FSNLP chap.8 Lexical Acquisition

Saku Sugawara

Univ. of Tokyo, Bungakubu

November 21, 2014

Precision and Recall

- **適合率と再現率**
- evaluation mesures in Information Retrieval
- selected items and target items
- true/false, positive/negative

	target	¬target
selected	tp	fp
¬selected	fn	tn

Precision and Recall

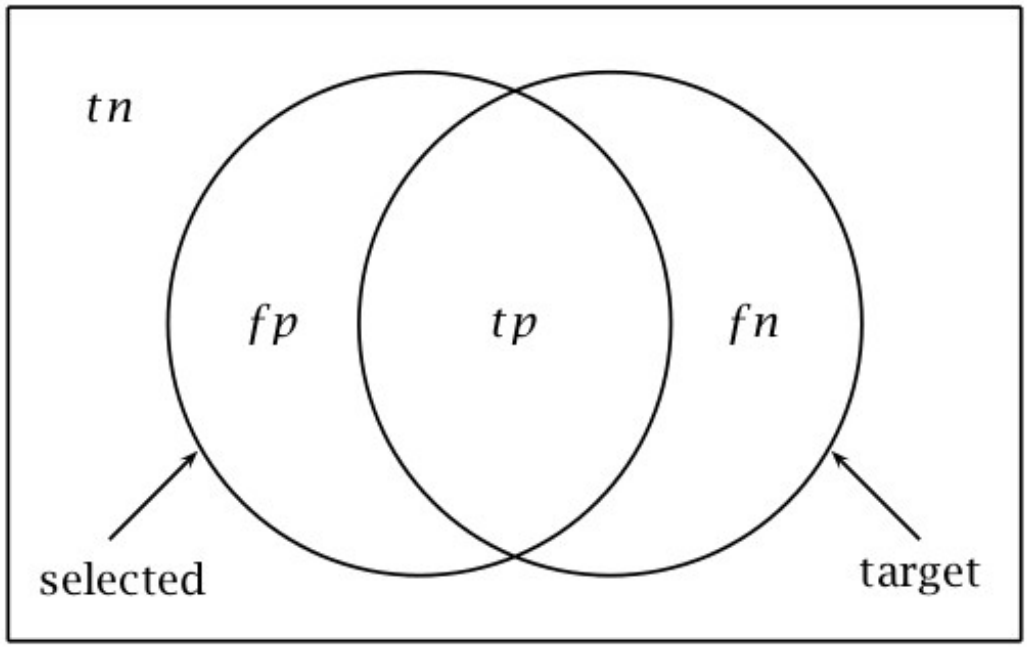


Figure 8.1 A diagram motivating the measures of precision and recall. The areas counted by the figures for true and false positives and true and false negatives are shown in terms of the target set and the selected set. Precision is $tp/|\text{selected}|$, the proportion of target (or correct) items in the selected (or retrieved) set. Recall is $tp/|\text{target}|$, the proportion of target items that were selected. In turn, $|\text{selected}| = tp + fp$, and $|\text{target}| = tp + fn$).

Precision and Recall

Precision

$$P = \frac{tp}{|\text{selected}|} = \frac{tp}{tp + fp}$$

Recall

$$R = \frac{tp}{|\text{target}|} = \frac{tp}{tp + fn}$$

F measure

F measure

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

- P is precision and R is recall
- α is a factor of the weighting of precision and recall
- A value of $\alpha = 0.5$ is often chosen for equal weighting of P and R. With this α value,

$$F = \frac{2PR}{R+P}.$$

Accuracy and Error

"Why don't we just report the percentage of things right or the percentage of things wrong?"

- things right:

$$\text{accuracy} = tp + tn$$

- things wrong:

$$\text{error} = fp + fn$$

These often aren't error good measures to use because the number of non-target and non-selected things, tn is huge, and dwarfs all the other numbers.

