical sentences and disfluencies, in most cases it is not clear where the boundaries are.

- Code switching –that is, the use of words, phrases, or sentences from multiple languages by multilingual speakers-is another problem that can affect the characteristics of sentences.

.

NATURAL LANGAUGE PROCESSING

2.1.1 Sentence Boundary Detection

- For example, when switching to a different language, the writer can either keep the punctuation rules from the first language or resort to the code of the second language.

- Conventional rule-based sentence segmentation systems in well-formed texts rely on patterns to identify potential ends of sentences and lists of abbreviations for disambiguating them.

- Forexample,ifthewordbeforetheboundaryisaknownabbreviation,suchas“Mr.”or“Gov.,”thetextisnotsegmentedatthatpositioneventhoughsomeperiodsareexceptions.

- Toimproveonsucharule-basedapproach,sentencesegmentationisstatedasaclassificationproblem.

- Giventhetrainingdatawhereallsentenceboundariesaremarked,wecantrainaclassifiertorecognizethem.

.

.

NATURAL LANGAUGE PROCESSING

2.1.2Topic Boundary Detection

- **Segmentation** (Discourse or text segmentation) is the task of automatically dividing a stream of text or speech into topically homogenous blocks.

- This is, given a sequence of(written or spoken) words, the **aim of topic segmentation** is to find the boundaries where topics change.

- Topic segmentation is an important task for various language understanding applications, such as information extraction and retrieval and text summarization.

- For example, in information retrieval, if a long documents can be segmented into shorter, topically coherent segments, then only the segment that is about the user’s query could be retrieved.

- During the late 1990s, the U.S defence advanced research project agency (DARPA) initiated the **topic detection and tracking program** to further the state of the art in finding and following **new topic** in a stream of broadcast news stories.

- One of the tasks in the TDT effort was segmenting a **news stream into individual stories**.

NATURAL LANGAUGE PROCESSING

2.2 Methods

Sentence segmentation and topic segmentation have been considered as a **boundary classification problem**.

Given a boundary candidate (between two word tokens for sentence segmentation and between two sentences for topic segmentation),the goal is to predict whether or not the candidate is an actual boundary (sentence or topic boundary).

Formally, let **x Ɛ X be the vector of features** (the observation) associated with a candidate and **y Ɛ Y be the label** predicted for that candidate.

The label y can be **b for boundary** and $\bar{b}$ **for non boundary**.

Classification problem: given a set of training examples $(x, y)_{train}$,find a function that will assign the most accurate possible label y of unseen examples x_{unseen}.

Alternatively to the binary classification problem, it is possible to model boundary types using finer-grained categories.

For segmentation in text be framed as a three-class problem: sentence boundary b^{a},with out an abbreviation and abbreviation b^{a} not as a boundary b^{-a}

Similarly spoken language, a three way classification can be made between non-boundaries $\bar{b}$ Statements b^{s}, and question boundaries b^{q}.

•For sentence or topic segmentation, the problem is defined as finding the most probable sentence or topic boundaries.

•The natural unit of sentence segmentation is words and of topic segmentation is sentence, as we can assume that topics typically do not change in the middle of a sentences.

NATURAL LANGAUGE PROCESSING

2.2 Methods

The words or sentences are then grouped into categories stretches belonging to one sentences or topic-that is word or sentence boundaries are classified into sentences or topic boundaries and-non-boundaries.

- The classification can be done at each potential boundary *i* (local modelling);then, the aim is to estimate the most probable boundary type y_i for each candidate xi

$$\hat{y} = \underset{y_i \, in \, Y}{argmax} P(y_i|x_i)$$

Here, the ^ is used to denote estimated categories, and a variable with out a ^ is used to show possible categories.

- In this formulation, a category is assigned to each example in isolation; hence, decision is made locally.

- However, the consecutive types can be related to each other.

For example, in broad cast news speech, two consecutive sentences boundaries that form a single word sentence are very in frequent.

- In local modelling, features can be extracted from surrounding example context of the candidate boundary to model such dependencies.

NATURAL LANGAUGE PROCESSING

2.2 Methods

It is also possible to see the candidate boundaries as a sequence and search for the sequence of boundary types $\bar{Y} = \overline{y1} \ldots \ldots .. \bar{y}n$

that have the maximum probability given the candidate examples, $\bar{X} = \overline{x1} \ldots \ldots .. \bar{x}n$

We categorize the methods into local and sequence classification.

Another categorization of methods is done according to the type of the machine learning algorithm: **generative** versus **discriminative**.

Generative sequence models estimate the **joint distribution** of the observations P(X,Y)(words, punctuation) and the labels (sentence boundary, topic boundary).

