

Knowledge-Free Induction of Morphology Using Latent Semantic Analysis

Patrick Schone and David Jurafsky

Presentation by Torsten Marek

May 13th, 2008

Outline

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 Introduction
- 2 Morphology Induction
- 3 Results

Rule-Based Morphological Analysis with FSTs

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Foundations

Kaplan & Kay 1981: Regular Models of Phonological Rule Systems

Rule-Based Morphological Analysis with FSTs

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Foundations

Kaplan & Kay 1981: Regular Models of Phonological Rule Systems

Rule-Based Finite State Morphology with XFST

```
define Syllabify C* V+ C* @-> ... "." || _ C V ;

define TernaryFeet BF "." Light @-> "(" ... ")"
// [{}].} | .#. ] [BF "."]* _
// [". Heavy "." S ] | .#. ;

define BinaryFeet BF @-> "(" ... ")" || .#.|"." _ .#.|"." ;
```

Rule-Based Analysis II

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Advantages

- scales to very complex morphological systems
- can be used for generation as well

Rule-Based Analysis II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Advantages

- scales to very complex morphological systems
- can be used for generation as well

Disadvantages

- time-consuming to write
- needs rule system experts that also know the language to be analyzed

Knowledge-Free Morphology Induction

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Definition

*... the task of inducing machine-readable dictionaries
using **no** human-provided information...*

Knowledge-Free Morphology Induction

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Definition

*... the task of inducing machine-readable dictionaries
using **no** human-provided information...*

Really knowledge-free?

① simplistic tokenization^a

^ano further information in the paper

Knowledge-Free Morphology Induction

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors

Results

Definition

*... the task of inducing machine-readable dictionaries using **no** human-provided information...*

Really knowledge-free?

- 1 simplistic tokenization^a
- 2 no stopword removal

^ano further information in the paper

Knowledge-Free Morphology Induction

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Definition

*... the task of inducing machine-readable dictionaries
using **no** human-provided information...*

Really knowledge-free?

- 1 simplistic tokenization^a
- 2 no stopword removal
- 3 no capitalization normalization

^ano further information in the paper

Knowledge-Freeness II: Assumptions

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Morphology Type

Cannot generalize roots in template morphology.

Example: Arabic

drs: /**dars**/ (lesson) - /**madr**asa/ (school) - /**dir**asa/ (the study) - /**mud**arris/ (professor)

Knowledge-Freeness II: Assumptions

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Morphology Type

Cannot generalize roots in template morphology.

Example: Arabic

drs: /**dars**/ (lesson) - /**madrasa**/ (school) - /**dirasa**/ (the study) - /**mudarris**/ (professor)

Affix Regularity

Cannot capture varying affixes.

Example: Turkish vowel harmony

dir (it is): kapı**dır** - palt**odur**

Knowledge-Freeness II: Assumptions

Morphology
Induction

Morphology Type

Cannot generalize roots in template morphology.

Example: Arabic

drs: /**dars**/ (lesson) - /**madr**asa/ (school) - /**dir**asa/ (the study) - /**mudarris**/ (professor)

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Affix Regularity

Cannot capture varying affixes.

Example: Turkish vowel harmony

dir (it is): kapı**dir** - palt**odur**

Whitespace Separation

Cannot work when tokens are not clearly separable from each other.

Example: Chinese

Knowledge-Freeness III: Problems

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Knowledge-Freeness III: Problems

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 valid affixes may be applied to the wrong words
 - *ally* \leftrightarrow *all*

Knowledge-Freeness III: Problems

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 valid affixes may be applied to the wrong words
 - *ally* \rightarrow *all*
- 2 words may be ambiguous
 - *rating* \rightarrow *rat* (correct: *rate*)

Knowledge-Freeness III: Problems

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 valid affixes may be applied to the wrong words
 - *ally* \rightarrow *all*
- 2 words may be ambiguous
 - *rating* \rightarrow *rat* (correct: *rate*)
- 3 non-productive affixes may be pruned
 - *dirty* \rightarrow *dirt* not captured

