Основы обработки текстов

Лекция 9

Лексическая семантика

Возможные взгляды на семантику

- Лексическая семантика
  - значение индивидуальных слов
- Композиционная семантика
  - как значения комбинируются и определяют новые значения для словосочетаний
- Дискурс или прагматика
  - как значения комбинируются между собой и другими знаниями, чтобы задать значение текста или дискурс

План

- Основные понятия
  - слова и отношения между ними
  - словари и тезаурусы
- Вычислительная семантика
  - Разрешение лексической многозначности
  - Семантическая близость слов
  - Некоторые современные направления

Основные понятия

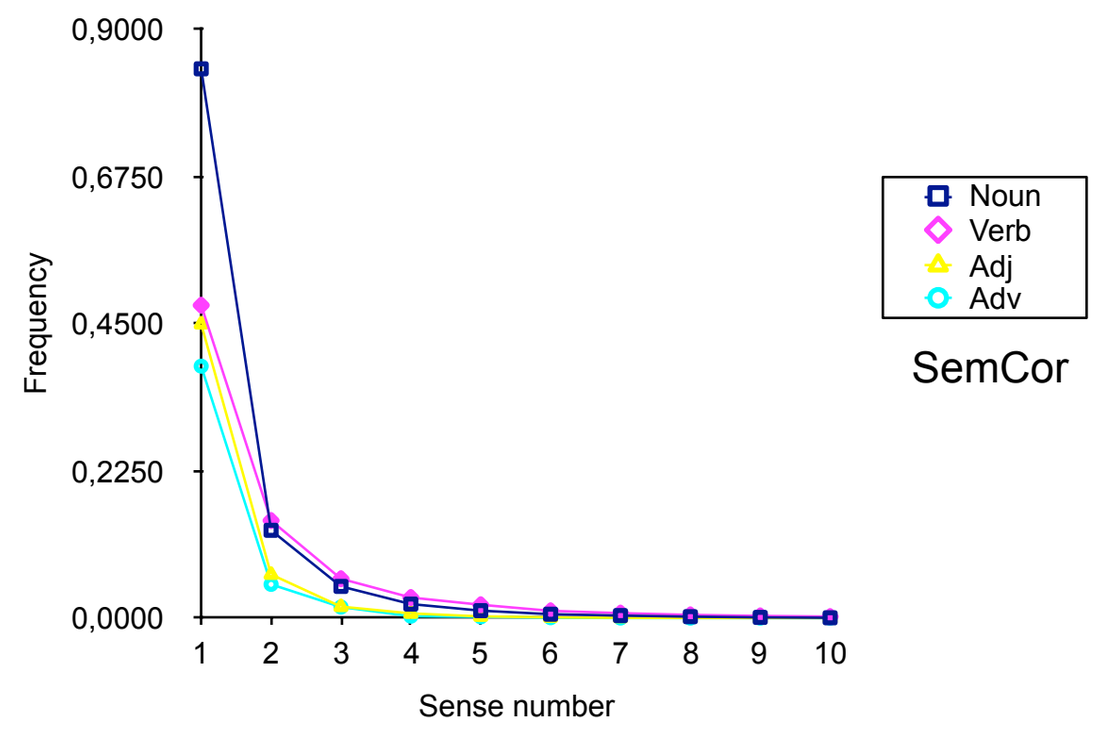
- Значение слова и многозначность
- Омонимия VS многозначность
  - –ключ
  - –платформа
- Метонимия
  - – Я три *тарелки* съел
- Зевгма
  - –За окном шел снег и рота красноармейцев
- Типы омонимов
  - –омофоны (луг-лук, плод-плот)
  - –омографы (м’ука - мук’а, гв’оздик-гвозд’ик)

Отношения между словами

- Синонимия
  - Машина / автомобиль
- Антонимия
  - большой / маленький, вверх / вниз, ложь / истина
- Обобщение и детализация (hyponym and hypernym/superordinate)
  - машина - транспорнтое средство
  - яблоко - фрукт
- Меронимы (партонимы) и холонимы
  - колесо - машина

Многозначность на практике

- Text-to-Speech
  - омографы
- Информационный поиск
- Извлечение информации
- Машинный перевод
- Закон Ципфа (Zipf law)



WordNet

- База лексических отношений
  - содержит иерархии
  - сочетает в себе тезаурус и словарь
  - доступен on-line
  - разрабатываются версии для языков кроме английского (в т.ч. для русского)

Категория	Уникальных форм
Существительные	117,097
Глаголы	11,488
Прилагательные	22,141
Наречия	4,601

- http://http://wordnet.princeton.edu/
- http://wordnet.ru/

Формат WordNet

The noun "bass" has 8 senses in WordNet.

1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.

