Doğal Dil İşlemeye Giriş

Temel Metin İşleme – Kelime Ayırma/Parçalama (Basic Text Processing-Word Tokenization)

Text Normalization

- Every NLP task requires text normalization:
  1. Tokenizing (segmenting) words
  2. Normalizing word formats
  3. Segmenting sentences

How many words?

- I do uh main- mainly business data processing
    - Fragments, filled pauses
- Seuss's cat in the hat is different from other cats!
    - **Lemma:** same stem, part of speech, rough word sense
        - `cat and cats = same lemma`
    - **WordForm:** the full inflected surface form
        - `cat and cats = different wordform`

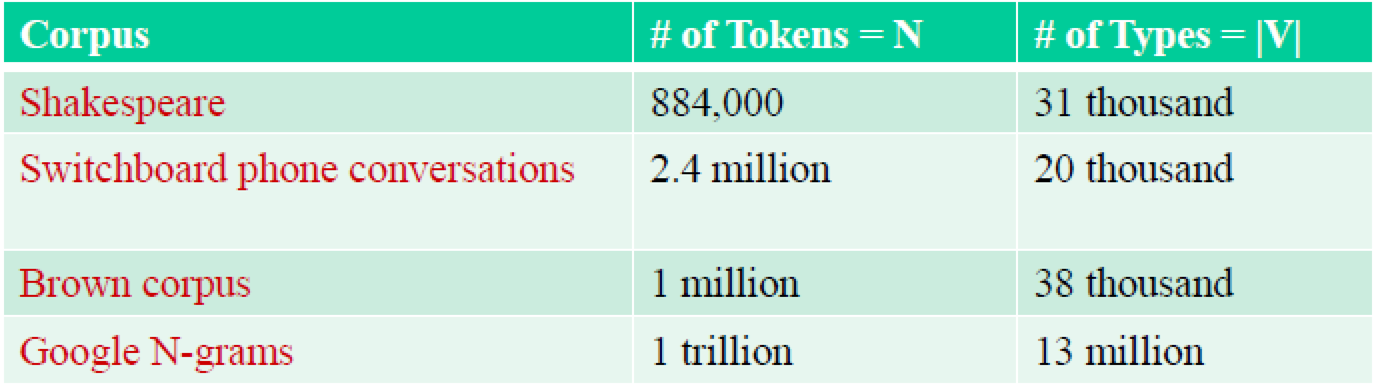
How many words?

they lay back on the San Francisco grass and looked at the stars and their

- **Type:** an element of the vocabulary.
- **Token:** an instance of that type in running text.
- How many?
  - 15 tokens (or 14)
  - 13 types (or 12) (or 11?)

How many words?

- ***N*** = number of token
- ***V*** = vocabulary = set of types
  - |V| is the size of th vocabulary



Corpus	# of Tokens = N	# of Types = |V|
Shakespeare	884,000	31 thousand
Switchboard phone conversations	2.4 million	20 thousand
Brown corpus	1 million	38 thousand
Google N-grams	1 trillion	13 million

Issues in Tokenization

- Finland’s capital → Finland Finlands Finland’s ?
- what’re, I’m, isn’t → What are, I am, is not
- Hewlett – Packard → Hewlett Packard ?
- state-of-the-art → state of the art
- Lowercase → lower-case lowercase lower case?
- San Francisco → one token or two ?
- m.p.h , PhD. → ??

Tokenization: language issues

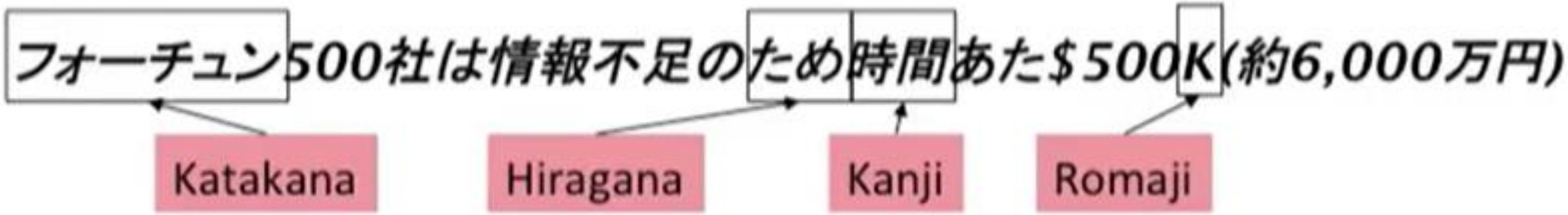
- French
    - ***L’ensemble*** → one token or two?
        - L ? L’? Le?
        - Want **l’ensemble** to match with un ensemble

- German noun compounds are not segmented
    - ***Lebensversicherungsgesellschaftsangestellter***
    - «life insurance company employee»
    - German information retrieval needs compound splitter

