Методы автоматического разрешения лексической многозначности

Семантический анализ текста

- Построение семантической интерпретации слов и конструкций
  - Разрешение многозначности слов
- Установление семантических отношений между элементами текста
  - Словари и правила
  - Машинное обучение по размеченным данным

Разрешение лексической многозначности

- Разрешение лексической многозначности – выбор значения слова из набора значений, описанных в том или ином источнике

- Кластеризация (дискриминация) значений слова – разделение употреблений слова на группы, соответствующие нескольким значениям безотносительно к предопределенному набору значений

- Конференция Senseval – 1,2,3
- 2007 конференция SemEval

Основные классы методов

- Методы, основанные на ручных правилах
- Методы, основанные на лингвистических ресурсах
  - Машинные словари (метод Леска и модификации)
  - Тезаурусы
  - Модели управления
- Методы, основанные на обучении по размеченному корпусу – обучение на примерах (обучение с учителем)
- Методы, основанные на неразмеченных корпусах, использование интернета
- Гибридные методы

История области

* **40е – зарождение машинного перевода**

– Warren Weaver, «The "Translation" memorandum» (1949)

– Yehoshua Bar-Hillel, скептик

•**70е – WSD – часть более крупных проектов**

– в основном, ручные правила

* **80е – появление электронных словарей**

– Oxford Advanced Learner's Dictionary of Current English,

– ручное выписывание правил – вытеснено автоматическим извлечением знаний из подобных источников

История области-2

- **90е** – "статистическая революция", обучение с учителем

- **00е** – смещение в сторону:
- → coarse-grained senses
- → domain adaptation
- → semi-supervised system и обучения без учителя
- → смешанные методы, обработка баз знаний

Тестирование автоматического разрешения многозначности

- Корпус текстов или предложений
  - Размечается экспертами правильными значениями – семантически размеченный корпус
- Оценки
  - Точность: число слов, размеченных правильно, по отношению к числу слов, обработанных системой
  - Полнота: число слов, размеченных правильно, по отношению к числу слов в тестовом множестве

Трудности задачи

- **Differency of dictionaries**
  - все словари разные и не эквивалентны друг другу
- **Part-of-speech tagging**
  - в некоторых языках проблема определения части речи слова (part-of-speech tagging) может быть очень близко связана с проблемой разрешения многозначностей
- **Inter-judge variance** -человеческий фактор
  - Системы разрешения лексической многозначности всегда оценивались сравнением результатов с результатом работы людей. А людям данная задача может оказаться не такой простой, как POS-tagging

Трудности задачи-2

- **Sense inventory** and **algorithms' task-dependency:**
  - для разных задач требуются и разные алгоритмы
  - для алгоритма разрешения лексической многозначности невозможно быть полностью уверенным, что он подойдёт под решение всех задач
- **Discreteness of senses**
  - Значения слов очень гибки, контекстно зависимы
  - не всегда строго делятся на несколько подзначений

Простые методы

Использование машинных словарей: алгоритм Леска

- (Michael Lesk 1986): пересечение контекста употребления слова с его словарным толкованием
- Выбираются значения слов, толкования которых имеют больше пересечений между собой

- Классический пример: PINE CONE
- PINE
  1. kinds of evergreen tree with needle-shaped leaves
  2. waste away through sorrow or illness
- CONE
  1. solid body which narrows to a point
  2. something of this shape whether solid or hollow
  3. fruit of certain evergreen trees

  Pine#1 ∩ Cone#3 = 2

Алгоритм Леска для текста

- **(Kilgarriff & Rosensweig 2000):**
  - измеряется пересечение между толкованиями и контекстом слова
  - Могут использоваться примеры из слов. статей
  - На тех же основаниях может использоваться размеченный корпус
- *Pine cones hanging in a tree*
- *Pine#1 ∩ Sentence = 1; Pine#2 ∩ Sentence = 0*
- Senseval-1: Метод Леска:
  - по определениям – ок. 0.3
  - по определениям и примерам – ок. 0.55
  - по определениям, примерам и корпусу – ок. 0.68

Учет наиболее частотного значения

- Определение наиболее частотного значения
  - Использование этого значения, если не удалось определить другим методом
  - Значение максимальной частотности по размеченному корпусу
  - Семантически размеченный корпус SemCor - размечен по значениям WordNet
- SemCor: употребление в наиболее частотном значении:
  - существительные – 85%
  - прилагательные – 45%
  - глаголы – 48%