1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
   "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

WordNet: отношения между словами

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Opposites	$leader^1 \rightarrow follower^1$

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \Longleftrightarrow decrease^1$

Иерархии WordNet

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

```
Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

Как “зачение” определяется в WordNet

- Множестно синонимов называется синсет
- Пример

```
from nltk.corpus import wordnet
for synset in wordnet.synsets('chick'):
    print synset.definition
    print [lemma.name for lemma in synset.lemmas]
```

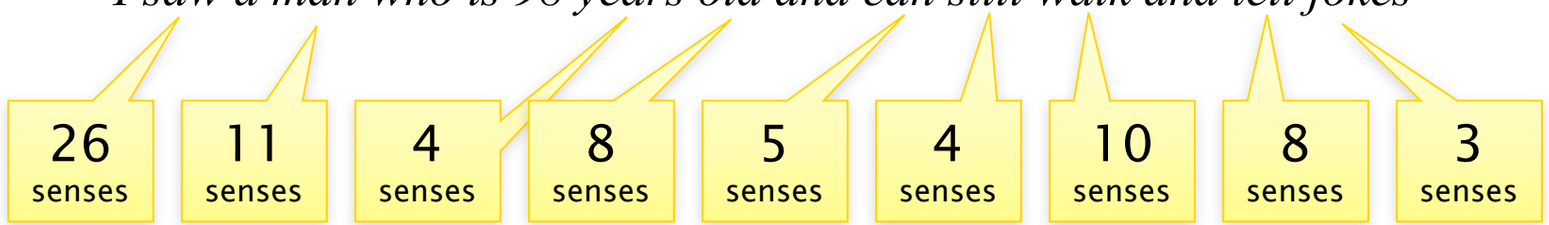
```
young bird especially of domestic fowl
['chick', 'biddy']
informal terms for a (young) woman
['dame', 'doll', 'wench', 'skirt', 'chick', 'bird']
```

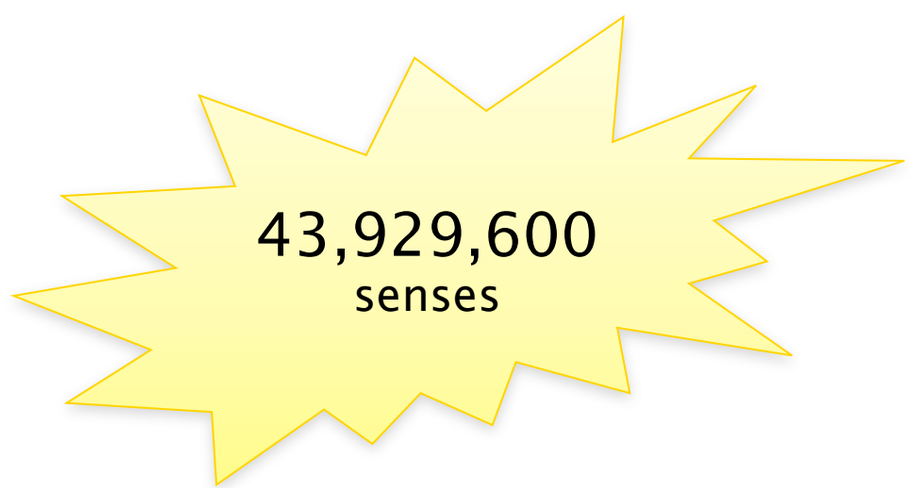
Вычислительная лексическая семантика

- Разрешение лексической многозначности
- Семантическая близость слов

Трудность разрешения лексической многозначности

I saw a man who is 98 years old and can still walk and tell jokes





Разрешение лексической многозначности (РЛМ)

- Word Sense Disambiguation (WSD)
  - определение значения слова в контексте
  - обычно предполагается фиксированный список значений (например WordNet)
- Сводится к задаче классификации
- Отличается от задачи разграничения значений (word sense discrimination)

РЛМ: варианты

- Определение значений только заранее выбранных слов (lexical sample task)
  - line - hard - serve; interest
  - Ранние работы
  - Обучение с учителем
- Определение значений всех слов (all-word task)
  - Проблема разреженности данных
  - Невозможно натренировать отдельный классификатор для каждого слова

Признаки

- Должны описывать контекст
