Unsupervised Morphological Analysis Using Tries

Koray Ak, Olcay Taner Yıldız

Department of Computer Engineering
Işık University
Istanbul, Turkey

ISCIS 2011

Outline

Languages

Two kinds of languages:

- Polysynthetic $\rightarrow$ connecting morphemes
  - Fusional languages (English, French, etc.)
    sing $\rightarrow$ sang
  - Agglutinative languages (Turkish, etc.)
    marmaradaki $\rightarrow$ marmara+da+ki
- Analytic (isolating) $\rightarrow$ stand alone morphemes (Mandarin, Chinese)

Types of Morphology

- Morpheme-based morphology
  - Morpheme $\rightarrow$ relationship between word forms.

$$word = morpheme + ... + morpheme$$

- Lexeme-based morphology

$$word_2 = word_1 + rule$$

- Word-based morphology

  Word-and-paradigm approach

  Inflectional paradigms to determine the word

Morphological Analysis

- Studies the structure of the words
- Machine Translation (MT), Information Retrieval (IE)
- Solution
  - Unsupervised algorithms that use machine learning approaches.
- Benefits
  - Solves problem in generalized manner.
  - Low Cost

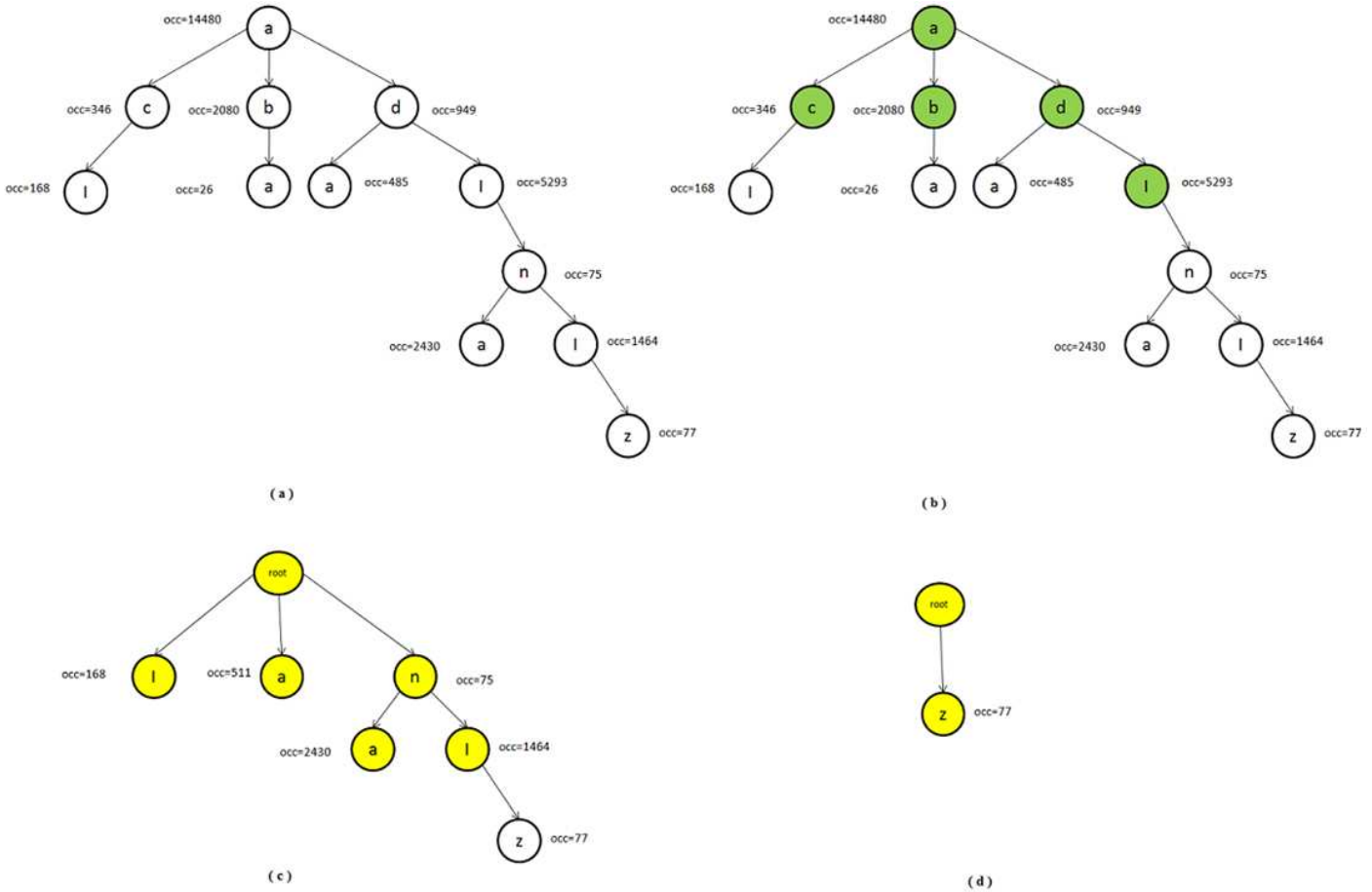
Morpho Challenge

- Competition, part of the EU Network of Excellence PASCAL2 Challenge Program.
- Started in 2005, arranged in each year except 2006.
- Objective: Design a statistical machine learning algorithm that discovers which morphemes words consist of.

Competitors

- Distinguish of stems and affixes by examining the differences in lengths and frequencies. [Bernhard06]
- Find substring words and transitional probabilities. [Keshava06]
- Paradigm-based approach [Zeman08]
  - Paradigm $\rightarrow$ all possible suffix-stem pairs.
  - Segmentation phase
- Character-based segmentation (ParaMor) [Monson,et al.09]
- Improved version of ParaMor [Monson, et al.oct09]

REC-TRIE


occ=14480
a
occ=346
c
occ=2080
b
d
occ=949
occ=168
l
occ=26
a
a
occ=485