Как определить наиболее частотное значение без размеченного корпуса

- (McCarthy et al. 2004) ACL 2004 Best Paper
- Синтаксический анализ корпуса, извлечение троек (R, W1,W2)
- Для каждого W можно определить список наиболее похожих по синтаксическому поведению слов {W_i}, i=1,k с некоторыми весами: star=(superstar, player, teammate, actor, galaxy, sun, planet, …)
- Для каждого w_i определяется близость к одному из значений w (метод Леска, по структуре WordNet)
- Для каждого значения насчитывается сумма
- Исходный вес* коэффициент близости
- Результат : 54% угаданных частотных значений из 2595 SemCor

Одно значение на документ

- **Гипотеза: большинство слов документа употребляются в** одном и том же значении
- Проверка по размеченному корпусу
- 8 слов с 2 основными значениями: plant, crane
- 98% вхождение слов имеют то же значение
- (Krovetz 1998)
- Точность разрешения многозначности, основанная на принципе одно значение на документ - 70% на корпусе SemCor

Одно значение на словосочетание

- Предположение: слово, чаще всего, сохраняет свое значение в словосочетании
- *Plant – industrial plant*
- 97% - слов с двумя значениями

- (Martinez and Agirre 2000)
  - значения WordNet
  - SemCor - 70%

Использование структуры тезауруса

Смотрим, насколько близки слова из контекста по тезаурусу значениям многозначного слова

Вычисление семантической близости – учет пути

- Вход: два понятия. Результат: мера близости
- (Leacock and Chodorow 1998)
- Similarity (C1, C2)= - log (path (C1,C2)/2D),
- D – глубина таксономии
- Similarity(wolf, dog) = 0.60
  Similarity(wolf, bear) = 0.42

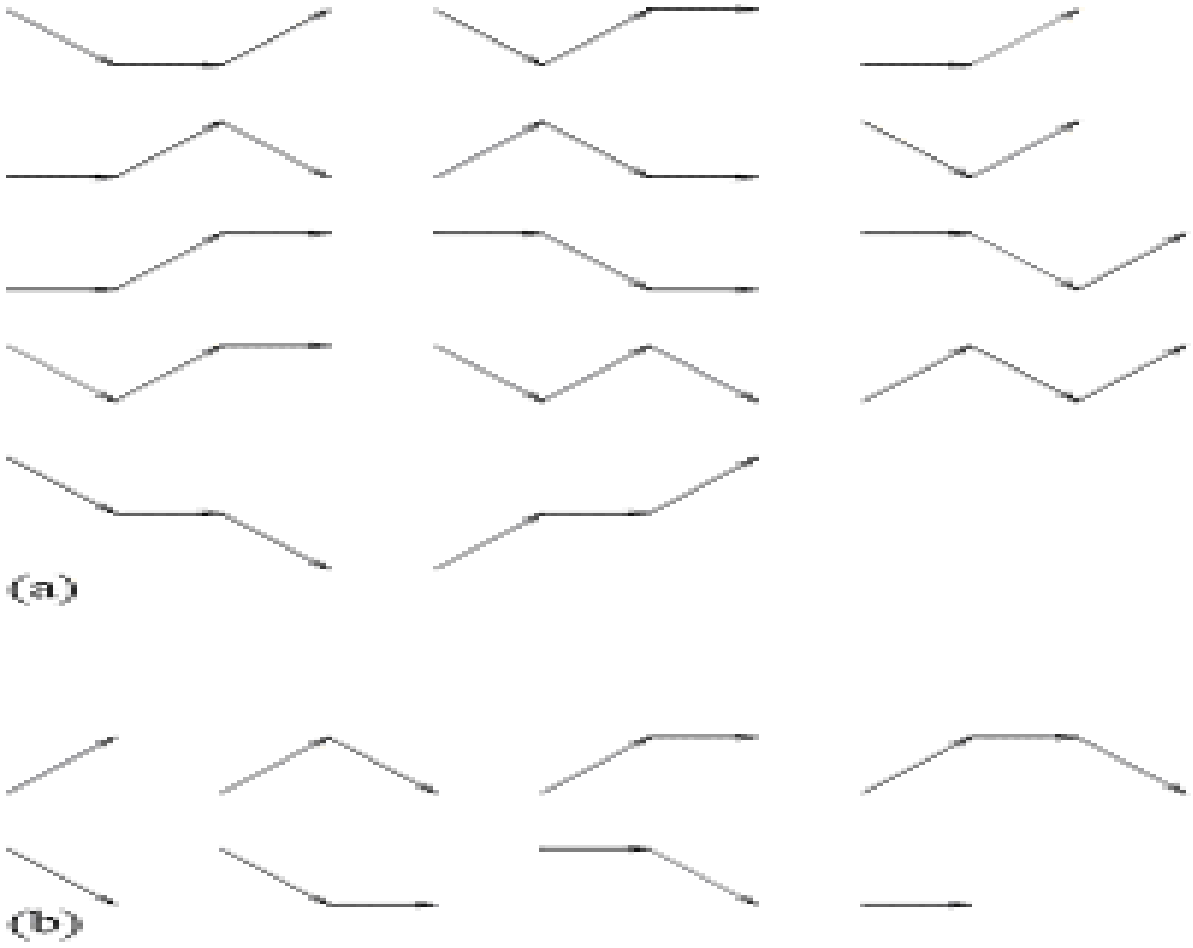
Вычисление семантической близости – учет пути-2