- Предварительная обработка текста
  - параграфы, предложения, части речи, леммы, синтаксический разбор?
- Признаки в словосочетаниях с позициями
- Множества соседей

- Проблема разреженности языка
  - Использовать семантическую близость (далее)

Пример

An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

Collocational features	
word_L3	electric
POS_L3	JJ
word_L2	guitar
POS_L2	NN
word_L1	and
POS_L1	CC
word_R1	player
POS_R1	NN
word_R2	stand
POS_R2	VB
word_R3	off
POS_R3	RB

Bag-of-words features	
fishing	0
big	0
sound	0
player	1
fly	0
rod	0
pound	0
double	0
runs	0
playing	0
guitar	1
band	0

Алгоритмы

- Любые методы классификации
  - (Пример) Наивный байесовский классификатор

Наивный байесовский классификатор

- Выбор наиболее вероятного значения

$$\hat{s} = \arg\max_{s \in S} P(s|f)$$

- По правилу Байеса

$$\hat{s} = \arg\max_{s \in S} \frac{P(s)P(f|s)}{P(f)} = \arg\max_{s \in S} P(s)P(f|s)$$

- Наивное предположение об условной независимости признаков

$$\hat{s} = \arg\max_{s \in S} P(s) \prod_{j=1}^{n} P(f_i|s)$$

Обучение наивного байесовского классификатора

- Метод максимального правдоподобия
- Другими словам, просто считаем

$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)} \qquad P(f_j \mid s) = \frac{count(f_j, s)}{count(s)}$$

- Алгоритм прост в реализации, но
  - Исчезновение значащих цифр → использовать сумму логарифмов вместо произведения
  - Нулевые вероятности → сглаживание

Вопрос на засыпку

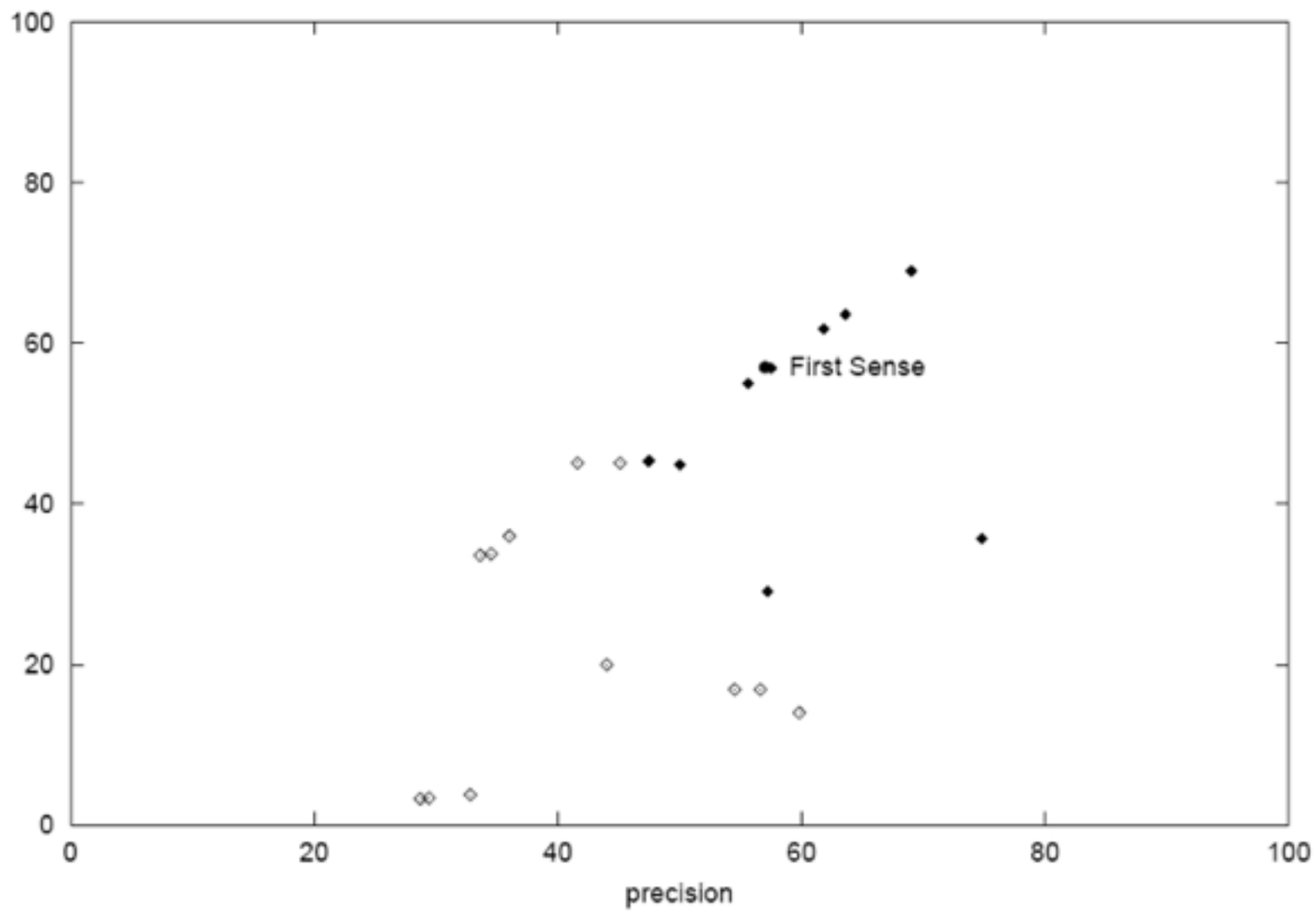
- Как сделать классификатор для задачи определения значений всех слов (all-word task)?

Методы оценки

- Внешние (in vivo)
  - Машинный перевод с/без РЛМ
- Внутренние (in vitro)
  - Применение к размеченным данным (SemCor, SENSEVAL, SEMEVAL)
  - Измерение точности и полноты в сравнении со стандартными значениями
- Нижняя граница