Fallout

fallout

$$\text{fallout} = \frac{fp}{|\neg target|} = \frac{fp}{fp + tn}$$

- Fallout is sometimes used as a measure of how hard it is to build a system that produces few false positives.

8.2.0 Introduction

Frame	Functions	Verb	Example
NP NP	subject,object	greet	She greeted me.
NP S	subject,clause	hope	She hopes he will attend.
NP INF	subject,infinitive	hope	She hopes to attend.
NP NP S	subject,object,clause	tell	She told me he will attend.
NP NP INF	subject,object,infinitive	tell	She told him to attend.
NP NP NP	subject,(direct)object,indirect object	give	She gave him the book.

- **下位範疇化フレーム**
- A particular set of syntactic categories that a verb can appear with is called a subcategorization frame.
- We sometimes omit subjects from subcategorization frames.

Importance of Subcategorization frame

8.7

a. She told the man where Peter grew up.

b. She found the place where Peter grew up.

If we know that *tell* has the subcategorization frame NP NP S (subject, object, clause), and that *find* lacks that frame, but has the subcategorization frame NP NP (subject, object), we can correctly parse the sentences:

Importance of Subcategorization frame

8.8

a. She told [the man] [where Peter grew up].
b. She [found the place [where Peter grew up]].

Verb	Frame	Functions
tell	NP NP S	subject,object,clause
find	NP NP	subject,object

Unfortunately...

- Most dictionaries do not contain information on subcategorization frames.
- According to one account, up to 50% of parse failures can be due to missing subcategorization frames.(John Carroll 1998)
- Even the most comprehensive source of subcategorization information does not have quantitative information such as the relative frequency of different subcategorization frames for a verb.

8.2.1 Lerner

Algorithm for learning some subcategorization frames (Brent 1993)
2 steps, Cues. and Hypothesis testing.

Step1: Cues.

- Define a regular pattern which indicates the presence of the frame
- For a particular cue c^j we define a probability of error ϵ_j that indicates how likely we are to make a mistake if we assign frame f to verb v based on cue c^j

c^j: regular pattern
ϵ_j: probability of error in assignment

8.2.1 Lerner

Step2: Hypothesis testing.

- Null hypothesis H_0: the frame is not appropriate for the verb
- We reject this hypothesis if the cue c^j indicate with high probability that our H_0 is wrong.

Cues.

8.9

Cue for frame "NP NP":
(OBJ | SUBJ_OBJ | CAP) (PUNC | CC)

OBJ	objective personal pronouns like me and him
SUBJ_OBJ	subjective and objective personal pronouns like you and it
CAP	capitalized word
PUNC	punctuation mark like ",.!?" etc.
CC	subordinating conjunction like if, $before$ or as

Cues.

Instantiations of “CAP PUNC” pattern:

8.10

[...] greet-V Peter-CAP ,-PUNC [...]

A verb indeed takes the frame “NP NP”.

8.11

I came Thursday, before the storm started.

The verb doesn’t allow the frame, but this case is very unlikely.

Hypothesis testing.

8.12 probability of error for null hypothesis

$$p_E = P(H_0 = \text{true} | C(v^i, c^j) \geq m)$$
$$= \sum_{r=m}^{n} \binom{n}{r} \epsilon_j^r (1 - \epsilon_j)^{n-r}$$

- verb v^i occurs a total of n times in the corpus
- there are $m \leq n$ occurrences with a cue for frame f^j
- we can reject H_0 that v^i does not permit f^j with p_E.

Hypothesis testing.

- We will reject the null hypothesis if $p_E < \alpha$ for an appropriate level of significance α, for example, $\alpha = 0.02$. For $p_E \geq \alpha$, we will assume that verb v^i does not permit frame f^j .

Lerner - cont.