- (Hirst and St.Onge, 1998)
  - Вес пути = C – длина_пути – k* (число поворотов пути)
  - C=8, K=1
  - Максимальный рассматриваемый путь – 5 шагов

(Hirst, St.Onge):

а) неразрешенные пути

б) разрешенные пути



Вычисление семантической близости

- (Resnik 1995): Информационное содержание
  - P(c) – вероятность нахождения понятия C в большом корпусе
  - если C1 вид для C2, то P(C1) =< P(C2); P(Top)=1
  - Информационное содержание: IC (C) = -log P(C)
- Чем более абстрактным является понятие, тем меньше величина его информационного содержания.
- LCS – наименьший родовой концепт

  Similarity (C1, C2) = IC (LCS (C1, C2)
- (Jiang and Conrath 1997)

  Similarity (C1, C2) = 2*IC (LCS (C1,C2)) –
  - (IC(C1)+IC(C2))

Эксперимент

- Разрешение многозначности слов на основе ближайшего соседа (для существительных)
  - (Patwardhan, Banerjee, Pedersen 2002)
  - Сравнение 5 метрик сходства - WordNet
- “Plant with flowers”
  - Plant, industrial plant
  - Plant, flora
- Similarity (plant, flower)
- 1723 вхождения многозначных существительных Senseval – 2
- (Jiang and Conrath 1997) – 39%