  - Выбор случайных значений работает плохо
  - Более сильные границы: наиболее частое значение, алгоритм Леска
- Верхняя граница: согласие экспертов
  - 75-80 для задачи определения значений всех слов со значениями из WordNet
  - до 90% с менее гранулированными значениями

Наиболее частое значение

- Сравнение методов на SENSEVAL-2



- McCarthy et. al. 2004 ACL - поиск наиболее частого значения по неразмеченному корпусу

Методы основанные на словорях и тезаурусах

- Алгоритм Леска (1986)
  - Взять все определения целевого слова из словаря
  - Сравнить с определениями слов в контексте
  - Выбрать значение с максимальным пересечением
- Пример
  - *pine*
    1. a kind of evergreen tree with needle-shaped leaves
    2. to waste away through sorrow or illness
  - *cone*
    1. A solid body which narrows to a point
    2. Something of this shape, whether solid or hollow
    3. Fruit of certain evergreen trees
  - Определить значение: *pine cone*

Варианты алгоритма Леска

- Упрощенный (Simplified Lesk)
  - Взять все определения целевого слова из словаря
  - Сравнить со ~~определениями~~ словами в контексте
  - Выбрать значение с максимальным пересечением
- Корпусный (Corpus Lesk)
  - Включить предложения из размеченного корпуса в сигнатуру каждого значения
  - Взвесить слова через IDF
  - IDF(w) = -log P(w)
  - Показывает лучшие результаты
  - Использовался как нижняя граница на SENSEVAL

Самонастройка (Bootstrapping)

- Yarowsky (1995)
  - Начать с маленького множества данных, размеченного вручную
  - Натренировать список принятия решений
  - Применить классификатор к неразмеченным данным
  - Переместить примеры в которых мы уверены в тренировочное множество
  - Повторить!
- Требует хорошей метрики уверенности
  - логарифмическое отношение правдоподобия
- Эвристики для получения начальных данных