Processing Stages

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Processing Stages

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Stage 1

Identify potential affixes

Processing Stages

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Stage 1

Identify potential affixes

Stage 2

Find word pairs that might be morphologically related

Processing Stages

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Stage 1

Identify potential affixes

Stage 2

Find word pairs that might be morphologically related

Stage 3

Create semantic vectors for words

Processing Stages

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Stage 1

Identify potential affixes

Stage 2

Find word pairs that might be morphologically related

Stage 3

Create semantic vectors for words

Stage 4

Select variants that have similar semantic vectors

Acquisition

Morphology Induction

Introduction

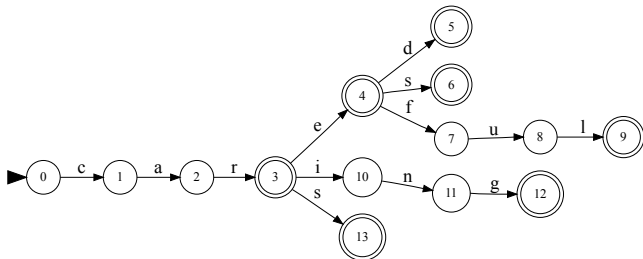
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors



Results

Identifying Affixes

- 1 Create a trie from words

Acquisition

Morphology Induction

Introduction

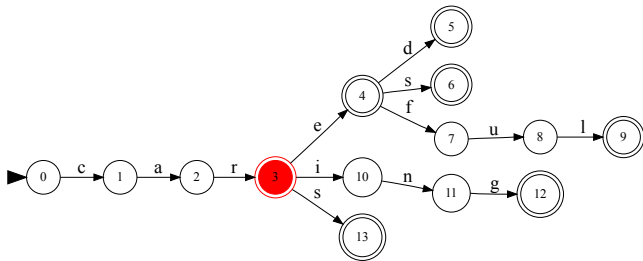
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors



Results

Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

Morphology Induction

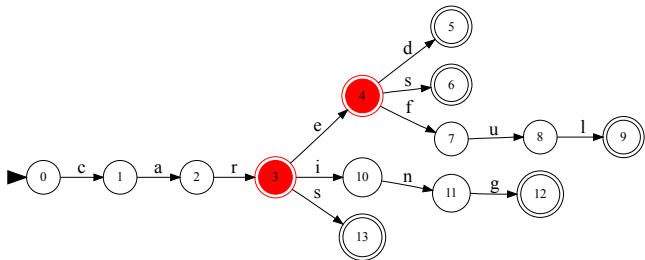
Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors

Results



Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

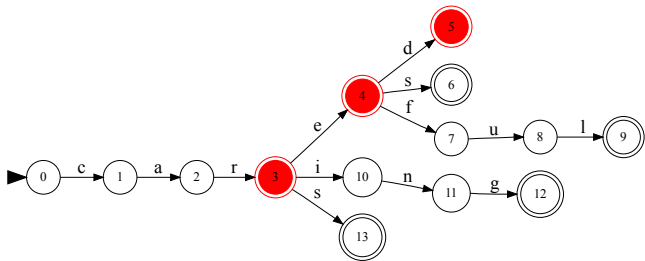
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors



Results

Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

Morphology Induction

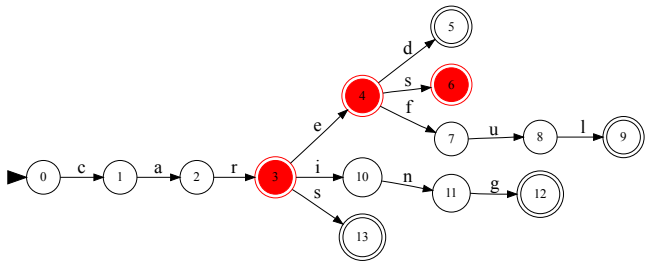
Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors

Results



Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

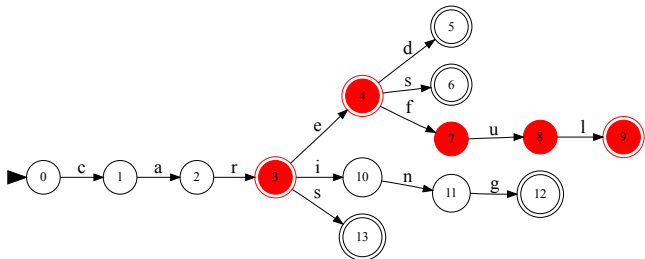
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors



Results

Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

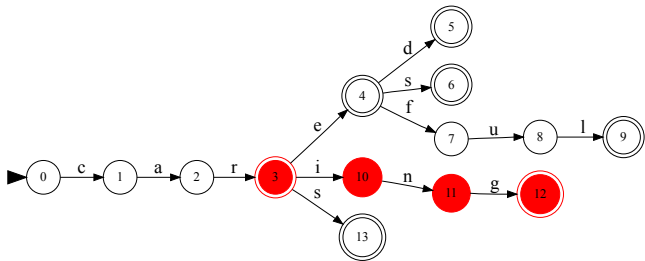
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors



Results

Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Acquisition

Morphology Induction

Introduction

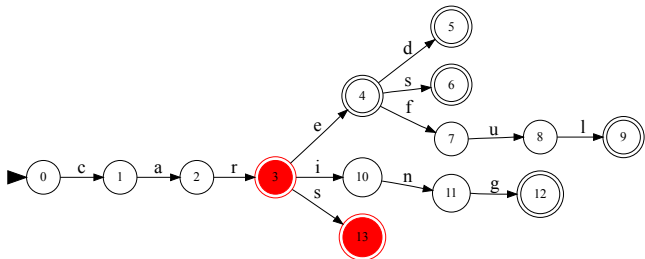
Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors



Results

Identifying Affixes

- 1 Create a trie from words
- 2 Collect suffixes from suffix languages:
NULL, "e", "ed", "es", "eful", "ing", "s"

Suffix Selection

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Create suffix statistics

- 1 Enumerate all suffix languages
- 2 Count suffix frequencies
- 3 Take only K most frequent suffixes

Suffix Selection

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Create suffix statistics

- 1 Enumerate all suffix languages
- 2 Count suffix frequencies
- 3 Take only K most frequent suffixes

Guidelines for K

- must include all regular affixes
- should include more frequent irregular affixes
- (Schone and Jurafsky, 2000): $K = 200$

Suffix Statistics

Morphology
Induction

Rank	Suffix	Sfx#	Sfx%	Type%
1	NULL	56057	11.16	100.00
2	s	12519	2.49	22.33
3	e	6529	1.30	11.65
4	d	6159	1.23	10.99
5	ed	4817	0.96	8.59
6	y	4569	0.91	8.15
7	n	4123	0.82	7.36
8	g	4011	0.80	7.16
9	ng	3831	0.76	6.83
10	ing	3726	0.74	6.65

Table: Suffixes in the Brown Corpus

(Schone and Jurafsky, 2000) do not give any statistics about the suffixes in their paper.

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Rule Format

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

**Stage 2: Rule
Hypotheses**

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Definition

Pairs of candidate affixes from Stage 1 that descend from the same ancestor.

Rule Format

Definition

Pairs of candidate affixes from Stage 1 that descend from the same ancestor.

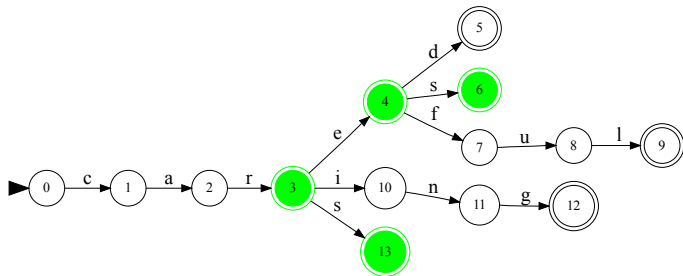


Figure: Sample Rule: ("s", NULL)

Rule Notions

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

PPMVs

pair of potential morphological variants, two words that share the same root and the same affix rule r .

Example: (*car*, *cars*)

Rule Notions

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

PPMVs

pair of potential morphological variants, two words that share the same root and the same affix rule r .

Example: (*car*, *cars*)

Rulesets

The set of all PPMVs of an affix rule r .

Example: $\{(car, cars), (care, cares)\}$

Rule Variations

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

**Stage 2: Rule
Hypotheses**

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Subrules

Specific rules may apply under certain conditions only.

Rule Variations

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Subrules

Specific rules may apply under certain conditions only.

Example: ("es", NULL)

- cares \rightarrow car
- flashes \rightarrow flash

Rule Variations

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Subrules

Specific rules may apply under certain conditions only.

Example: (“es”, NULL)

- cares \nrightarrow car
- flashes \rightarrow flash

Subrules

Consider potential subrules like (“shes”, “sh”)

Latent Semantic Analysis

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

**Stage 3:
Semantic
Vectors**

Stage 4: Find
similar semantic
vectors

Results

Recap

... find significant semantic relations between words and documents in a corpus with virtually no human intervention ...

The Term \times Term Matrix

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Matrix Layout

Size: $N \times 2N$

$$M(i, pN + j)$$

N $N - 1$ most frequent words from the corpus
 $i = j = N$ contains all other words

i, j word indices the vector f , words sorted by
descending frequency

p 0 for words j that occur within 50 tokens before
 i , 1 for 50 tokens after.