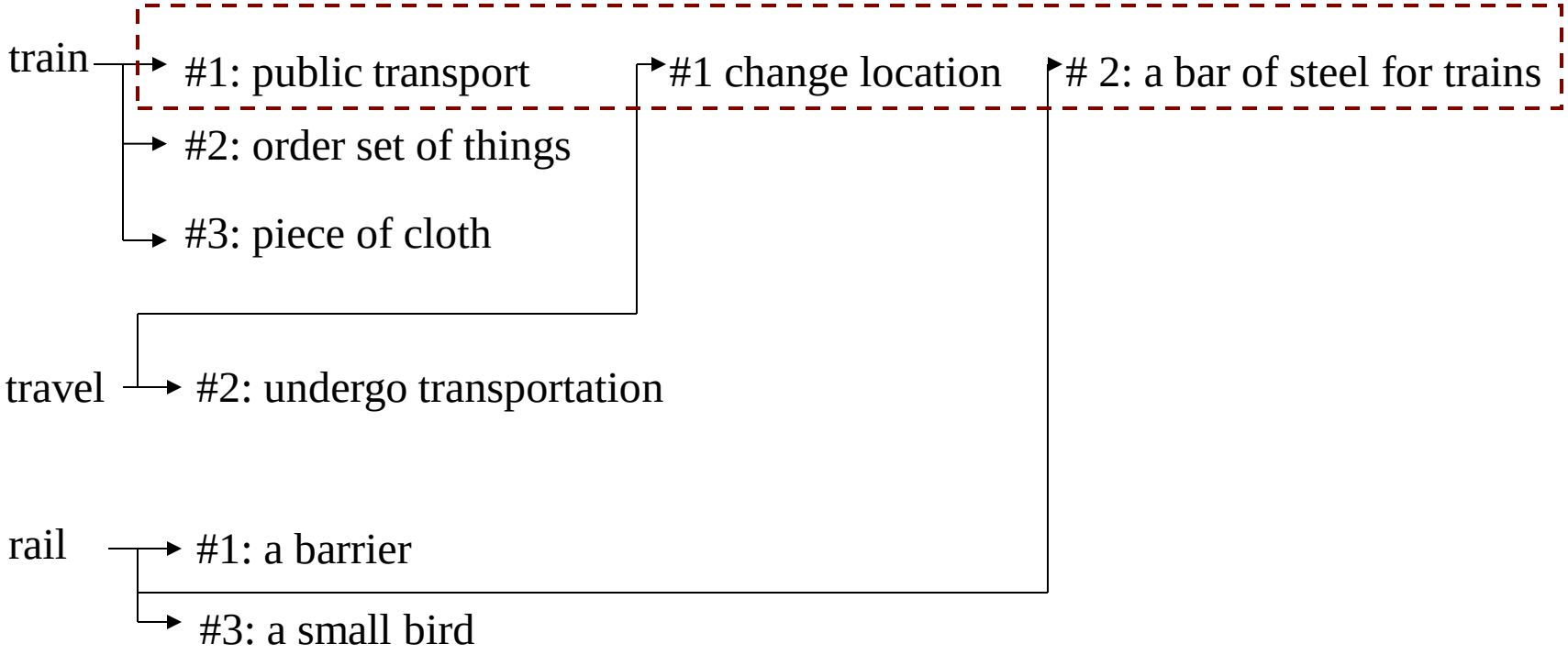
Учет глобального контекста

Лексические цепочки

- (Hirst and St-Onge 1988), (Haliday and Hassan 1976)
- Лексические цепочки – последовательность близких по смыслу слов, в которых проявляется лексическая связность связного текста
- Алгоритм создания лексических цепочек
  1. Извлечение слов из текста, между которыми может быть определена мера семантической близости
  2. Двигаясь сначала текста, для каждого очередного слова просматриваются имеющиеся цепочки
  3. Если есть цепочка, то слово присоединяется
  4. Если несколько, то цепочка, в которой близость больше
  5. Обычно есть ограничения на расстояние (предложение, абзац) до последнего слова цепочки

Пример: лексические цепочки и разрешение многозначности

A very long **train traveling** along the **rails** with a constant **velocity** v in a certain **direction** …



Лексические цепочки для разрешения многозначности: Оценки

- Если слово многозначное, то выбирается значение, на основе которого произошло присоединение к цепочке
- (Galley and McKeown 2003):
  74 текстов SemCor, 35000 сущ., – 62.09
- Этапы обработки:
  - сопоставление с WordNet, отмечаются все возможные значения
  - находятся отношения между значениями – синонимы, гипонимы, гиперонимы, понятия- «сестры»
- Предположение: одно значение на документ
- Подсчет весов, полученных по всему документу: зависимость от типа связи, расстояние
  (1 предложение, 3 предложения, абзац)

Обучение на примерах с учителем (supervised learning)

Обучение с учителем

- Набор совокупности примеров, которые иллюстрируют различные возможные классификации
- Идентификация образцов, соответствующих каждому классу
- Обобщение образцов в правила
- Применение правила для классификации нового примера

Разрешение многозначности на основе обучения с учителем

- Ресурсы
  - Размеченный корпус
  - Набор значений словаря
  - Синтаксический анализ
- Результат
  - Обычно одно целевое слово
- Задача разрешения лексической многозначности как задача автоматической классификации по заданному набору классов (=значений)

Примеры, размеченные по значениям

Bonnie and Clyde are two really famous criminals, I think they were **bank/1** robbers
My **bank/1** charges too much for an overdraft.
I went to the **bank/1** to deposit my check and get a new ATM card.
The University of Minnesota has an East and a West **Bank/2** campus right on the Mississippi River.
My grandfather planted his pole in the **bank/2** and got a great big catfish!
The **bank/2** is pretty muddy, I can’t walk there.

