Chapter 8 in Foundations of Statistical NLP

Lexical Acquisition

LCT/LST Master Bridge Course

Sibel Ciddi, Miloš Ercegovcevic, Evi Kiagia,
Mariya Koleva, Antonia Scheidel

Outline

- Introduction to Lexical Acquisition
- Verb Subcategorization
- Selectional Preferences
- Semantic Similarity

Introduction

Introduction

What should you do with this?

A) Take it back to the zoo.

B) Score a goal with it.

C) Have it for dessert.

Introduction

What should you do with this?

A) Take it back to the zoo.

B) Score a goal with it.

C) Have it for dessert.

Introduction

You shall know a word by the company it keeps.
(John Rupert Firth)

Durian

Introduction

You shall know a word by the company it keeps.
(John Rupert Firth)

Durian IS good if you have GOOD durian.
Fresh durian is slightly sweet and smooth and it has a tender chew to it. Bad durian is mushy!
I wouldn't trust durian unless it comes fresh from a tree.

retrieved from http://bit.ly/oHSsjZ

Introduction

Lexicon: That part of the grammar of a language which includes the lexical entries **for all the words** and/or morphemes in the language and which may also include **various other information**, depending on the particular theory of grammar.

Introduction

Lexicon: That part of the grammar of a language which includes the lexical entries for all the words and/or morphemes in the language and which may also include various other information, depending on the particular theory of grammar.

What do we want for MRDs?

quantitative information

syntactically relevant lexical properties

Introduction

Lexicon: That part of the grammar of a language which includes the lexical entries **for all the words** and/or morphemes in the language and which may also include **various other information**, depending on the particular theory of grammar.

What do we want for MRDs?

Verb Subcategorization

Usage: We say a verb **subcategorizes for** different syntactic categories.

Sem. arguments:
Theme, Recipient *Category*

NP, PP *Subcategory*

NP, NP *Subcategory*

He donated his money to the church.

He gave the church his money.

Verb Subcategorization

...helps us parse sentences like these:

She told the man where Peter grew up.

She found the place where Peter grew up.

...because we know this:

Tell	NP NP S
Find	NP NP

subcategorization frame

Verb Subcategorization - Example

We want to learn subcategorization frames.

Frame	Functions	Verb	Example
NP NP	Subj, obj	Greet	She greeted me.
NP S	Subj, clause	Hope	She hopes he will attend.
NP INF	Subj, infinitive	Hope	She hopes to attend.
NP NP S	Subj, obj, clause	Tell	She told me he will attend.
NP NP INF	Subj, obj, infinitve	Tell	She told him to attend.
NP NP NP	Subj, (dir) obj, indir obj	Give	She gave him the book.

Verb Subcategorization - Example

Use frame learning algorithm ("Lerner", Brent, 1993)

This is the question we need to answer.

→ Does verb **v** take frame **f**?

Two Steps:

1: Define regular patterns which indicate the presence of the frame with high certainty. → **Cue**

2: Perform Hypothesis Testing

Lerner Algorithm - Cues

Example for a cue for frame "NP NP":

(OBJ | SUBJ_OBJ | CAP) (PUNC | CC)

I greet Peter .

Lerner Algorithm - Cues

Example for a cue for frame "NP NP":

(OBJ | SUBJ_OBJ | CAP) (PUNC | CC)

I greet Peter .

I arrived on Thursday , as (...)

Lerner Algorithm - Cues

Example for a cue for frame "NP NP":

(OBJ | SUBJ_OBJ | CAP) (PUNC | CC)

✔ I greet Peter .

✘ I arrived on Thursday , as (...)

Lerner Algorithm - Cues

Example for a cue for frame "NP NP":

(OBJ | SUBJ_OBJ | CAP) (PUNC | CC)

✔ I greet Peter .

✘ I arrived on Thursday , as (...)

From the definition:
Cues indicate the presence of a frame with **high certainty**.

Certainty =
Probability of error, ε^j.
How likely is it that we make a mistake if we assign frame f^j to verb v based on cue c^j?