Matrix Example

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

**Stage 3:
Semantic
Vectors**

Stage 4: Find
similar semantic
vectors

Results

Example Sentence: *The dog chased the cat with an airplane.*

Matrix Example

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Example Sentence: *The dog chased the cat with an airplane.*

Frequency Vector

$f = [the, an, with, dog, cat, chased, airplane]$

Matrix Example

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Example Sentence: *The dog chased the cat with an airplane.*

Frequency Vector

$f = [the, an, with, dog, cat, chased, airplane]$

$i = 0$ (the), $j = 3$ (dog)

$M(0, 0N + 3) = 0$

$M(0, 1N + 3) = 1$

Related Vectors

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vector Similarity

Angle between two vectors v_1, v_2 :

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Related Vectors

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vector Similarity

Angle between two vectors v_1, v_2 :

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Correlation

Compute μ_w, σ_w by comparing Ω_w to 200 randomly chosen semantic vectors.

Normalized Cosine Score

Semantic Vectors

$$\Omega_w = \xi_w^T V_k$$

ξ_w^T : word row from the matrix M

V_k : projection matrix into the k-dimensional latent semantic space

Ω_w : semantic vector

$k = 300$ (*usual value*)

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Normalized Cosine Score

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Semantic Vectors

$$\Omega_w = \xi_w^T V_k$$

ξ_w^T : word row from the matrix M

V_k : projection matrix into the k-dimensional latent semantic space

Ω_w : semantic vector

$k = 300$ (*usual value*)

$NCS(\Omega_w, \Omega_v)$

$$NCS(\Omega_w, \Omega_v) = \min_{y \in \{v, w\}} \frac{\cos(\Omega_w, \Omega_v) - \mu_y}{\sigma_y}$$

Sample NCSs

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding Suffixes

Stage 2: Rule Hypotheses

Stage 3: Semantic Vectors

Stage 4: Find similar semantic vectors

Results

PPMVs	NCS	PPMVs	NCS
car/cars	5.6	ally/allies	6.5
car/caring	-0.71	ally/all	-1.3
car/cares	-0.14	dirty/dirt	2.4
car/cared	-0.96	rating/rate	0.97

Table: Normalized Cosines for various PPMVs

(Schone and Jurafsky, 2000, p. 4)

Ruleset-Level Statistics

Morphology
Induction

Assumption

Random NCSs are distributed according to $\mathcal{N}(0, 1)$.

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Ruleset-Level Statistics

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Assumption

Random NCSs are distributed according to $\mathcal{N}(0, 1)$.

Rule Legitimacy

$$Pr(true) = \frac{n_T \Phi_Z(\mu_T, \sigma_T)}{(n_R - n_T) \Phi_Z(0, 1) + n_T \Phi_Z(\mu_T, \sigma_T)}$$

n_R number of PPMVs in a particular ruleset

n_T number of PPMVs in a true correlation

μ_T mean NCS of a true correlation

σ_T standard deviation of a true correlation

$$\Phi_Z(\mu, \sigma) = \int_Z^{\infty} e^{-\left(\frac{X-\sigma}{\sigma}\right)^2}$$

Subrule Selection

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Recap: (“es”, NULL)

- cares \rightarrow car
- flashes \rightarrow flash

Solution: introduce subrules like (“ches”, “ch”)

Subrule Selection

Recap: (“es”, NULL)

- cares \rightarrow car
- flashes \rightarrow flash

Solution: introduce subrules like (“ches”, “ch”)

Rule/Subrule	Average	StdDev	#instances
(“es”, NULL)	1.62	2.43	173
(“ches”, “ch”)	2.20	1.66	32
(“shes”, “sh”)	2.39	1.52	15
(“res”, “r”)	-0.69	0.47	6
(“tes”, “t”)	-0.58	0.93	11

Table: Subrules Analysis

Experiments

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Data

8 mio. words from the TREC data, only words of frequency
10+ considered.

Experiments

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Data

8 mio. words from the TREC data, only words of frequency
10+ considered.

Morphological Gold Standard: CELEX

- 1 MRD for English
- 2 hand-tagged
- 3 morphological, phonetical and syntactical information

Reference System: Linguistica

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Linguistica

- unsupervised learning of morphology
- segments words into (several) morphemes

Reference System: Linguistica

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Linguistica

- unsupervised learning of morphology
- segments words into (several) morphemes

Minimum Description Length

MDL tries to find a model that optimally compresses morphological information of a word, using information-theoretical notions.