  - одно значение на словосочетание
  - одно значение на дискурс

Алгоритм Yarowsky

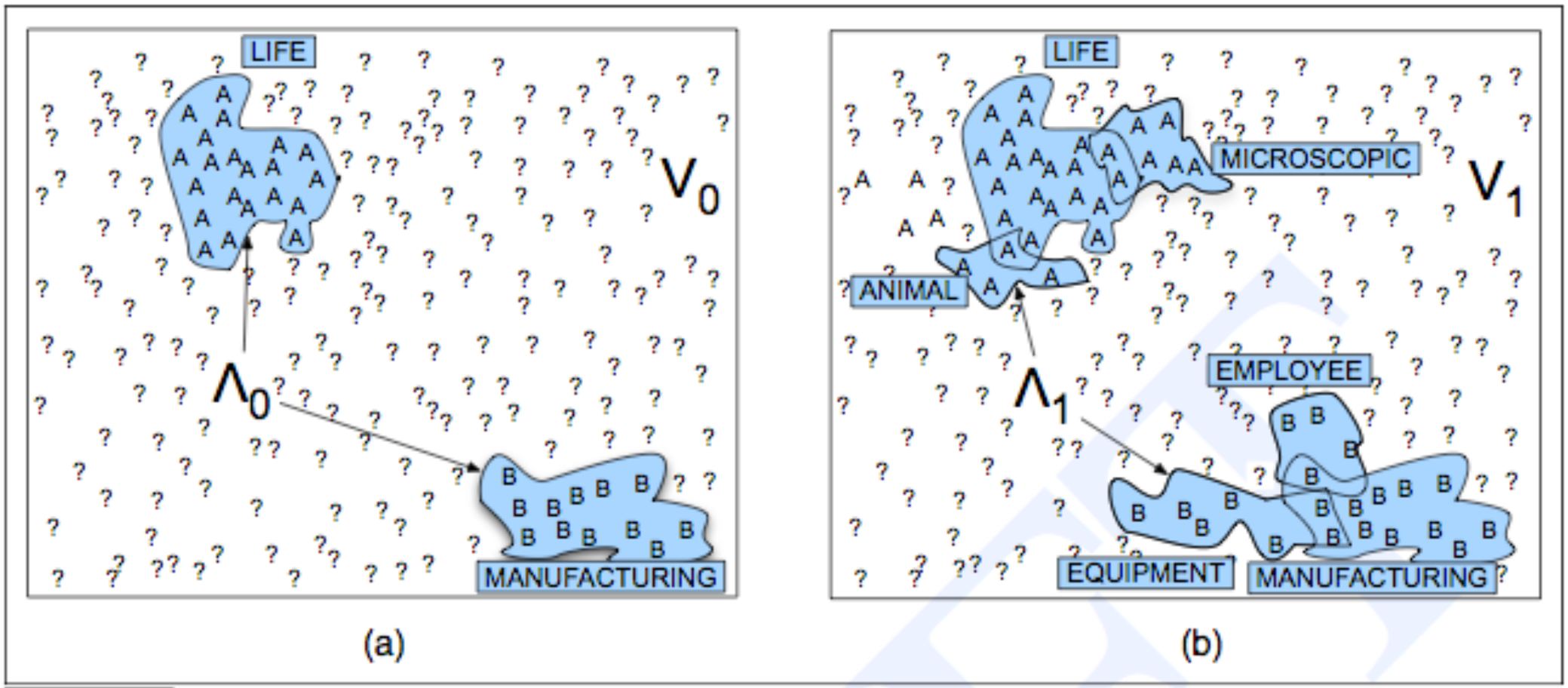


Figure 20.4 The Yarowsky algorithm disambiguating "plant" at two stages; "?" indicates an unlabeled observation, A and B are observations labeled as SENSE-A or SENSE-B. The initial stage (a) shows only seed sentences Λ_0 labeled by collocates ("life" and "manufacturing"). An intermediate stage is shown in (b) where more collocates have been discovered ("equipment", "microscopic", etc.) and more instances in V_0 have been moved into Λ_1, leaving a smaller unlabeled set V_1. Figure adapted from Yarowsky (1995).

Семантическая близость слов

- Подходы на основе тезаурусов
- Подходы на основе статистики

Мотивация

- Хороший признак для многих задач
- Позволяет бороться с разреженностью языка
- Имеет прикладное применение
  – поиск опечаток (с учетом семантики)
  – поиск плагиата
  – извлечение информации

Подход на основе тезаурусов

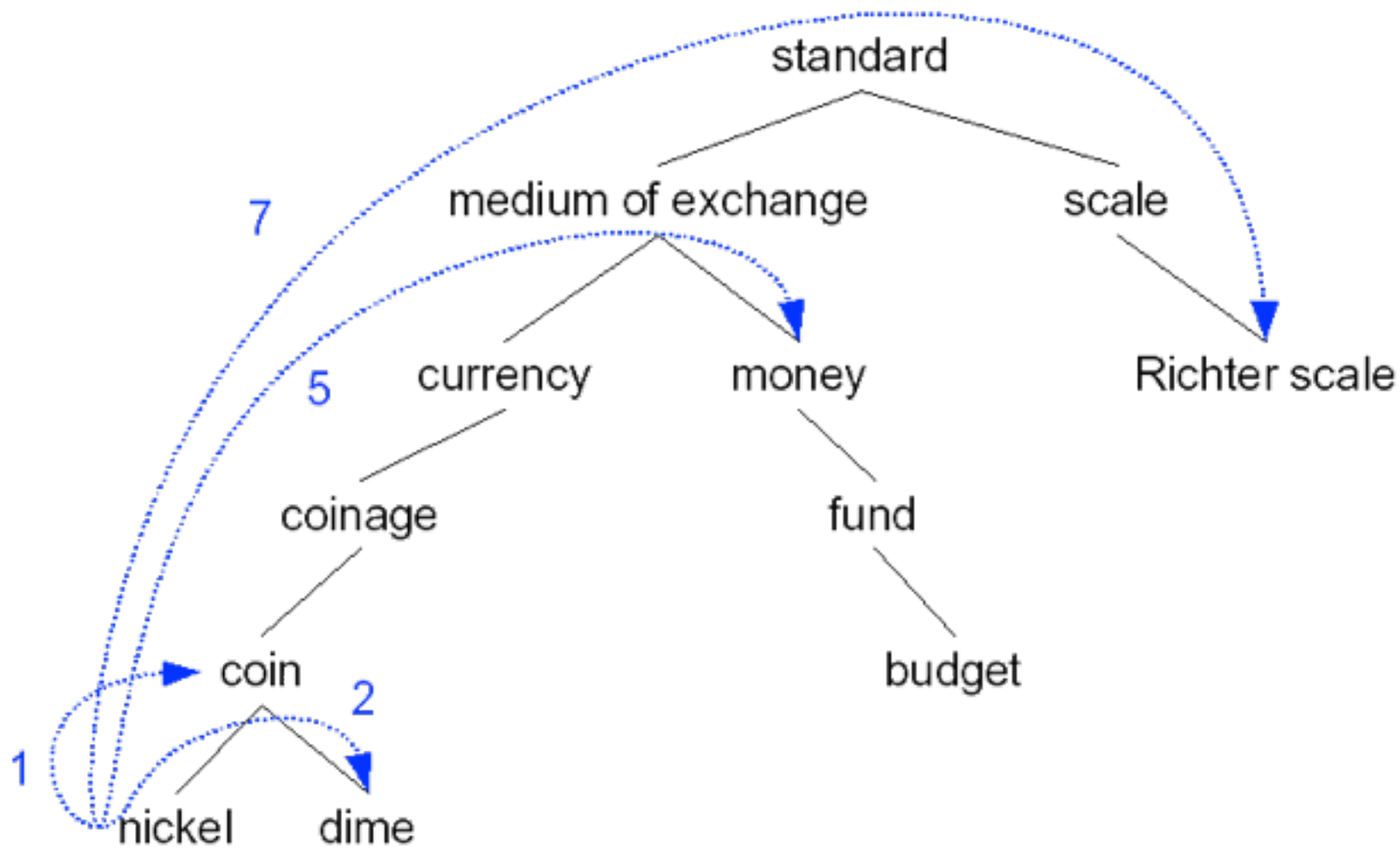
- Близость по пути
- Метод Резника
- Метод Лина
- Расширенный алгоритм Леска

Семантическая близость слов в тезаурусах

- Можно использовать любые отношения между словами
- На практике используется иерархическая структура и иногда описания значений
- Похожесть (similarity) VS связность (relatedness)
  - машина и топливо: не похожи но связаны
  - машина и велосипед: похожи

Близость по пути в иерархии

- Два понятия семантически близки, если они находятся рядом в иерархии



Близость между словами

- Только что мы посчитали близость между понятиями
- Перейдем ко словам

- simpath(c1,c2) = -log (pathlen(c1,c2))
- wordsim(w1,w2) = $\max_{c1 \in senses(w1), c2 \in senses(w2)}$ sim(c1,c2)

Другие методы

- Сначала немного определений...
  - Информационное содержимое
  - Наименьший общий предок

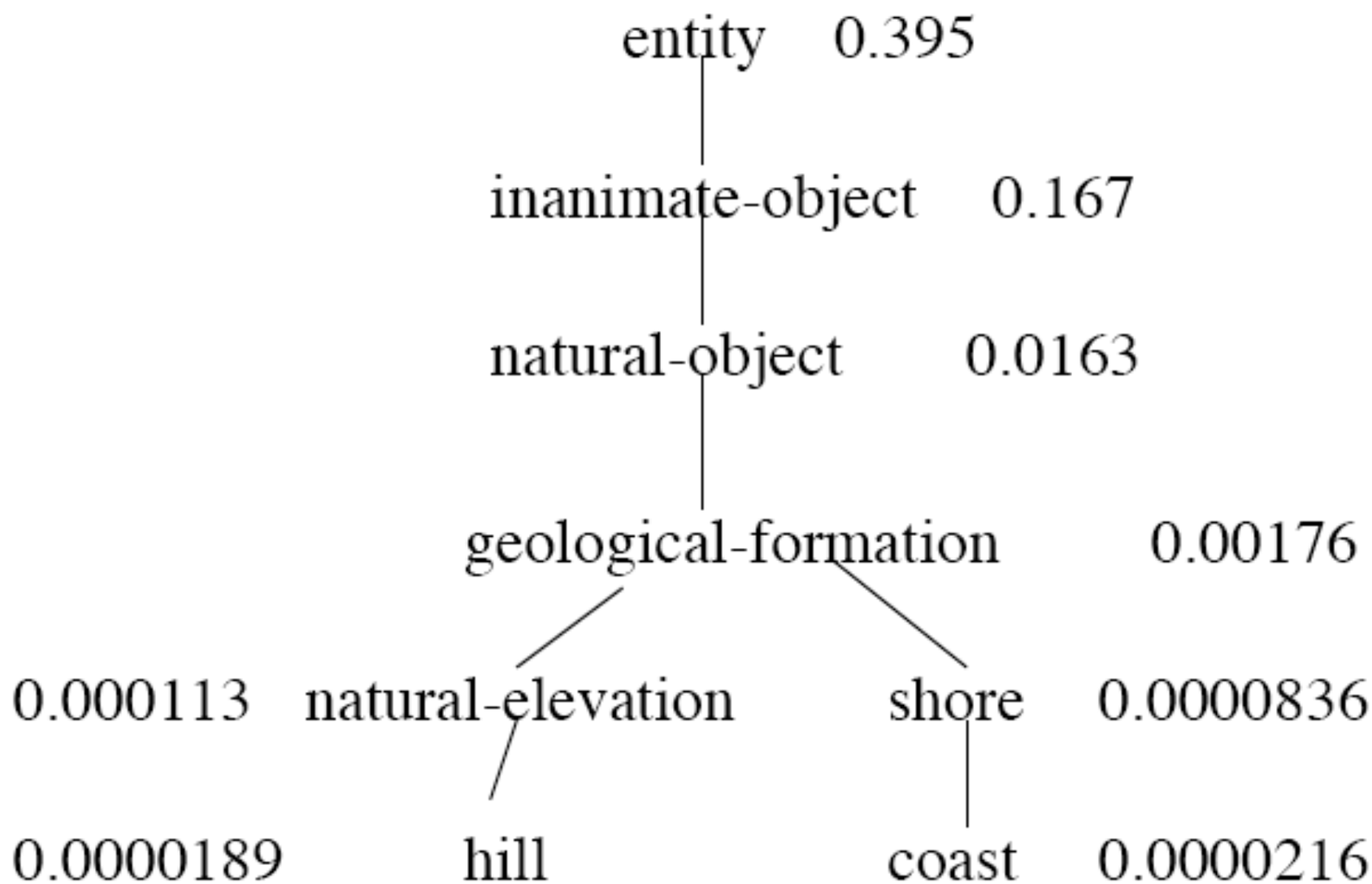
