Tokenization and Word Segmentation

Daniel Zeman, Rudolf Rosa

March 3, 2022
Tokenization and Word Segmentation

• IMPORTANT because:
  • Training tokenization ≠ test tokenization
  • ⇒ accuracy goes down

• Not always trivial
• May interact with morphology

• May include normalization (character-level)
Tokenization and Word Segmentation

• Issues of orthography of individual languages

• Issues caused by design decisions of individual corpora

• We will refer to the Universal Dependencies project (UD; https://universaldependencies.org/); more info in following weeks

• Due to limited time, we will probably skip some slides at the end
"María, I love you!" Juan exclaimed.

«¡María, te amo!», exclamó Juan.

X PRON X VERB X

« ¡María, te amo! », PUNCT PUNCT PROPN PUNCT PRON VERB PUNCT PUNCT

• Classic tokenization:
  • Separate punctuation from words
  • Recognize certain clusters of symbols like “...”
  • Perhaps keep together things like user@mail.x.edu
Using Unicode Character Categories

- https://perldoc.perl.org/perlunicode.html

```perl
$text =~ s/\pP/ $1 /g;
$text =~ s/^\s+//;
$text =~ s/\s+$//;
```

- Optionally recombine email addresses, URLs etc.
Using Unicode Character Categories

- https://perldoc.perl.org/perlunicode.html

```perl
$text =~ s/\pP/ $1 /g;
$text =~ s/^\s+//;
$text =~ s/\s+$//;
```

- Optionally recombine email addresses, URLs etc.

- Some problems
  - haven’t (English; should be have n’t)
  - instal·lació (Catalan; should be 1 token)
  - single quote (punctuation) misspelled as acute accent (modifier letter)

- writing systems without spaces
Normalization

- Often part of tokenization

- Decimal comma to decimal point; separator of thousands
Normalization

- Often part of tokenization

- Decimal comma to decimal point; separator of thousands

- Unicode directed quotes and long hyphens to undirected ASCII
  - Sometimes mistaken for ACUTE ACCENT, PRIME (math) etc.
Normalization

- Often part of tokenization

- Decimal comma to decimal point; separator of thousands

- Unicode directed quotes and long hyphens to undirected ASCII
  - Sometimes mistaken for ACUTE ACCENT, PRIME (math) etc.

- \TeX\-like ASCII directed quotes ` ` and ' ' and hyphens -- and ---
Normalization

- Often part of tokenization

- Decimal comma to decimal point; separator of thousands
- Unicode directed quotes and long hyphens to undirected ASCII
  - Sometimes mistaken for ACUTE ACCENT, PRIME (math) etc.
- \TeX\-like ASCII directed quotes ` ` and ' ' and hyphens -- and ---
- English/ASCII punctuation in foreign writing systems
  - 「你看過《三國演義》嗎？」他問我。
  - “你看過‘三國演義’嗎?”他問我.
Normalization

- Often part of tokenization

- Decimal comma to decimal point; separator of thousands

- Unicode directed quotes and long hyphens to undirected ASCII
  - Sometimes mistaken for ACUTE ACCENT, PRIME (math) etc.

- \TeX\-like ASCII directed quotes ` ` and ’ ’ and hyphens -- and ---

- English/ASCII punctuation in foreign writing systems
  - 「你看過《三國演義》嗎？」他問我。
  - “你看過‘三國演義’嗎?”他問我.

- European/ASCII digits in Arabic, Devanagari etc.
  - 0 1 2 3 4 5 6 7 8 9 (Western Arabic/European)
  - ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩ (Eastern Arabic)
  - ० १ २ ३ ४ ५ ६ ७ ८ ९ (Devanagari)
Let’s go to the sea.

Vámonos al mar. Vamos nos a el mar.