Lerner Algorithm - Corpus

Table 2
Lexical categories used in the definitions of the cues.

SUBJ:	I | he | she | we | they
OBJ:	me | him | us | them
SUBJ_OBJ:	you | it | yours | hers | ours | theirs
DET:	a | an | the | her | his | its | my | our | their | your | this | that | whose
+TNS:	has | have| had | am | is | are | was | were | do | does | did | can | could | may | might | must | will | would
CC:	when | before | after | as | while | if
PUNC:	. | ? | ! | , | ; | :

Lexical categories:
Regular expressions!

↓

Used to "tag" every word in the corpus.

Lerner Algorithm - Corpus

Table 2
Lexical categories used in the definitions of the cues.

SUBJ:	I | he | she | we | they
OBJ:	me | him | us | them
SUBJ_OBJ:	you | it | yours | hers | ours | theirs
DET:	a | an | the | her | his | its | my | our | their | your | this | that | whose
+TNS:	has | have| had | am | is | are | was | were | do | does | did | can | could | may | might | must | will | would
CC:	when | before | after | as | while | if
PUNC:	. | ? | ! | , | ; | :

Lexical categories:
Regular expressions!

↓

Used to "tag" every word in the corpus.

Mark up every verb occurrence with the corresponding frame

Lerner Algorithm - What now?

1) Define cues for all frames of interest
2) For every verb-frame combination:

How often does a cue c^j for the frame f^j occur with the verb v^i?

We will call this the *occurrence count* for later reference

$C(v^i,c^j)$

Lerner Algorithm - Testing

Null Hypothesis H_0: **Verb v^i does not permit frame f^j**

Probability P_E of error for rejecting H_0:

P(no_permission | occurrence_count big enough)

$\mathbf{v^i(f^j) = 0}$ $\quad$ $\mathbf{C(v^i,c^j)}$ $\quad$ $\mathbf{\geq m^*}$

* m: threshold

Lerner Algorithm - Testing

Null Hypothesis H_0: **Verb v^i does not permit frame f^j**

Probability P_E of error for rejecting H_0:

P(no_permission | occurrence_count big enough)

$$= \sum_{r=m}^{n} \binom{n}{r} \epsilon_j{}^r (1-\epsilon_j)^{n-r}$$

ε^j: error rate for cue **f^j**

If $P_E < \alpha$, we reject H_0: **v^i does permit frame f^j!**

Lerner Algorithm - Question

Revision: **Confusion Matrix**

		Predicted Outcome	
		True	False
Actual Value	True	TP	FN
	False	FP	TN

Question: What are TP, TN, FN and FP for our example?

Lerner Algorithm - Question

Confusion Matrix for Lerner Algorithm

		Cue Found?	
		True	False
Permit Frame?	True	TP	FN
	False	FP	TN

Lerner Algorithm - Addition

Manning 1993: Introduction of a tagger

What is different?

- Real tagset instead of regex categories
- Run cue detection on tagger output

What does that change?

- Two error-prone systems
- Unreliable cues are detected → More errors?

Lerner Algorithm - Addition

Unreliable cues are detected → More errors?

Example:

- $\mathbf{c}^j$ with error rate $\boldsymbol{\varepsilon}^j = 0{,}25$
- $C(\mathbf{c}^j) = 11$ (out of 80)
- $P_E \approx 0.011$, $\boldsymbol{\alpha} = 0{,}02$

→ reject H_0 and permit frame

Lerner Algorithm - Addition

Unreliable cues are detected → More errors?

