Unsupervised Morphological Segmentation Based on Word Segments Predictability and Alignment

Delphine Bernhard

Laboratoire TIMC-IMAG, Grenoble, France

Unsupervised Segmentation of Words into Morphemes
Pascal Challenge Workshop, April 12, 2006

Part I

Motivation

Why ?

Context

Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic

Why ?

Context

Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic
- slammograms

Why ?

Context

Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic
- slammograms (refers to mammograms)

Why ?

Context

Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic
- slammograms (refers to mammograms)
- pneumonoultramicroscopicsilicovolcanoconiosis

Why ?

Context

Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic
- slammograms (refers to mammograms)
- pneumonoultramicroscopicsilicovolcanoconiosis
  "a lung disease caused by the inhalation of very fine silica dust, mostly found in volcanoes" = pneumoconiosis

Why ?

Context

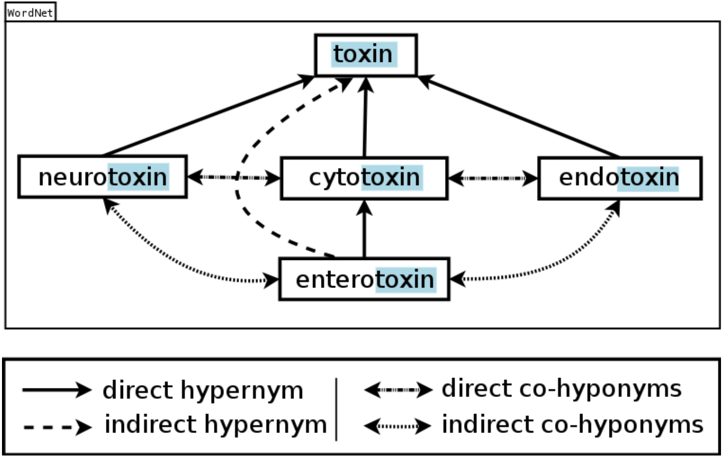
Work on the morphology of domain-specific vocabulary, esp. medical language (many neoclassical compounds)

Examples

- dermatofibrosarcoma
- glomeroporphyritic
- slammograms (refers to mammograms)
- pneumonoultramicroscopicsilicovolcanoconiosis
  "a lung disease caused by the inhalation of very fine silica dust, mostly found in volcanoes" = pneumoconiosis
  (But this is a hoax !)

Objectives

- Automatic acquisition of semantic relationships thanks to morphological relatedness



WordNet

toxin

neurotoxin

cytotoxin

endotoxin

enterotoxin

→ direct hypernym

- - -> indirect hypernym

↔ direct co-hyponyms

↔ indirect co-hyponyms

Part II

Method

Constraints

- Take into account all of the following word formation processes:
  - inflection
  - derivation
  - compounding
- Method not limited to French or English.
- Distinguish between different types of word segments:
  - prefix
  - suffix
  - stem
  - linking element

Overview of the method

Input

List of words

Stages

1. Acquisition of prefixes and suffixes
2. Acquisition of stems
3. Alignment of word segments
4. Selection of the best segmentation for each word

Acquisition of prefixes and suffixes [1]

Input

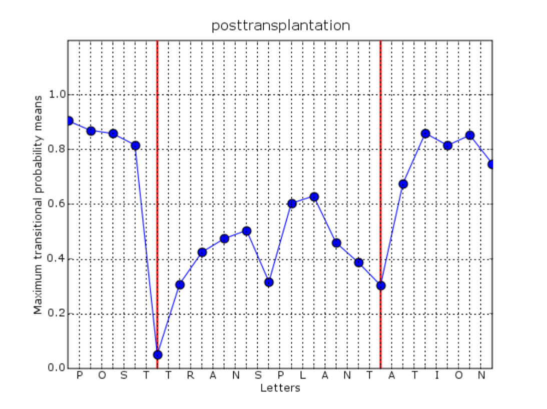
Longest words

Acquisition of prefixes and suffixes [1]

Input

Longest words

Locate positions with low segment predictability

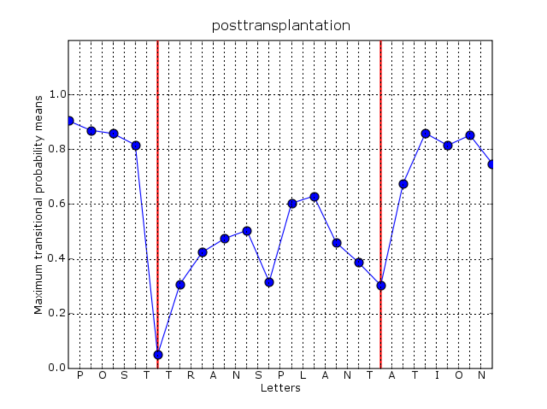


Acquisition of prefixes and suffixes [1]

Input

Longest words

Locate positions with low segment predictability



Output

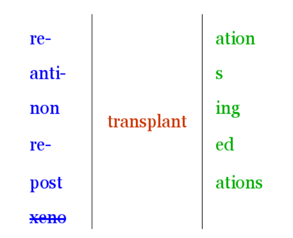
Segments

Acquisition of prefixes and suffixes [2]