Discriminative sequence models, however, **focus on features** that categorize the differences between the labelling of that examples.

NATURAL LANGAUGE PROCESSING

2.2.1GenerativeSequenceClassificationMethods

□Most commonly used generative sequence classification method for topic and sentence is the hidden Markov model(HMM).

□The probability in equation2.2 is rewritten as the following, using the Bayes rule:

$$Y = yargmaxP(Y|X) \qquad 2.1$$

$$Y = yargmaxPYX = yargmax\mathrm{T}PXYP(Y)P(X) = yargmaxP(X|YP(Y) \qquad 2.2$$

Here Y=Predicted class(boundary) label

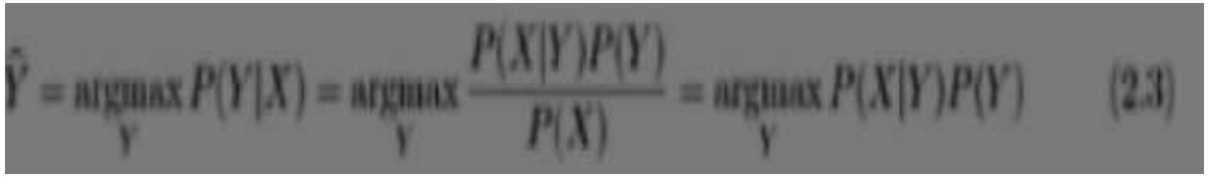
Y=(y1,y2,….yk)=Set of class (boundary) labels

X=(x1,x2,….xn)=set of feature vectors

P(Y|X)=the probability of given the X (feature vectors), what is the probability of X belongs to the class (boundary) label.

P(x)=Probability of word sequence

P(Y) = Probability of the class(boundary)

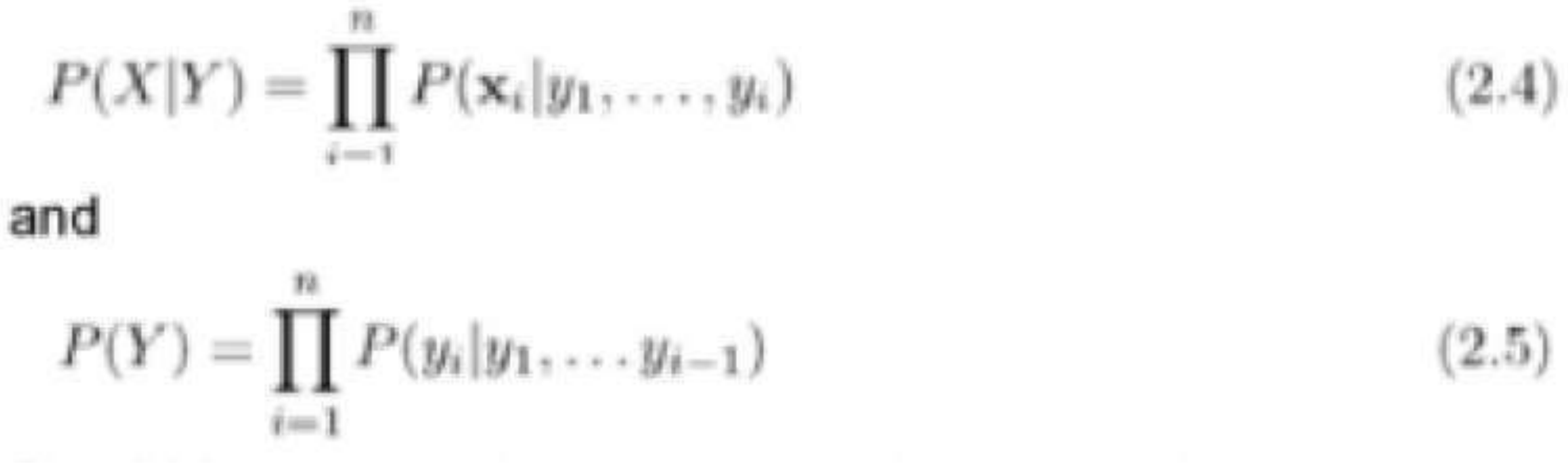


$$\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y|X) = \underset{Y}{\operatorname{argmax}} \frac{P(X|Y)P(Y)}{P(X)} = \underset{Y}{\operatorname{argmax}} P(X|Y)P(Y) \qquad (2.3)$$

NATURAL LANGAUGE PROCESSING

P(X) in the denominator is dropped because it is fixed for different Y and hence does not change the argument of max.