Example:

- $\mathbf{c}^j$ with error rate $\boldsymbol{\varepsilon}^j = 0{,}25$
- $C(\mathbf{c}^j) = 11$ (out of 80)
- $P_E \approx 0.011$, $\boldsymbol{\alpha} = 0{,}02$

→ reject H_0 and permit frame

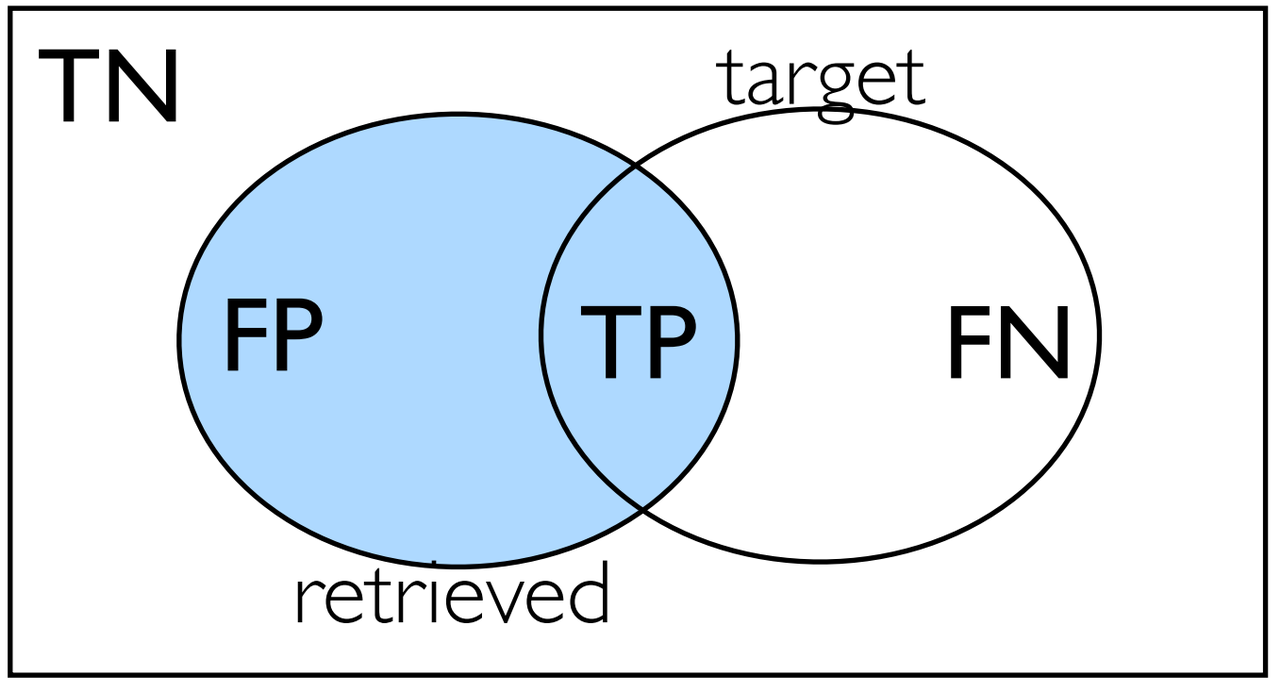
Actually, no.

Number of available cues significantly increased by allowing **low-reliability cues + additional cues based on tagger output.**

→ More verb occurrences have cues for a given frame.

Lerner Algorithm - Question

Revision? **Precision and Recall**



$$\textbf{Precision} = \frac{TP}{TP + FP} \qquad \textbf{Recall} = \frac{TP}{TP + FN}$$

Lerner Algorithm - Question

Compute Precision and Recall for the following results:

Verb	Correct Frames	Incorrect Frames	OALD Frames
bridge	1	1	1
burden	2		2
depict	2		3
emanate	1		1
leak	1		5
occupy	1		3
remark	1	1	4
retire	2	1	5
shed	1		2
troop	0		3
	12	3	29

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Selectional Preferences

Selectional Preferences

- Selectional preferences are semantic constraints that regulate the nature of the arguments of a word.
- Selectional Preferences and NOT selectional rules
- Analogous to subcategorization frames but have to do with the semantic organisation of words.