Tokenization: language issues

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida

- Further complicated in Japanese, with multiple alphabets intermingled



フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

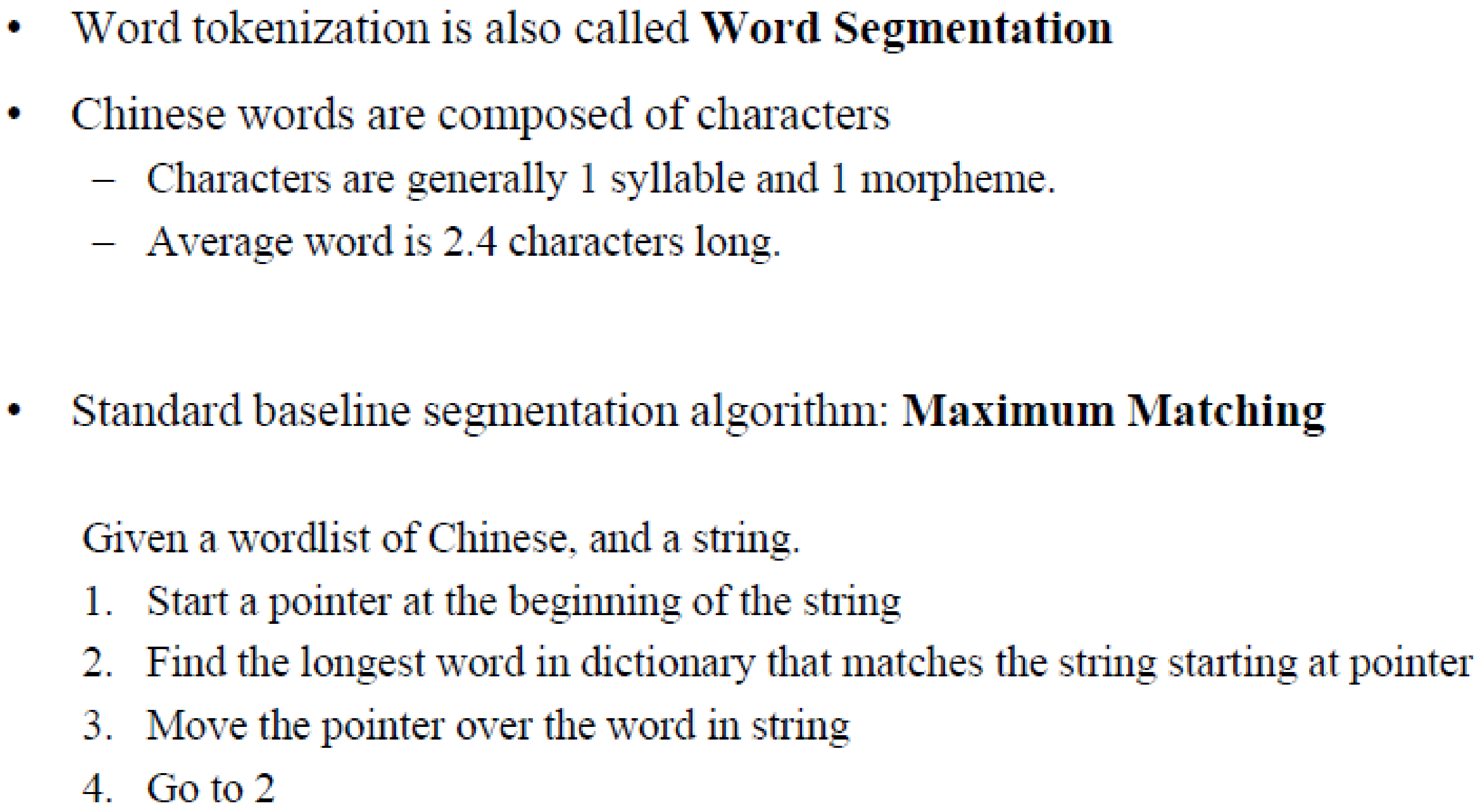
Katakana

Hiragana

Kanji

Romaji

Word tokenization in Chinese



- Word tokenization is also called **Word Segmentation**
- Chinese words are composed of characters
  - Characters are generally 1 syllable and 1 morpheme.
  - Average word is 2.4 characters long.