VERB? X NOUN PUNCT VERB PRON ADP DET NOUN PUNCT

• **Syntactic word** vs. orthographic word
• **Multi-word tokens**
• Two-level scheme:
  • Tokenization (low level, punctuation, concatenative)
  • Word segmentation (higher level, not necessarily concatenative)
• Orthographic vs. syntactic word
  • Syntactically autonomous part of orthographic word
  • Contractions \((al = a + el)\)
  • Clitics \((vámonos = vamos + nos)\)
    • ¿A qué hora \textit{nos vamos} mañana? 
      “What time do we leave tomorrow?”
    • \textit{Nos despertamos a las cinco}. 
      “We wake up at five.”
    • \textit{Nuestro guía nos despierta a las cinco}. 
      “Our guide wakes us up at five.”
He abdicated in favour of his son Baudouin.

yatanāzalu ʿan al-ʿarši li+ibni+hi būdūān
surrendered on the throne to son his Baudouin

VERB ADP NOUN ADP+NOUN+PRON PROPN
Segmentation as Part of Morphological Analysis

- Arabic
  - Select **Resolve**
  - Enter "ﻻﺑﻨﻪ" (labnh)

- Sanskrit
  - Sanskrit Reader Companion: [https://sanskrit.inria.fr/DICO/reader.fr.html](https://sanskrit.inria.fr/DICO/reader.fr.html)
  - Select Input convention = Devanagari
  - Enter “सकलार्थशास्त्रसारं जगति समालोक्य विष्णुश्रमेदम्” (sakalārthaśāstrasāraṁ jagati samālokya viṣṇuśarmedam)

- German compound splitting (unsupervised)
  - Not split in Universal Dependencies
We are now in Valencia.

現在我們在瓦倫西亞。

Xiàn zài wǒ men zài wǎ lún xī yǎ.

Now we in Valencia.

ADV  PRON  ADP  PROPN  PUNCT
I went to the beauty salon of Kyōdō [, Beyond-R.]
I went to the beauty salon of Kyōdō [, Beyond-R.]

経堂の美容室にきました

Kyōdō no miyōshitsu ni itte kimashita
I went to the beauty salon of Kyōdō [, Beyond-R.]

経堂の美容室に行ったきました

Kyōdōno miyōshitsuni itte kimashita

経堂美容室

Kyōdō to-beauty-salon going come

PROPN NOUN VERB

Case=Gen Case=Dat VerbForm=Conv VerbForm=Fin

Tense=Past Polite=Form
All the concrete country roads are the result of...

越南文：Emails are the result of...
Il touche environ 100 000 sesterces par an.
One syntactic word spans several orthographic words?

### text = Bin nach wie vor sehr zufrieden.

### text_en = I am still very satisfied.

1 Bin sein AUX ... 6 cop _ __
2 nach nach ADP ... 6 obl _ __
3 wie wie ADV ... 2 fixed _ __
4 vor vor ADP ... 2 fixed _ __
5 sehr sehr ADV ... 6 advmod _ __
6 zufrieden zufrieden ADJ ... 0 root _ SpaceAfter=No
7 . . PUNCT ... 6 obl _ __
Fixed Expressions

One syntactic word spans several orthographic words?

*I am still very satisfied.*
Some corpora use the underscore character to glue MWEs together.

*I am still very satisfied.*
Some corpora use the underscore character to glue MWEs together.

- Durante la presentación del libro "La_prosperidad_por_medio_de_la_investigación._La_investigación_básica_en_EEUU", editado por la Comunidad_de_Madrid, el secretario general de la Confederación_Empresarial_de_Madrid-CEOE (CEIM), Alejandro_Couceiro, abogó por la formación de los investigadores en temas de innovación tecnológica.

- Lemmas?
- Tags?
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.
  • Border case? Keep orthographic words (if they exist).
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.
  • Border case? Keep orthographic words (if they exist).