Identification of a stem among the segments

Segments	post	transplant	ation
Frequency	278	> 42 <	1,163
Length	4	< 10 >	5

Prefixes and suffixes



re-
anti-
non
re-
post
~~xeno~~

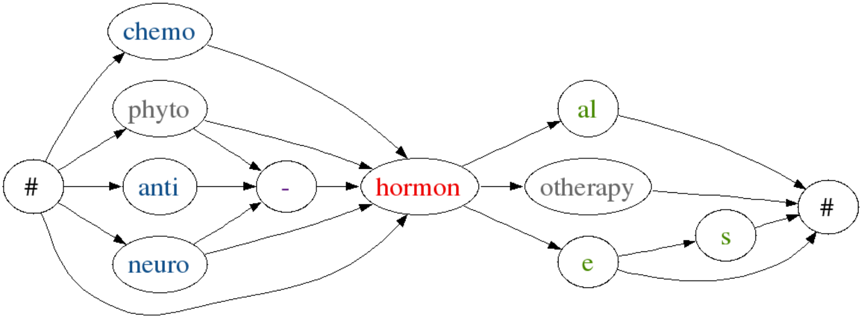
transplant

ation
s
ing
ed
ations

Extraction of stems

Subtract prefixes and suffixes from all words

Alignment of word segments [1]



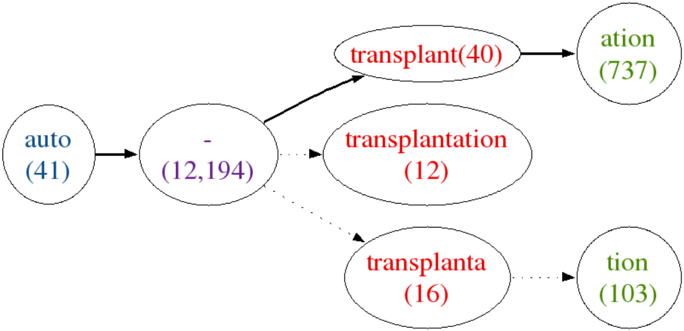
Alignment of word segments [2]

Validation of new prefixes and suffixes

Words	Known suffixes A_1	Potential stems A_2	New suffixes A_3
hormonal	-al		
hormonotherapy		-otherapy	
hormone	-e		
hormones			-es

$$\frac{|A_1| + |A_2|}{|A_1| + |A_2| + |A_3|} \geq a \text{ and } \frac{|A_1|}{|A_1| + |A_2|} \geq b$$

Selection of the best segmentation



Segmentation of new words

- For each word, select segments so that the total cost is minimal
- Cost functions used:

$$cost_1(s_i) = -log \frac{f(s_i)}{\sum_i f(s_i)}$$

$$cost_2(s_i) = -log \frac{f(s_i)}{\mathrm{max_i}[f(s_i)]}$$

Part III

Results and conclusion

Evaluation

Position of boundaries

MorphoChallenge evaluation

Conflation sets

Check if word forms containing the same stem are related

- Test on an English corpus on breast cancer (about 86,000 word types).
- Manually built morphological families for the top 5,000 key words
- Results: F-measure $\sim$ 50% (Recall = 40% $\pm$ 7, Precision = 66% $\pm$ 7)

Examples [1]

Words	Segmentations
chondrosarcomas	chondro + sarcoma + s
cystosarcoma	cyst + o + sarcoma
dermatofibrosarcomas	derm + at + o + fibro + sarcoma + s
fibroxanthosarcoma	fibroxanthosarcoma
leiomyosarcoma	leiomyo + s + arc + oma
leiomyosarcomas	leiomyo + sarcoma + s
liposarcoma	lipo + sarcoma
lymphangiosarcomas	lymph + angiosarcoma + s
myxofibrosarcoma	myxo + fibro + sarcoma
myxosarcomas	myxo + sarcoma + s
neurofibrosarcoma	neur + o + fibro + sarcoma
osteosarcoma	osteo + sarcoma
osteosarcomatous	osteosarcoma + tous
sarcoma	sarcoma
sarcomatoid	sarcoma + t + oid

Examples [2]

Words	Segmentations
auto-transplant	auto + - + transplant
auto-transplantation	auto + - + transplant + ation
autotransplantation	auto + transplant + ation
post-transplantation	post + - + transplant + ation
posttransplantation	post + transplant + ation
retransplantation	re + transplant + ation
transplantability	transplantability
transplant	trans + plant
transplanted	trans + plant + e + d
transplanting	trans + plant + ing
transplants	trans + plant + s
xenotransplantation	xenotransplant + ation
xenotransplanted	xenotransplant + ed
xenotransplants	xeno + transplants

Main issues

Over-segmentation

- leiomyo + s + arc + oma
- g + lobul + e

⇒ Low precision

Under-segmentation

- transplantability
- xenotransplant + ation
- xenotransplant + ed

⇒ Low recall

Conclusion

Summary

- Method usable for languages other than French and English
- Performs segmentation + distinguishes between different kinds of segments

Future work

- Use other data structures to deal with very, very large corpora
- Deal with variations within stems (accents, alternations)
- Evaluate how well word segments predict semantic relationships between terms

Thank you

Further information:

Delphine.Bernhard@imag.fr
http://www-timc.imag.fr/Delphine.Bernhard/