- Precision is high, but recall is low.
- Even an unreliable indicator is helpful.
  - For example, if cue c^j with error rate $\epsilon_j = 0.25$ occurs 11 out of 80 times, then we can still reject the null hypothesis with $p_E \approx 0.011 < 0.02$ despite the low reliability of c^j .
- One way to improve these results would be to incorporate prior knowledge about a verb ' s subcategorization frame

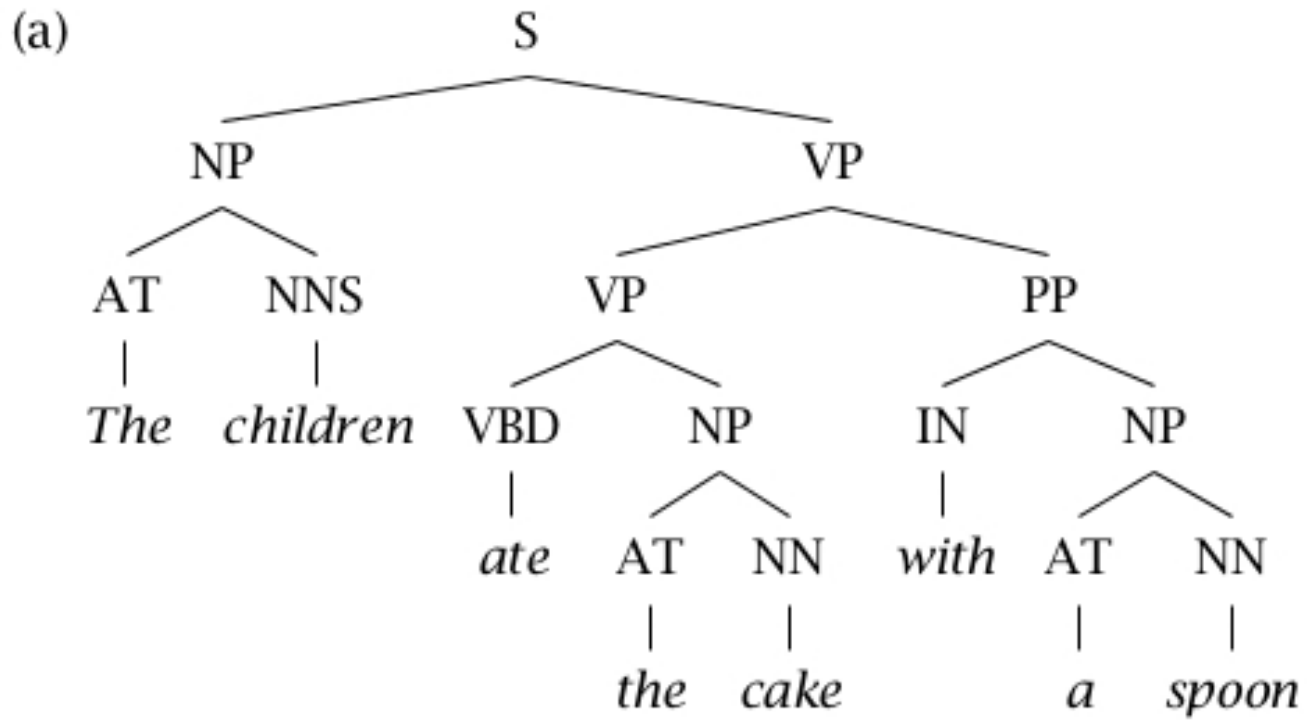
Introduction

- **連結曖昧さ**
- PP(prepositional phrase) attachment is the attachment ambiguity problem that has received the most attention in the Statistical NLP literature.
- In this section, we introduce a method for determining the attachment of prepositional phrases based on lexical information.