• Words with spaces
  • Vietnamese writing system
  • Very restricted set of exceptions (numbers)
  • Special relations elsewhere (fixed, compound)
Recoverability: CoNLL-U Format

```
# text = Vámonos al mar.
# text_en = Let’s go to the sea.

| ID | FORM  | LEMMA    | UPOS   | ... | HEAD | _ | MISC         |
|----|-------|----------|--------|-----|------|__|--------------|
| 1-2| Vámonos | _        | _      | ... | _    |   |              |
| 1  | Vamos  | ir       | VERB   | ... | 0    | root | __           |
| 2  | nos    | nosotros | PRON   | ... | 1    | obj  | __           |
| 3-4| al     | _        | _      | ... | _    |   |              |
| 3  | a      | a        | ADP    | ... | 5    | case | __           |
| 4  | el     | el       | DET    | ... | 5    | det  | __           |
| 5  | mar    | mar      | NOUN   | ... | 1    | obl  | __ SpaceAfter=No |
| 6  | .      | .        | PUNCT  | ... | 1    | punct | __          |
```

Tokenization and Word Segmentation
Recoverability: CoNLL-U Format

# text = Vámonos al mar.
# text_en = Let’s go to the sea.

| ID | FORM | LEMMA | UPOS | ... | HEAD | _ | MISC |
|----|------|-------|------|-----|------|__|______|
| 1-2 | Vámonos | _ | _ | ... | 0 | root | __ | __ |
| 1 | Vamos | ir | VERB | ... | 0 | root | __ | __ |
| 2 | nos | nosotros | PRON | ... | 1 | obj | __ | __ |
| 3-4 | al | _ | _ | ... | 1 | obj | __ | __ |
| 3 | a | a | ADP | ... | 5 | case | __ | __ |
| 4 | el | el | DET | ... | 5 | det | __ | __ |
| 5-6 | mar. | _ | _ | ... | 1 | obl | __ | __ |
| 5 | mar | mar | NOUN | ... | 1 | obl | __ | __ |
| 6 | . | . | PUNCT | ... | 1 | punct | __ | __ |
Tokenization vs. Multi-word Tokens

- Parallelism among closely related languages
  - ca: informar-se sobre el patrimoni cultural
  - es: informarse sobre el patrimonio cultural
  - en: learn about cultural heritage

- ca: L’único que veig és => L’únic que veig és
- en: don’t => do n’t

- No strict guidelines for tokenization (yet)
  - UD English: non-stop, post-war: single-word tokens
  - UD Czech: non-stop would be split to three tokens
  - Abbreviations: etc.
    - End of sentence...
Tokenization vs. Multi-word Tokens Summary

• Punctuation involved? Low level!
  • Exceptions: Spanish-Catalan parallelism.
Tokenization vs. Multi-word Tokens Summary

• Punctuation involved? Low level!
  • Exceptions: Spanish-Catalan parallelism.

• Boundary between two letters? Typically high level.
  • Exceptions: Chinese, Japanese.
• Punctuation involved? Low level!
  • Exceptions: Spanish-Catalan parallelism.

• Boundary between two letters? Typically high level.
  • Exceptions: Chinese, Japanese.

• Non-concatenative? High level!
Errors in Underlying Text

- We do not want to hide errors (learning robust parsers!)
  - But: reference corpora (linguistic research) may want to hide them.
Errors in Underlying Text

- We do not want to hide errors (learning robust parsers!)
  - But: reference corpora (linguistic research) may want to hide them.
- Possibilities:
- Typo not involving word boundary
  - FORM = *anotation*; LEMMA = *annotation*; FEATS: Typo=Yes; MISC: Correct=annotation
- Wrongly split word: *ann\_otation*
- Wrongly merged words: *the\_car*
- Fix tokenization (i.e. two lines); first line MISC: SpaceAfter=No | CorrectSpaceAfter=Yes
- Sentence segmentation can be affected, too!
• We do not want to hide errors (learning robust parsers!)
  • But: reference corpora (linguistic research) may want to hide them.

• Possibilities:

• Typo not involving word boundary
  • FORM = anotation; LEMMA = annotation; FEATS: Typo=Yes; MISC: Correct=annotation

• Wrongly split word:
Errors in Underlying Text

• We do not want to hide errors (learning robust parsers!)
  • But: reference corpora (linguistic research) may want to hide them.