- P(X|Y) and P(Y) can be estimated as



$$P(X|Y) = \prod_{i=1}^{n} P(\mathbf{x}_i|y_1, \ldots, y_i) \qquad (2.4)$$

and

$$P(Y) = \prod_{i=1}^{n} P(y_i|y_1, \ldots y_{i-1}) \qquad (2.5)$$

NATURAL LANGAUGE PROCESSING

2.2.2DiscriminativeLocalClassificationMethods

- Discriminative classifiers aim to model **P(yi|xi) equation 2.1 directly**.
- The most important distinction is that where as c**lass densities P(x|y)** are model assumptions **in generative approaches**, such as naïveBayes, in discriminative methods, discriminant functions of the feature space define the model.
- A number of discriminative classification approaches, such as **support vector machines**, boosting, **maximum entropy**, and regression. Are based on very different machine learning algorithms.
- While discriminative approaches have been shown to out perform generative methods in many speech and language processing tasks.
- For **sentence segmentation, supervised learning methods have primarily been applied to news paper articles**.
- Stamatatos, Fakotakis and Kokkinakis used **transformation based learning** (TBL)to infer rules for **finding sentence boundaries**.

NATURAL LANGAUGE PROCESSING

Many classifiers have been tried for the task: regression trees, neural networks, classification trees, maximum entropy classifiers, support vector machines, and naïve Bayes classifiers.

The most Text tiling method Hearst for topic segmentation uses a l**exical cohesion metric** in a word vector space as an indicator of topic similarity.

Figure depicts a typical graph of similarity with respect to **consecutive segmentation units**

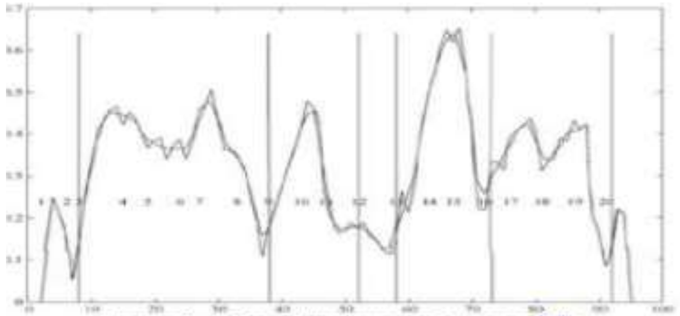


Figure 2–4. Text Tiling example (from [22])

Originally, two methods for computing the similarity scores were

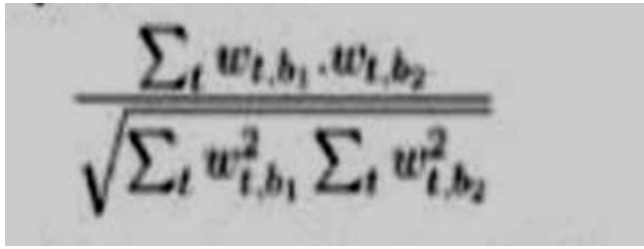
The document is chopped when the similarity is below some threshold.

Originally, **two methods** for computing the similarity scores were proposed: **block comparison** and **vocabulary introduction**.

NATURAL LANGAUGE PROCESSING

The first, **block comparison, compares adjacent blocks of text to see how similar** they are according to how many words the adjacent blocks have in common.

Given two blocks, b1 and b2, each having k tokens (sentences or paragraphs), the **similarity (or topical cohesion) score** is computed by the **formula**:



$$\frac{\sum_t w_{t,b_1} \cdot w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}}$$

Where w t,b is the weight assigned to term t in block b.

The weight scan be binary or may be computed using other information retrieval-metrics such as term frequency.

The second, the **vocabulary introduction method**, **assigns a score to a token-sequence gap** on the basis of **how many new words are seen in the interval in which it is the midpoint**

NATURAL LANGAUGE PROCESSING

Similar to the **block comparison formulation, given two consecutive blocks b1 and b2, of equal number of words w**, the topic alcohesion score is computed with the following formula:

- Where **Num NewTerms (b) returns the number of terms in block b seen the first time in text**.

2.2.3DiscriminativeSequenceClassificationMethods

- In segmentation tasks, the sentence or topic decision for a given example (word, sentence, paragraph) highly depends on the decision for the examples in its vicinity.
- Discriminative sequence classification methods are in general extensions of local discriminative models with additional decoding stages that find the best assignment of labels by looking at neighbouring decisions to label.
- **Conditional random fields** (CRFs) are extension of maximum entropy, **SVM** struct is an extension of SVM, and **maximum margin Markov networks** (M3N) are extensions of HMM.
- CRFs are a class of log-linear models for labelling structures

Contrary to local classifiers that predict sentences or topic boundaries independently, CRF scan over see the whole sequence of boundary hypotheses to make their decisions.