8.14

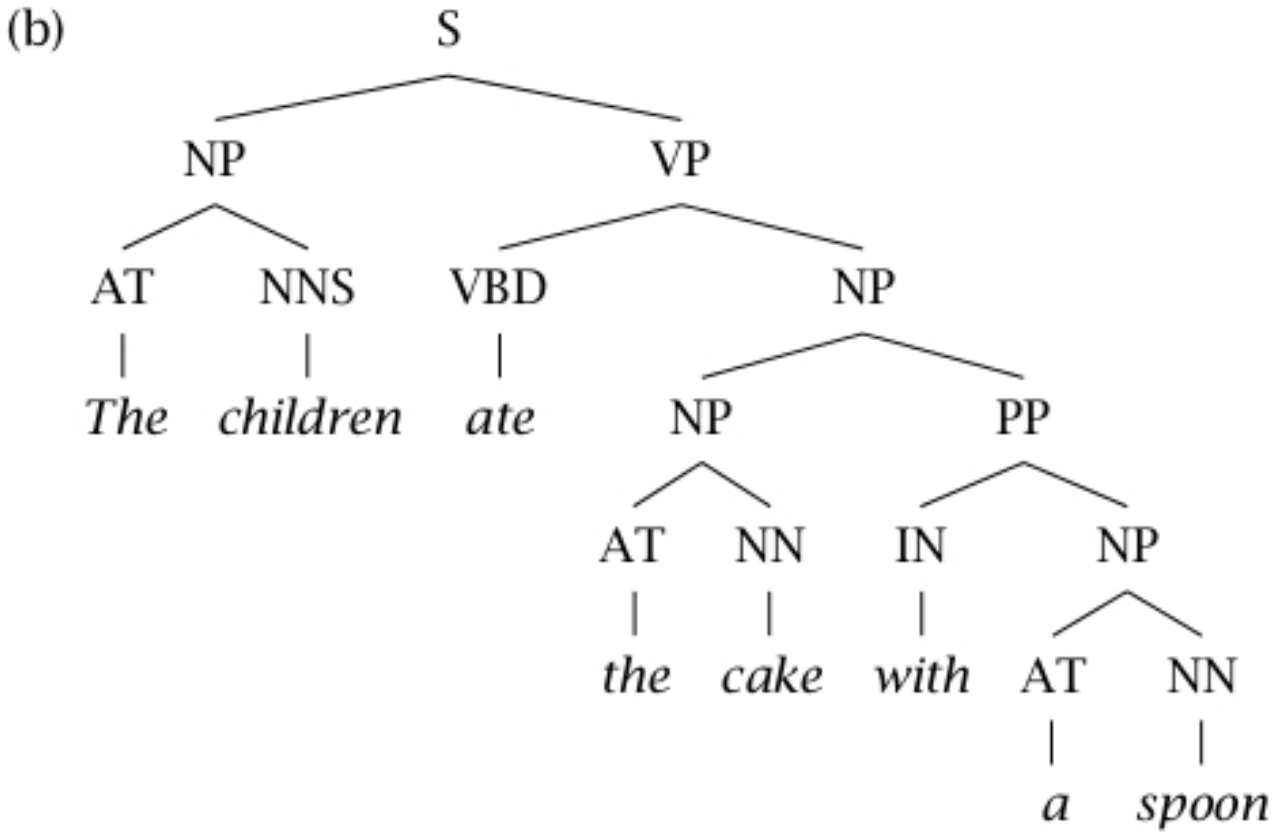
The children ate the cake with a spoon.

Introduction



The children [ate the cake with a spoon].

Introduction



The children ate [the cake with a spoon].

Introduction

8.15

a. Moscow sent more than 100,000 soldiers into Afghanistan...
b. Sydney Water breached an agreement with NSW Health...

In cases like these, lexical preferences can be used to disambiguate. These simple statistics are basically co-occurrence counts between the verb/noun and the preposition.

A simple model: likelihood ratio λ

8.16

$$\lambda(v, n, p) = \log \frac{P(p|v)}{P(p|n)}$$

$P(p|v)$: the probability of seeing a PP with p after the verb v
$P(p|n)$: the probability of seeing a PP with p after the noun n We can then attach to the verb for $\lambda(v, n, p) > 0$ and to the noun for $\lambda(v, n, p) < 0$.

Low attachment

- There is a preference for attaching phrases "low" in the parse tree.
- For PP attachment, the lower node is the NP node.