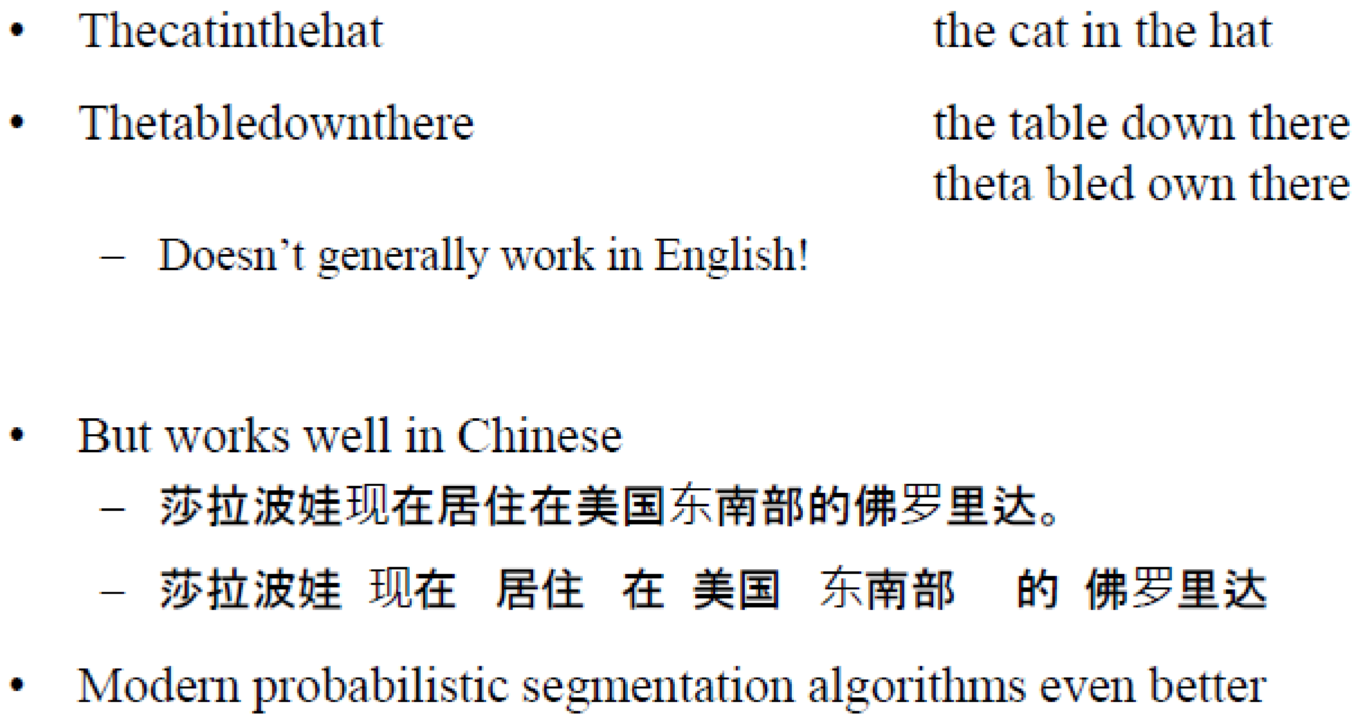
- Standard baseline segmentation algorithm: **Maximum Matching**

Given a wordlist of Chinese, and a string.

1. Start a pointer at the beginning of the string
2. Find the longest word in dictionary that matches the string starting at pointer
3. Move the pointer over the word in string
4. Go to 2

Max-match Segmentation algoritması İngilizce üzerinde çalışır mı?

Max-match segmentation



- Thecatinthehat the cat in the hat
- Thetabledownthere the table down there
  theta bled own there
  - Doesn't generally work in English!

- But works well in Chinese
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Modern probabilistic segmentation algorithms even better

Word Normalization and Stemming

Normalization

- Need to «normalize» terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match U.S.A. and USA and US

- We implicity define equivalence classes of terms
  - e.g., deleting periods in a term

Case folding

- Applications like IR: reduce all letters to lower case
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
- For sentiment analysis, MT, Information extraction
  - Case is helpful (***US*** versus ***us*** is important)

Lemmatization

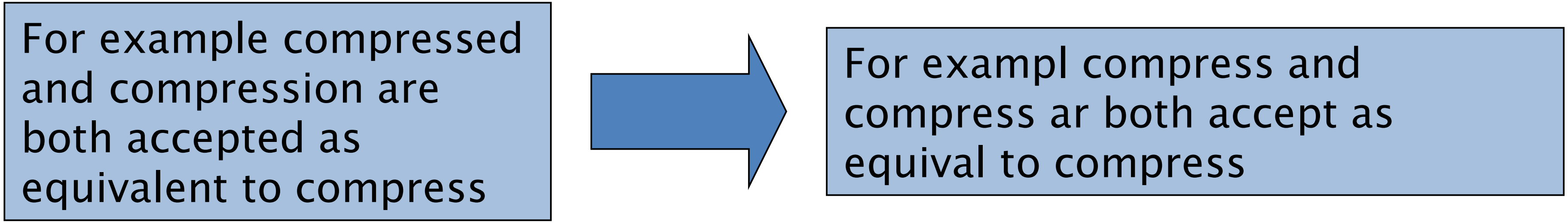
- Reduce inflections or variant forms to base form
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*
- **Lemmatization:** have to find correct dictionary headword form