Examples:

- Drink: Drink coffee, drink tea, drink water e.t.c drink+beverage
- Bark: animate subject(dog)+bark
- Metaphorical and figurative use of languagee.g *"fear's eating the soul"*

Selectional Preferences

Are important because:

- Help us infer the category of a word even though we may not know what it means

Example:

"Susan had never eaten a durian before"

- What can we infer about the durian?

Selectional Preferences

Resnik's model(1993,1996)

Selectional Preference Strength:

- Illustrates how strong is the relationship between the verb and the direct object.
- Distinction between head of the Phrase

  e.g. the green apple → apple
- Distinction into classes of words, something that help us generalize and parametrize words according to their class.

$$S(v) = D(P(C|v) \| P(C)) = \sum_c P(c|v) \log\left(\frac{P(c|v)}{P(c)}\right)$$

$$S(v) = D(P(C|v) \| P(C)) = \sum_c P(c|v) \log\left(\frac{P(c|v)}{P(c)}\right)$$

- P(C) is the overall probability distribution of Noun Classes
- P(C/v): probability distribution of noun clauses in the direct object position of v

*Nouns are taken from any lexical resource that groups nouns into verb, e.g. *Wordnet*

Selectional Association

Selectional Association Strenght”

- Of a noun ”n” and a single class ”c”

A(v,n) = A(v,c)

- Of a noun ”n” belonging to more than one classes ”c”

A(v,n)= max A(v,c)

Its association strenght is the highest association strenght of any of its classes

Example

"Susan interrupted the chair"

- "Chair": Polysemous as it belongs to more than one classes (furniture) and (people)
- A(interrupt, people)>> A (interrupt,furniture)

A(interrupt, chair) belongs to the class of people so we can easily disambiguate chair.

Selectional Preferences Strength

- Help us calculate the strenght between noun classes and verbs
- Understand which arguments are more preferred and which are dispreferred

Semantic Similarity

Semantic Similarity
Intuitive Notion

- How we usually think about it:
  - Synonymy – we see the words as largely interchangeable:

    carpet/rug, *drink/beverage*

- Expanded notion:
  - Words of the same semantic domain: *boy/youth*
  - Words about entities that co-occur in the real world, even if they are of different syntactic categories: *chef, sautee, savoury*

Intuitive Notion (cont'd)

- Miller and Charles (1991) – words are similar if:
  - They are more or less interchangeable in the same context (cf. *carpet/rug*);
  - A word is „similar to the appropriate sense“ of another (ambiguous) word, i.e. a word *w1* would usually be similar only to one sense of an ambiguous word *w2:*

a record/an account of an event

**a world record/account*

**a criminal record/account*

Two assumptions:

- Semantic properties of a word can be acquired on the basis of semantic similarity or dissimilarity;

- Semantically similar words behave similarly.

Uses

- Generalisation:
  - Similarity-based – the nearest neighbours can help us generalize about an unknown element:
    cf. the *durian* from *selectional preference*;
  - Class-based – we look not only at the nearest neighbours; we speculate about the entire class to which the word might belong;

- K Nearest Neighbours classification task:
  - A training set of words, assigned to categories;
  - Task: assign an element to a category that is prevalent for *K*'s nearest neighbours;

Types of Similarity Measures

- Vector Space measures:
  - (+) Conceptually simple;
  - (-) Lack clear interpretation of the computed measure;

- Probabilistic measures:
  - (+) solid theoretical footing;
  - (-) rely on some additional transformations;

Vector Space Measures (1)

- Words can be represented as vectors in space;
- Possible spaces:
  - Document space;
  - Word space;

Give topical similarity

  - Modifier space;

Give different semantic properties

Vector Space Measures (2)

	cosmonaut	astronaut	moon	car	truck
d_1	1	0	1	1	0
d_2	0	1	1	0	0
d_3	1	0	0	0	0
d_4	0	0	0	1	1
d_5	0	0	0	1	0
d_6	0	0	0	0	1

Figure 8.3 A document-by-word matrix A.

	cosmonaut	astronaut	moon	car	truck
cosmonaut	2	0	1	1	0
astronaut	0	1	1	0	0
moon	1	1	2	1	0
car	1	0	1	3	1
truck	0	0	0	1	2

Figure 8.4 A word-by-word matrix B.