Вероятность класса

- Определим P(C) как:
  - Вероятность, что случайно выбранное слово в корпусе является экземпляром класса C
  - P(root)=1
  - Чем ниже узел в иерархии, тем ниже вероятность

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

Информационное содержимое

- Расширяем иерархию WordNet вероятностями P(C)



Определения

- Информационное содержимое
  - IC(c)=-log(P(c))
- Наименьший общий предок
  - LCS (c1,c2)

Метод Резника

- Resnik (1995)
  - Чем больше общего между понятиями, тем более они похожи
  - $sim_{resnik}(c1,c2) = IC(LCS(c1,c2)) =$

$$= -\log P(LCS(c1,c2))$$

Метод Лина

- Dekang Lin (1998)
  - При вычислении близости также надо учитывать различие между понятиями
- Идея может быть выражена как

$$sim_{Lin}(c_1, c_2) = \frac{2 \times log(P(LCS(c_1, c_2)))}{log(P(c_1)) + log(P(c_2))}$$

$$sim_{Lin}(hill, coast) = \frac{2 \times log(P(geological\_information))}{log(P(hill)) + log(P(coast))} = 0.59$$

Расширенный алгоритм Леска

- Два концепта похожи, если их описания содержат похожие слова
  - *Drawing paper*: **paper** that is **specially prepared** for use in drafting
  - *Decal*: the art of transferring designs from **specially prepared paper** to a wood or glass or metal surface
- Каждому общему словосочетанию длины n назначить вес n^2
- paper + specially prepared: 1+4 = 5

Резюме: методы, основанные на тезаурусах

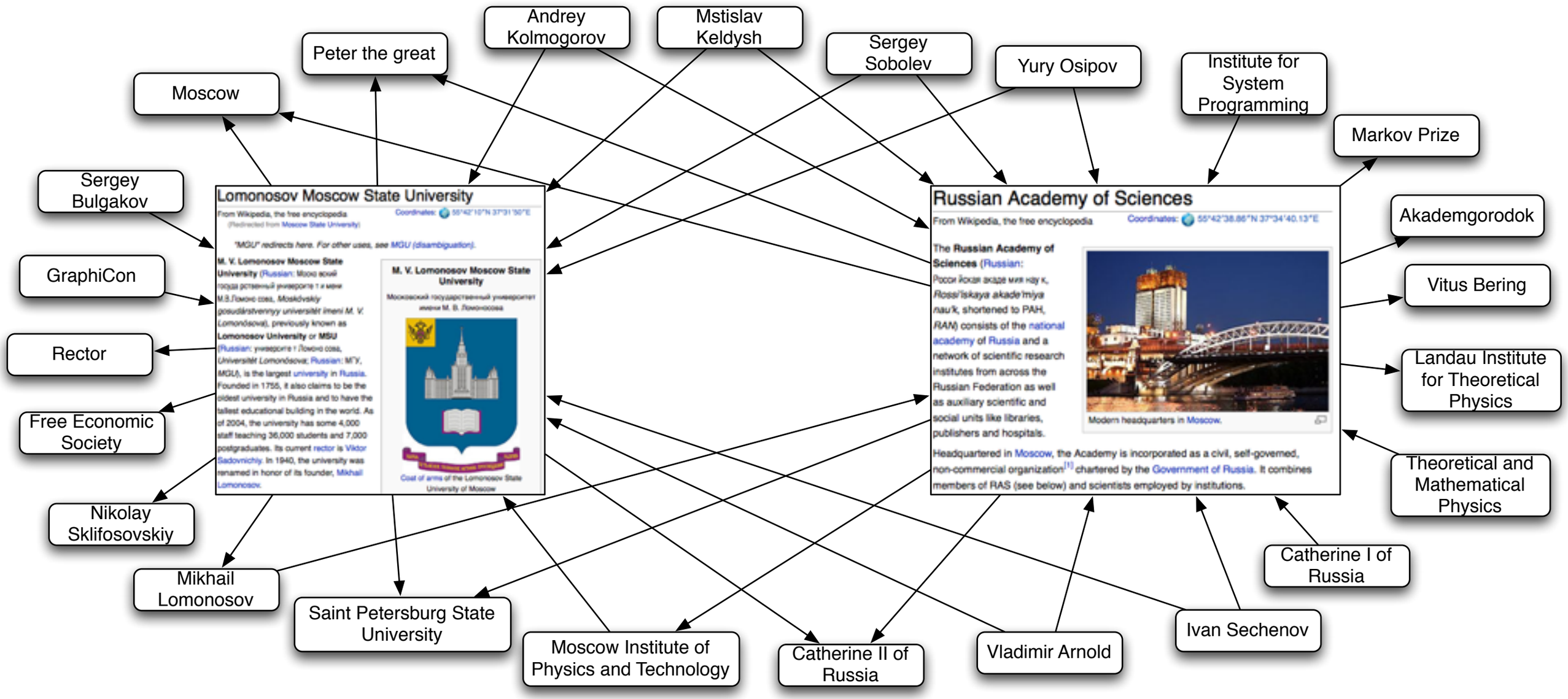
$$
\begin{aligned}
\mathrm{sim}_{\mathrm{path}}(c_1, c_2) &= -\log \mathrm{pathlen}(c_1, c_2) \\
\mathrm{sim}_{\mathrm{Resnik}}(c_1, c_2) &= -\log P(\mathrm{LCS}(c_1, c_2)) \\
\mathrm{sim}_{\mathrm{Lin}}(c_1, c_2) &= \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \\
\mathrm{sim}_{\mathrm{jc}}(c_1, c_2) &= \frac{1}{2 \times \log P(\mathrm{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))} \\