Morphology

- Morphemes:
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Parts that adhere to stems, often with grammatical functions

Stemming

- Reduce terms to stems in information retieval
- *Stemming* is crude chopping off affixes
  - language dependent
  - e.g. **automate(s), automatic, automation** all reduced to **automat**

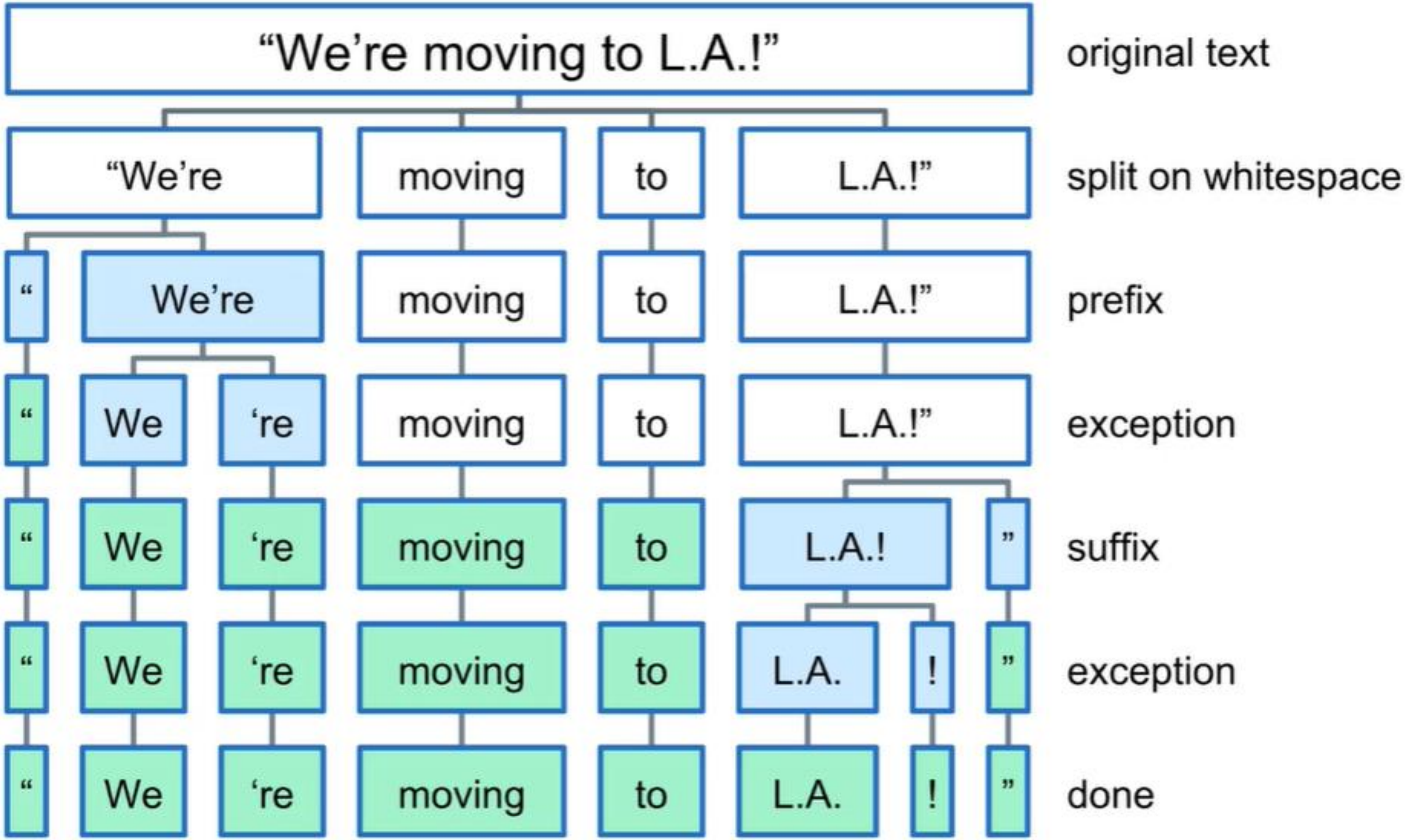


For example compressed and compression are both accepted as equivalent to compress

For exampl compress and compress ar both accept as equival to compress

ÖRNEKLER

Tokenization



“We’re moving to L.A.!” — original text

“We’re | moving | to | L.A.!” — split on whitespace

“ | We’re | moving | to | L.A.!” — prefix

“ | We | ‘re | moving | to | L.A.!” — exception

“ | We | ‘re | moving | to | L.A.! | ” — suffix

“ | We | ‘re | moving | to | L.A. | ! | ” — exception

“ | We | ‘re | moving | to | L.A. | ! | ” — done