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Figure 8.5 A modifier-by-head matrix C. The nouns (or heads of noun phrases) in the top row are modified by the adjectives in the left column.

Vector Space Measures (3)

- How do we use vectors?
  - With binary or real values;
  - (binary) vector – the set of non-zero values;
  - Use set operations to calculate similarities;

- e.g. The vector for *cosmonaut* in matrix B is {*Soviet, spacewalking*}

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Vector Space Measures (4)

Similarity measure	Definition
matching coefficient	$X \cap Y$
Dice coefficient	$\frac{2\|X \cap Y\|}{\|X\| + \|Y\|}$
Jaccard (or Tanimoto) coefficient	$\frac{\|X \cap Y\|}{\|X \cup Y\|}$
Overlap coefficient	$\frac{\|X \cap Y\|}{\min(\|X\|, \|Y\|)}$
cosine	$\frac{\|X \cap Y\|}{\sqrt{\|X\| \times \|Y\|}}$

Table 8.7 Similarity measures for binary vectors.

Matching coefficient

- It only counts the number of dimensions which are not zero in both vectors:

  cosmonaut = {Soviet, spacewalking}

  astronaut = {American, spacewalking}

  matching coefficient = 1

- The vector length is irrelevant;
- The total number of non-zero dimension in each vector is irreloevant;

Dice Coefficient

- Normalizes for length
- Divides by total number of non-zero entries

$$\frac{2|X \cap Y|}{|X| + |Y|}$$

Shared entries

Total-non zero

Total non-zero

cosmonaut = {1,0,1} ({Soviet, American, Spacewalking})

astronaut = {0,1,1}

Dice = ?

Jaccard coeffiecient

- Penalizes more than Dice if the number of shared entries is small

$$\frac{|A \cap B|}{|A \cup B|}$$

cosmonaut = {1,0,1} ({Soviet, American, Spacewalking})

astronaut = {0,1,1}

Jaccard = ?

Overlap coefficient

- Checks if the two sets overlap – i.e., if every non-zero entry in one vector is also non-zero in the second vector

$$\frac{|X \cap Y|}{\min(|X|, |Y|)}$$

Cosine

- Useful when we compare words for which we have different amounts of data;
- If the vectors have different number of non-zero entries, the cosine penalizes less than Dice.
- Can be used with real value vectors; the angle between two vectors.

$$\frac{|v \cap w|}{\sqrt{|v| \times |w|}}$$

$$\frac{\vec{v} \cdot \vec{w}}{|v| \times |w|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \times \sqrt{\sum_{i=1}^{n} w_i^2}}$$

Probabilistic measures (1)

- Problem: vector space measures operate on binary value vectors;
- We need similarity measures which deal with counts and probabilities;
- Semantic similarity can be viewed as the similarity or dissimilarity between two probability distributions;

Probabilistic measures (2)

- Matrices of counts can be transformed into matrices of probabilities:

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

- P (spacewalking | astronaut) = ½ = 0.5
- P (red | truck) = ?

Probabilistic measures (3)

- Three measures of dissimilarity between probability distributions:

(Dis-)similarity measure	Definition
KL divergence	$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$
information radius (IRad)	$D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})$
L_1 norm	$\sum_i \|p_i - q_i\|$

Kullback-Leibler Divergence

- Measures how well distribution q approximates distribution p;
- It is assimetric, although we intuitively see similarity as symmetric;
- May be undefined, if the denominator is 0, which is often the case;

Information Radius

- Based on KL-divergence, but it is always finite;
- Compares total divergence between *p* and *q* to the average of p and q;
- Is symmetric;

L_1 (Manhattan) Norm

- Measures the expected proportions of different events;
- Is symmetrical;
- Is well-defined;

Probabilistic Measures in Comparison

- Dagan et al. (1997) find that lRad performs better than the other two metrics and recommend the use of it;