Low attachment

8.17

Chrysler confirmed that it would end its troubled venture with Maserati.

- The preposition $with$ occurs frequently after both end and $venture$.
- The λ model is wrong because equation (8.16) ignores a bias for low attachment in cases where a preposition is equally compatible with the verb and the noun.

8.3.1 Hindle and Rooth(1993)

- We define the event space to consist of all clauses that have a transitive verb , an NP following the verb (the object noun phrase) and a PP following the NP.
- To simplify the probabilistic model, we will estimate how likely it is in general for a preposition to attach to a verb or noun.

8.3.1 Hindle and Rooth(1993)

We will look at the following two questions, formalized by the sets of indicator random variables VA_p and NA_p:

- VA_p: Is there a PP headed by p and following the verb v which attaches to v ($VA_p = 1$) or not ($VA_p = 0$)?
- NA_p: Is there a PP headed by p and following the noun n which attaches to n ($NA_p = 1$) or not ($NA_p = 0$)?

8.3.1 Hindle and Rooth(1993)

8.19 and 8.20

$$\begin{aligned} P(VA_p, NA_p|v,n) &= P(VA_p|v,n)P(NA_p|v,n) \\ &= P(VA_p|v)P(NA_p|n) \end{aligned}$$

The advantage of the independence assumption is that it is easier to derive empirical estimates for the two variables separately rather than estimating their joint distribution.

8.3.1 Hindle and Rooth(1993)

Attach(p)

$$P(Attach(p) = n|v, n) = P(NA_p = 1|n)$$
$$P(Attach(p) = v|v, n) = P(VA_p = 1|v)P(NA_p = 0|n)$$

A likelihood ratio λ

$$\lambda = \log \frac{P(Attach(p) = v|v, n)}{P(Attach(p) = n|v, n)}$$
$$= \log \frac{P(VA_p = 1|v)P(NA_p = 0|v)}{P(NA_p = 1|n)}$$

8.3.1 Hindle and Rooth(1993)

where

$$P(VA_p = 1|v) = \frac{C(v,p)}{C(v)}$$

$$P(NA_p = 1|n) = \frac{C(n,p)}{C(n)}$$

The remaining difficulty is to determine the attachment counts from an unlabeled corpus.

8.3.1 Hindle and Rooth(1993) - cont.

Hindle and Rooth (1993) propose a heuristic for determining $C(v, p)$ and $C(n, p)$ from unlabeled data that has essentially three steps.

1. Build an initial model by counting all unambiguous cases.
2. Apply the initial model to all ambiguous cases and assign them to the appropriate count if λ exceeds a threshold (for example, $\lambda > 2.0$ for verb attachment and $\lambda < -2.0$ for noun attachment).
3. Divide the remaining ambiguous cases evenly between the counts (that is, increase both $C(v, p)$ and $C(n, p)$ by 0.5 for each ambiguous case).

8.3.1 Hindle and Rooth(1993) - cont.

In general, the procedure is accurate in about 80% of cases. We can trade higher precision for lower recall if we only make a decision for values of λ that exceed a certain threshold. For example, Hindle and Rooth (1993) found that precision was 91.7% and recall was 55.2% for $\lambda = 3.0$.

8.3.2 a first major limitation

- Sometimes other information is important (studies suggest human accuracy improves by around 5% when they see more than just a v,n,p triple).
- In particular, in sentences like those in (8.25), the identity of the noun that heads the NP inside the PP is clearly crucial:

8.25

a. I examined the man with a stethoscope.

b. I examined the man with a broken leg.

v, n, p **以外の意味、とくに前置詞句の中の名詞句の意味も大事**

8.3.2 a second major limitation

Hindle and Rooth (1993) consider only the most basic case of a PP immediately after an NP object which is modifying either the immediately preceding noun or verb. But there are many more possibilities for PP attachments than this.