NATURAL LANGAUGE PROCESSING

Complexity of the Approaches

- The approaches described here have **advantages** and **disadvantages**.
- In a **given context** and under a set of observation features, one **approach may be better than other**.
- These approaches can be rated in terms of **complexity** (time and memory) of **their training** and **prediction algorithms** and in terms of their **performance on real-world datasets**.
- In terms of complexity, **training of discriminative approaches** is **more complex than training of generative ones** because they require multiple passes over the training data to adjust for feature weights.
- However, generative models such as HELM scan handle **multiple orders of magnitude larger training sets** and benefits, for instance, from decades of news wire transcripts.
- On the other hand, they work with **only a few features** (only words for HELM) and do not cope well with unseen events.

NATURAL LANGAUGE PROCESSING

Performance of approaches

NLP approaches

- **Statistical approach**: Uses patterns in large amounts of text to understand human language

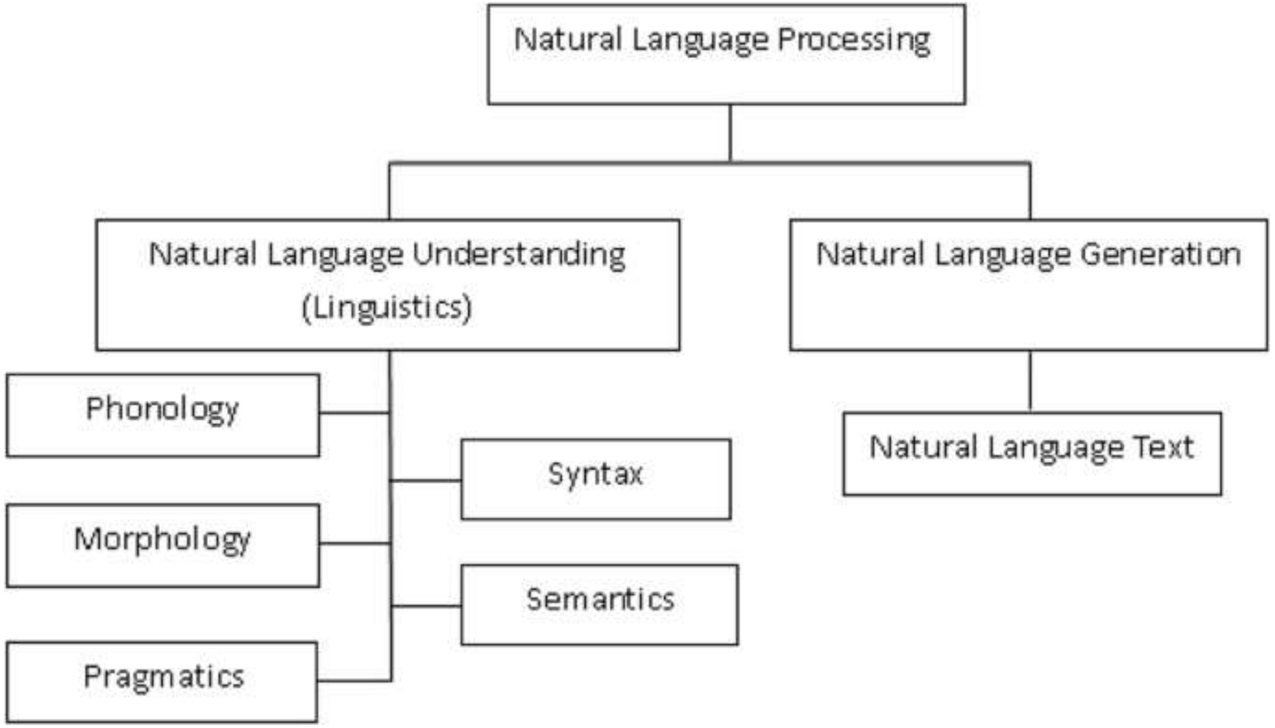
- **Symbolic approach**: Uses human-developed rules, such as grammar rules, to define how the system behaves

- **Connectionist approach**: Combines statistical and symbolic approaches

- **Deep learning**: Uses artificial neural networks to learn from data and recognize complex patterns

NATURAL LANGAUGE PROCESSING

Performance of approaches



Natural Language Processing
Natural Language Understanding (Linguistics)
Natural Language Generation
Phonology
Syntax
Morphology
Semantics
Pragmatics
Natural Language Text

NATURAL LANGAUGE PROCESSING

Features

List the applications in NLP.

Applications of NLP:

•Information retrieval & web search

•Grammar correction & Question answering

•Sentiment Analysis.

•Text Classification.

•Chatbots & Virtual Assistants.

•Text Extraction.

•Machine Translation.

•Text Summarization.

•Market Intelligence.

•Auto-Correct.