l
occ=5293
n
occ=75
occ=2430
a
l
occ=1464
z
occ=77
(a)
occ=14480
a
occ=346
c
occ=2080
b
d
occ=949
occ=168
l
occ=26
a
a
occ=485
l
occ=5293
n
occ=75
occ=2430
a
l
occ=1464
z
occ=77
(b)
root
occ=168
l
occ=511
a
n
occ=75
occ=2430
a
l
occ=1464
z
occ=77
(c)
root
z
occ=77
(d)

Dataset

Dataset

- Word list (gathered from different sources) with word frequencies.
- Most of the root words appear in the word list (66%).
- Turkish $\rightarrow$ 15545 root words among 617298 words.

Experiments

- Two Perl scripts are provided by the challenge to calculate these values.
- Comparison with linguistic gold standard
- Evaluation based on F-measure
- Metrics:
  - Hit: A valid cut that at the right place.
  - Insertion: A wrong cut at the wrong place.
  - Deletion: A missed cut where a valid cut is ignored.

$$Precision = H/(H + I)$$
$$Recall = H/(H + D)$$
$$F - Measure = 2H/(2H + I + D)$$

Results-Turkish

Precision, Recall, and F-Measure of REC-TRIE compared with other algorithms in Morpho Challenge 2009 for Turkish.

Author	Method	Precision	Recall	F-Measure
Monson et al.	Paramor-Morfessor Mimic	48.07%	60.39%	53.53%
Monson et al.	ParaMor-Morfessor Union	47.25%	60.01%	52.88%
Monson et al.	Paramor Mimic	49.54%	54.77%	52.02%
Our Algorithm	**REC-TRIE**	**53.40%**	**43.06%**	**47.68%**
Lavallée &Langlais	RALI-COF	48.43%	44.54%	46.40%
-	Morfessor CatMAP	79.38%	31.88%	45.49%
Spiegler et al.	PROMODES 2	35.36%	58.70%	44.14%
Spiegler et al.	PROMODES	32.22%	66.42%	43.39%
Bernhard	MorphoNet	61.75%	30.90%	41.19
Can & Manandhar	2	41.39%	38.13%	39.70%
Spiegler et al.	PROMODES committee	55.30%	28.35%	37.48%
Golénia et al.	UNGRADE	46.67%	30.16%	36.64%
Virpioja &Kohonen	Allomorfessor	85.89%	19.53%	31.82%
-	Morfessor Baseline	89.68%	17.78%	29.67%
Lavallée &Langlais	RALI-ANA	69.52%	12.85%	21.69%
-	Letters	8.66%	99.13%	15.93%
Can & Manandhar	1	73.03%	8.89%	15.86%