直後にくる前置詞句にしか着目しておらず、距離が離れた複雑な前置詞関係に対応できない

noun compounds

the left-branching structure [[N N] N]

door bell manufacturer = [[door bell] manufacturer]

the right-branching structure [N [N N]]

woman aid worker = [woman [aid worker]].

noun compounds

- The left-branching case roughly corresponds to attachment of the PP to the verb ([V N P]), while the right-branching case corresponds to attachment to the noun ([V [N P]]).
- We could directly apply the formalism we ' ve developed for prepositional phrases to noun compounds.
- However, data sparseness tends to be a more serious problem for noun compounds than for prepositional phrases because prepositions are high-frequency words whereas most nouns are not.
- For this reason, one approach is to use some form of semantic generalization based on word classes in combination with attachment information.

indeterminacy

A large proportion of prepositional phrases exhibit "indeterminacy" with respect to attachment.

8.26

We have not signed a settlement agreement with them.

Lauer (1995a) found that a significant proportion of noun compounds also had this type of attachment indeterminacy.

8.4.0 Introduction

Most verbs prefer arguments of a particular type.

- the objects of the verb eat tend to be food items
- the subjects of think tend to be people
- the subjects of bark tend to be dogs

8.4.0 Introduction

We use the term preferences as opposed to rules because the preferences can be overridden in metaphors and other extended meanings. For example, eat takes non-food arguments in eating one ' s words or fear eats the soul.
preferences **は** rules **よりも弱い制約として存在する（暗喩や意味の拡張のために破られることが許される）**

8.4.0 Introduction

The acquisition of selectional preferences is important in Statistical NLP for a number of reasons.

1. If a word like durian is missing from our machine-readable dictionary, then we can infer part of its meaning from selectional restrictions. =**語義推定**
2. for ranking the possible parses of a sentence
   自動化された言語処理においては、文の意味を理解しようとするよりも選択制限に基づいてランク付けさせるほうが容易

8.4.1 Resnik (1993,1996)

- We will now introduce the model of selectional preferences proposed by Resnik (1993, 1996).
- We will only consider the case ‘ verb direct object ’ here, that is, the case of verbs selecting a semantically restricted class of direct object noun phrases.
- The model formalizes selectional preferences using two notions: selectional preference strength and selectional association.

Selectional Preference Strength (SPS)

SPS: how strongly the verb constrains its direct object

two assumptions

1. We only take the head noun of the direct object into account(for example, $apple$ in Susan ate the green apple).
2. Instead of dealing with individual nouns, we will instead look at classes of nouns.

Selectional Preference Strength (SPS)

Selectional Preference Strength $S(v)$

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

Selectional Preference Strength (SPS)

where

$$P(c|v) = \frac{P(v,c)}{P(v)}$$

$$P(v,c) = \frac{1}{N} \sum_{n \in \mathsf{words}(c)} \frac{1}{|\mathsf{classes}(n)|} C(v,n)$$

$P(C)$: the overall probability distribution of noun classes
$P(c|v)$: the probability distribution of noun classes in the direct object position of v.

Selectional Preference Strength (SPS)

Nounclass :c	$P(c)$	$P(c\|eat)$	$P(c\|see)$	$P(c\|find)$
people	0.25	0.01	0.25	0.33
furniture	0.25	0.01	0.25	0.33
food	0.25	0.97	0.25	0.33
action	0.25	0.01	0.25	0.01
SPS :$S(v)$		1.76	0.00	0.35

- **ひとつの動詞に対して定義され、名詞のクラスの分布に対する選好性の強さを表す。0 に近いほど選好性が弱く、どのクラスの名詞に対しても使用されやすくなり、値が大きいほど選好性が強く、特定のクラスの名詞に対してのみ使用されるようになる。**

Selectional Association

Selectional association

$$A(v,c) = \frac{P(c|v)\log\frac{P(c|v)}{P(c)}}{S(v)} A(v,n) \quad = \max c \in \mathsf{classes}(n) A(v,c)$$