\mathrm{sim}_{\mathrm{eLesk}}(c_1, c_2) &= \sum_{r,q \in \mathrm{RELS}} \mathrm{overlap}(\mathrm{gloss}(r(c_1)), \mathrm{gloss}(q(c_2)))
\end{aligned}
$$

Проблемы с подходом, основанном на тезаурусе

- Не доступен для многих языков
- Много слов пропущено
- Используются только обобщения и детализация
  - Хорошо работает для имен существительных
  - Для прилагательных и глаголов намного хуже

Использование Википедии

- Нормализованное количество общих соседей



- Близкие концепты чаще встречаются вместе

Статистический подход к оценки близости слов

- Firth (1957): “You shall know a word by the company it keeps!”
- Пример

Бутылка **tezgüino** стоит на столе

Все любят **tezgüino**

Tezgüino делает тебя пьяным

Мы делаем **tezgüino** из кукурузы

- Идея:
  – из контекста можно понять значение слова
  – надо взять контекст и посмотреть, какие еще слова имеют такой же контекст

Векторное представление контекста

- Для каждого слова из словаря определим бинарный признак, показывающий встречаемость вместе с целевым словом **w**
- $w=(f_1,f_2,f_3,...,f_N)$
- w= tezgüino, v1=бутылка, v2=кукуруза, v3=матрица
- w = (1,1,0,...)

Идея

- Задать два слова через разреженный вектор признаков
- Применить метрику близости векторов
- Два слова близки, если векторы близки

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Статистический подход к оценки близости слов

- Необходимо определить 3 вещи:
  - контекст (совместная встречаемость с другими терминами)
  - вклад термина в близость (вес термина)
  - близость между векторами

Совместная встречаемость

- Проблема разреженности
  - Нужны большие корпуса

Вес термина

- Manning and Schuetze (1999)

$$
\begin{aligned}
\text{assoc}_{\text{prob}}(w,f) &= P(f|w) \\
\text{assoc}_{\text{PMI}}(w,f) &= \log_2 \frac{P(w,f)}{P(w)P(f)} \\