Results-English

Precision, Recall, and F-Measure of REC-TRIE compared with other algorithms in Morpho Challenge 2009 for English.

Author	Method	Precision	Recall	F-Measure
Virpioja&Kohonen	Allomorfessor	68.98%	56.82%	62.31%
-	Morfessor Baseline	74.93%	49.81%	59.84%
Monson, et al.	Paramor-Morfessor Union	55.68%	62.33%	58.82%
Lingos, et al.	-	83.49%	45.00%	58.48%
Monson, et al	ParaMor-Morfessor Mimic	54.80%	60.17%	57.36%
Monson, et al	Paramor Mimic	53.13%	59.01%	55.91%
Bernhard	MorphoNet	65.08%	47.82%	55.13
Monson, et al	Paramor Mimic	53.13%	59.01%	55.91%
Monson, et al	Paramor Mimic	53.13%	59.01%	55.91%
Bernhard	MorphoNet	65.08%	47.82%	55.13
Lavallée &Langlais	RALI-COF	68.32%	46.45%	55.30%
Our Algorithm	**REC-TRIE**	**50.80%**	**53.86%**	**52.29%**
Can & Manandhar	-	58.52%	44.82%	50.76%
-	Morfessor CatMAP	84.75%	35.97%	50.50%
Spiegler, et al.	PROMODES	36.20%	64.81%	46.46%
Lavallée &Langlais	RALI-ANA	64.61%	33.48%	44.10%
Spiegler,et al.	PROMODES 2	32.24%	61.10%	42.21%
Spiegler et al.	PROMODES committee	32.24%	61.10%	42.21%
Tchoukalov, et al.	MetaMorph	68.41%	27.55%	39.29%
Golénia, et al.	UNGRADE	28.29%	51.74%	36.58%
-	Letters	3.82%	99.88%	7.35%

Results-Finnish

Precision, Recall, and F-Measure of REC-TRIE compared with other algorithms in Morpho Challenge 2009 for Finnish.

Author	Method	Precision	Recall	F-Measure
Monson, et al.	Paramor-Morfessor Union	47.89%	50.98%	49.39%
Monson, et al	ParaMor-Morfessor Mimic	51.75%	45.42%	48.38%
-	Morfessor CatMAP	79.01%	31.08%	44.61%
Spiegler et al.	PROMODES committee	41.20%	48.22%	44.44%
Monson, et al	Paramor Mimic	47.15%	40.50%	43.57%
Spiegler,et al.	PROMODES 2	33.51%	61.32%	43.34%
Spiegler, et al.	PROMODES	33.86%	51.41%	42.25%
Lavallée &Langlais	RALI-COF	74.76%	26.20%	38.81%
Golénia, et al.	UNGRADE	40.78%	33.02%	36.49%
Our Algorithm	**REC-TRIE**	**45.09%**	**27.05%**	**33.81%**
Bernhard	MorphoNet	63.35%	22.62%	33.34
Virpioja&Kohonen	Allomorfessor	86.51%	19.96%	32.44%
-	Morfessor Baseline	89.41%	15.73%	26.75%
Tchoukalov, et al.	MetaMorph	37.17%	15.15%	21.53%
Lavallée &Langlais	RALI-ANA	60.06%	10.33%	17.63%
-	Letters	5.17%	99.89%	9.83%

Conclusion

- No prefix detection method, no control for the irregular changes of the words or umlauts
- Encouraging results with respect to the competitors' performance $\rightarrow$ 4th in Turkish, 12th in English, 10th in Finnish.