Conflation Sets

Explanation

Morphological relations modeled as directed graphs

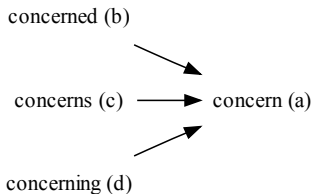


Figure: Conflation set $\{a, b, c, d\}$

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Evaluation

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Process

- 1 Extract gold standard conflation sets Y_w for each word w from CELEX

Evaluation

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Process

- 1 Extract gold standard conflation sets Y_w for each word w from CELEX
- 2 Induce conflation sets X_w from text data

Evaluation

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Process

- 1 Extract gold standard conflation sets Y_w for each word w from CELEX
- 2 Induce conflation sets X_w from text data
- 3 Compare X_w and Y_w

Evaluation II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vertex Comparison

Correct:
$$C = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|}$$

Evaluation II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vertex Comparison

$$\text{Correct: } \mathcal{C} = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|}$$

$$\text{Deletions: } \mathcal{D} = \sum_w \frac{|Y_w - X_w|}{|Y_w|}$$

Evaluation II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vertex Comparison

$$\text{Correct: } \mathcal{C} = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|}$$

$$\text{Deletions: } \mathcal{D} = \sum_w \frac{|Y_w - X_w|}{|Y_w|}$$

$$\text{Insertions: } \mathcal{I} = \sum_w \frac{|X_w - Y_w|}{|Y_w|}$$

Evaluation II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vertex Comparison

$$\text{Correct: } \mathcal{C} = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|}$$

$$\text{Deletions: } \mathcal{D} = \sum_w \frac{|Y_w - X_w|}{|Y_w|}$$

$$\text{Insertions: } \mathcal{I} = \sum_w \frac{|X_w - Y_w|}{|Y_w|}$$

Derived Scores

$$\text{Precision: } \frac{\mathcal{C}}{\mathcal{C} + \mathcal{I}}$$

Evaluation II

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Vertex Comparison

$$\text{Correct: } \mathcal{C} = \sum_w \frac{|X_w \cap Y_w|}{|Y_w|}$$

$$\text{Deletions: } \mathcal{D} = \sum_w \frac{|Y_w - X_w|}{|Y_w|}$$

$$\text{Insertions: } \mathcal{I} = \sum_w \frac{|X_w - Y_w|}{|Y_w|}$$

Derived Scores

$$\text{Precision: } \frac{\mathcal{C}}{\mathcal{C} + \mathcal{I}}$$

$$\text{Recall: } \frac{\mathcal{C}}{\mathcal{C} + \mathcal{D}}$$

Results

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

	<i>Linguistica</i>	<i>LSA</i> <i>pr</i> \geq 0.5	<i>LSA</i> <i>pr</i> \geq 0.7	<i>LSA</i> <i>pr</i> \geq 0.9
#Correct ¹	10515	10529	10203	9863
#Inserts	2157	1852	1138	783
#Deletes	2571	2341	2667	3007
Precision	83.0%	85.0%	90.0%	92.6%
Recall	80.4%	81.8%	79.3%	76.6%
F-Score	81.6%	83.4%	84.3%	83.9%

Table: Performance compared to English CELEX

(Schone and Jurafsky, 2000, p. 6)

¹Not \mathcal{C} , but precision and recall stay the same

Conclusion

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Conclusion

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

① Knowledge-free induction of morphology

Conclusion

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 Knowledge-free induction of morphology
- 2 simple suffix induction using a trie

Conclusion

Morphology Induction

Introduction

Morphology Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

- 1 Knowledge-free induction of morphology
- 2 simple suffix induction using a trie
- 3 LSA for clustering words into paradigms

References

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results



Goldsmith, J. (2001).

Unsupervised learning of the morphology of a natural language.

Computational Linguistics, pages 153–189.



Schone, P. and Jurafsky, D. (2000).

Knowledge-free induction of morphology using latent semantic analysis.

In Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, 2000, pages 67–72.

Association for Computational Linguistics, Somerset, New Jersey.

Thank You for Your Attention

Morphology
Induction

Introduction

Morphology
Induction

Stage 1: Finding
Suffixes

Stage 2: Rule
Hypotheses

Stage 3:
Semantic
Vectors

Stage 4: Find
similar semantic
vectors

Results

Questions?