\text{assoc}_{\text{Lin}}(w,f) &= \log_2 \frac{P(w,f)}{P(w)P(r|w)P(w'|w)} \\
\text{assoc}_{\text{t-test}}(w,f) &= \frac{P(w,f)-P(w)P(f)}{\sqrt{P(f)P(w)}}
\end{aligned}
$$

Близость между векторами

$$\mathrm{sim}_{\mathrm{cosine}}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} \max(v_i,w_i)}$$

$$\mathrm{sim}_{\mathrm{Dice}}(\vec{v},\vec{w}) = \frac{2\times\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} (v_i+w_i)}$$

$$\mathrm{sim}_{\mathrm{JS}}(\vec{v}||\vec{w}) = D(\vec{v}|\frac{\vec{v}+\vec{w}}{2}) + D(\vec{w}|\frac{\vec{v}+\vec{w}}{2})$$

Современное направление

- Использование нейронных сетей для получения векторного представления слов
  - word2vec, GloVe, fasttext
  - http://code.google.com/p/word2vec/
- Близость к слову france ->

Word	Cosine distance
spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291
catalonia	0.534176

- Задача поиска аналогий
  - v(king)-v(man)+v(woman)=?
- Gensim: реализация на Python

Оценка качества

- Внутренняя
  - Коэффициент корреляции между
    - результатами алгоритма и
    - значениями, поставленными людьми
- Внешняя
  - Встроить в приложение
    - Поиск опечаток
    - Поиск плагиата
    - Разрешение лексической многозначности

Заключение

- **Лексическая семантика** изучает значения отдельных слов
- **WordNet** содержит различные отношения между словами, синсеты задают **значения** слов
- **Разрешение лексической многозначности** - задача определения значений слов
- **Семантическая близость** между словами - полезный инструмент для монгих приложений

Что не было рассказано

- Композиционная семантика
- Представление знаний
- Семантические поля и семантические роли
  - PropBank
  - FrameNet
- Задача разграничения значений
- Автоматическое извлечение отношений между словами
- ...

Следующая лекция

- Информационный поиск
- Вопросно-ответные системы
- Автоматическое реферирование