8.32, 8.33

$$A(eat, \mathsf{food}) = 1.08$$
$$A(find, \mathsf{action}) = -0.13$$

Selectional Association

Verb :v	Noun :n	$A(v,n)$	Class	Noun :n	$A(v,n)$	Class
$answer$	$request$	4.49	speechact	$tragedy$	3.88	communication
$find$	$label$	1.1	abstraction	$fever$	0.22	psych.feature
$hear$	$story$	1.89	communication	$issue$	1.89	communication
$remember$	$reply$	1.31	statement	$smoke$	0.2	articleofcommerce
$repeat$	$comment$	1.23	communication	$journal$	1.23	communication
$read$	$article$	6.8	writing	$fashion$	−0.20	activiy
see	$friend$	5.79	entity	$method$	−0.01	method
$write$	$letter$	7.26	writing	$market$	0	commerce

implicit object alternation

8.35

a. Mike ate the cake.

b. Mike ate.

- selectional preference strength is a good predictor of the permissibility of the implicit-object alternation for verbs.
- **選好性が強いものほど省略しやすくなる（直観に合ってる）**

8.5.0 Introduction

- Automatically acquiring a relative measure of how similar a new word is to known words (or how dissimilar) is much easier than determining what the meaning actually is.
  - **意味を直接獲得するのはよくわからないので類似度でやります**
- not synonymy but the same semantic domain or topic.
- not $dwelling/abode$, but $doctor, nurse, fever, andintravenous$

8.5.1 Vector Space Measures

Document Space

	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

Table : Fig 8.3 A document-by-word matrix A

- Entry a_{ij} contains the number of times word j occurs in document i.
- Matrix A defines similarity between documents.

8.5.1 Vector Space Measures

Word Space

	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

Table : Fig 8.4 A word-by-word matrix B

- Entry b_{ij} contains the number of times word j co-occurs with word i.
- Co-occurrence can be defined with respect to documents, paragraphs or other units.

8.5.1 Vector Space Measures

Modifier Space

	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

Table : Fig 8.5 A modifier-by-head matrix C

- Entry c_{ij} contains the number of times that head j is modified by modifier i.
- Matrix C defines similarity between modifiers.

Similarity measures for binary vectors

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

Table : Similarity measures for binary vectors.

Similarity measures for binary vectors

Cosine is useful for Statistical NLP

- The cosine penalizes less in cases where the number of non-zero entries is very different.
- This property of the cosine is important in Statistical NLP since we often compare words or objects that we have different amounts of data for, but we don ' t want to say they are dissimilar just because of that.
  - **データの次元の数が極端に異なっていても、一致しているものがあれば他の指標より高めの数値を出してくれる**
- The cosine is also the most important one for real-valued vectors.
- Intuitive simplicity and computational efficiency

8.5.2 Probabilistic measures

- Computing the cosine assumes a Euclidean space.
- The Euclidean distance is appropriate for normally distributed quantities, not for counts and probabilities.
  - **回数や確率を比較する手段が欲しい**
- KL divergence / information radius / L_1 norm

8.5.2 Probabilistic measures

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

Table : Measures of (dis-)similarity between probability distributions.

KL divergence

KL divergence indicates how much information is lost if we assume distribution q when the true distribution is p,
and has two problems:

1. we get a value of ∞ if there is a dimension with $q_i = 0$ and $p_i \neq 0$
2. KL divergenc is asymmetric.

IRad, L_1 norm

- information radius: overcomes these KL divergence's problems.
  how much information is lost if we describe the two words
- L1 norm:
  a measure of the expected proportion of different events
- Conclusion: Dagan et al. (1997b) show that IRad consistently performs better than KL and L_1. Consequently, they recommend IRad as the measure that is best to use in general.

8.6 The Role of Lexical Acquisition in Stat. NLP

Lexical acquisition plays a key role in Statistical NLP

1. the cost of building lexical resources manually
2. quantitative information
   - "one cannot learn a new language by reading a bilingual dictionary".
3. inherent productivity of language