• Possibilities:

  • Typo not involving word boundary
    • FORM = annotation; LEMMA = annotation; FEATS: Typo=Yes; MISC: Correct=annotation

  • Wrongly split word:

  • Wrongly merged words: thecar
    • Fix tokenization (i.e. two lines); first line MISC: SpaceAfter=No | CorrectSpaceAfter=Yes
    • Sentence segmentation can be affected, too!
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo (*the car iss produced...*)
Errors in Underlying Text

• Wrong morphology: the cars is produced in Detroit
  • Not like normal typo (the car iss produced...)
  • Not obvious what is correct
    • the car is
    • the cars are
Errors in Underlying Text

• Wrong morphology: *the cars is produced in Detroit*
  • Not like normal typo (*the car iss produced...*)
  • Not obvious what is correct
    • *the car is*
    • *the cars are*

• Suggestion: select which word to fix, e.g. *cars to car*

• FORM = *cars*; FEATS: Number=Plur; MISC: Correct=car | CorrectNumber=Sing
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo (*the car iss produced...*)
  - Not obvious what is correct
    - *the car is*
    - *the cars are*

- Suggestion: select which word to fix, e.g. *cars to car*
- FORM = *cars*; FEATS: Number=Plur; MISC: Correct=car | CorrectNumber=Sing
- cs: *viděl moří* “he saw the sea”
  - Should be *moře*
  - Would be Case=Acc (disambiguated from Case=Acc,Gen,Nom,Voc)
  - This form is Case=Dat,Loc (but which one?)
- *cestoval k moři* “he traveled to the sea” Case=Dat
- *plavil se po moři* “he sailed the sea” Case=Loc
Tokenization Alignment

- If you need to match two different tokenizations
- Use case: evaluation of end-to-end parsing systems
If you need to match two different tokenizations
Use case: evaluation of end-to-end parsing systems

Normalization involved? Bad luck…
- Normalization rules needed
- Or: Longest common subsequence (LCS) algorithm
Tokenization Alignment

- If you need to match two different tokenizations
- Use case: evaluation of end-to-end parsing systems

- Normalization involved? Bad luck...
  - Normalization rules needed
  - Or: Longest common subsequence (LCS) algorithm

- Otherwise easy
  - Non-whitespace character offsets
Evaluation Metrics

• Align system-output tokens to gold tokens

*Al-Zaman* : *American forces killed Shaikh Abdullah al-Ani, the preacher at the mosque in the town of Qaim, near the Syrian border.*

**GOLD:**  
Al - Zaman : American forces killed Shaikh  
OFFSET:  0-1  2  3-7  8  9-16  17-22  23-28  29-34

• All characters except for whitespace match => easy align!

**SYSTEM:**  
Al-Zaman : American forces killed Shaikh  
OFFSET:  0-7  8  9-16  17-22  23-28  29-34
Evaluation Metrics

- Align system-output tokens to gold tokens

_Die Kosten sind definitiv auch im Rahmen._

**GOLD:** Die Kosten sind definitiv auch _im_ Rahmen .

**SPLIT:** Die Kosten sind definitiv auch _in dem_ Rahmen .

**OFFSET:** 0-2 3-8 9-12 13-21 22-25 26-27 28-33 34

- Corresponding but not identical spans?

- Find longest common subsequence

**SYSTEM:** Kosten sind definitiv auch _im_ Rahmen .

**SPLIT:** Kosten sind _de finitiv_ auch _im_ Rahmen .

**OFFSET:** 3-8 9-12 13-21 22-25 26-27 28-33 34
Evaluation Metrics

• Align system-output tokens to gold tokens

Die Kosten sind definitiv auch im Rahmen.

GOLD:  Die Kosten sind definitiv auch im Rahmen .
SPLIT:  Die Kosten sind definitiv auch in dem Rahmen .
OFFSET: 0-2  3-8  9-12  13-21  22-25  26-27  28-33  34

• Corresponding but not identical spans?
• Find longest common subsequence

SYSTEM:   auch im Rahmen .
SPLIT:    auch in einem, dem alle zustimmen, Rahmen .
OFFSET:  22-25  26-27  28-33  34