Контексты слова *bank*

FINANCIAL_BANK_BAG:

a an and are ATM Bonnie card charges check Clyde criminals deposit famous for get I much My new overdraft really robbers the they think to too two went were

RIVER_BANK_BAG:

a an and big campus cant catfish East got grandfather great has his I in is Minnesota Mississippi muddy My of on planted pole pretty right River The the there University walk West

Примитивный алгоритм: проверять вхождение слова в списки, добавлять 1 к счетчику, выбирать наибольшее

Подходы, основанные на машинном обучении

- Создание обучающей выборки, в которой целевое слово вручную размечено значениями из списка
  - Одно размеченное слово на пример
- Признаки для представления контекста
  - Соседние слова, коллокации, части речи, синтаксические отношения и др
- Применяемые методы машинного обучения
  - Метод опорных векторов, KNN
  - Деревья решений, Списки решений (Decision Lists)
  - Наивный байесовский классификатор
  - Персептроны, Нейронные сети

От текста к векторам признаков

- My/pronoun grandfather/noun used/verb to/prep fish/verb along/adv the/det **banks/SHORE** of/prep the/det Mississippi/noun River/noun. (S1)
- The/det **bank/FINANCE** issued/verb a/det check/noun for/prep the/det amount/noun of/prep interest/noun. (S2)

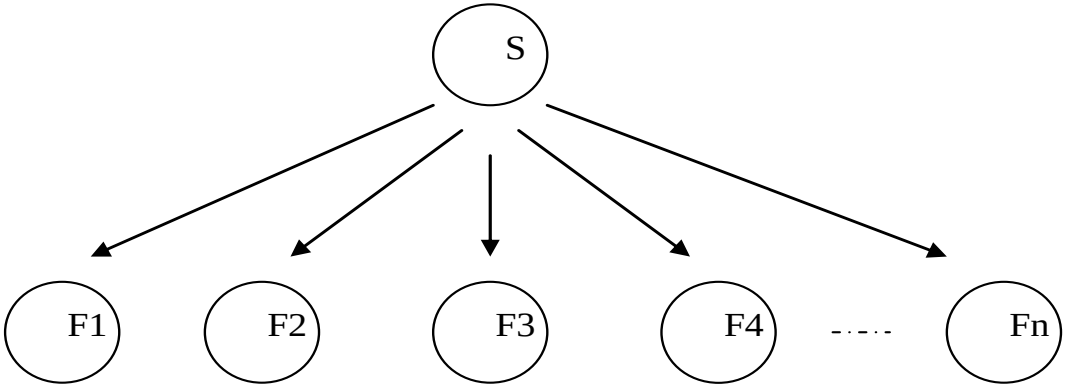
	P-2	P-1	P+1	P+2	fish	check	river	interest	SENSE TAG
S1	adv	det	prep	det	Y	N	Y	N	SHORE
S2		det	verb	det	N	Y	N	Y	FINANCE

Байесовский классификатор

$$p(S \mid F1, F2, F3, \ldots, Fn) = \frac{p(F1, F2, F3, \ldots, Fn \mid S) * p(S)}{p(F1, F2, F3, \ldots, Fn)}$$

- Оценка вероятности встретить значение – p (S)
- Оценка вероятности встречаемости признаков при условии заданного значения
- Знаменатель не влияет на результат

Наивная байесовская модель



$$P(F1, F2, ..., Fn \mid S) = p(F1 \mid S) * p(F2 \mid S) * ... * p(Fn \mid S)$$

Decision Lists and Trees

- Представляют проблему разрешения многозначности как серию вопросов, на которые нужно ответить
  - Список выбирает между двумя значениями после одного позитивного ответа
  - деревья позволяют сделать выбор между несколькими значениями после серии ответов
- Обычно меньший набор признаков, чем при мешке слов или методе Байеса
  - Легче интерпретировать

Список решений для выбора значений (Yarowsky, 1994)

- Используется встречаемость слов в контексте
- Слова непосредственно слева или справа:
  - I have my bank/1 *statement*.
  - The *river* bank/2 is muddy.

- Или слова, найденные в k позициях слева или справа (k=10-50):
  - My *credit* is just horrible because my bank/1 has made several mistakes with my *account* and the *balance* is very low.

Построение списка решений

- Сортировка контекстных слов на основе логарифма условных вероятностей
- Слова, которые наиболее связаны с одним из значний, и не встречающиеся с другим, получают высокий ранг

$$Abs(\log\frac{p(S=1|F_i=Collocation_i)}{p(S=2|F_i=Collocation_i)})$$

Вычисление DL ранга

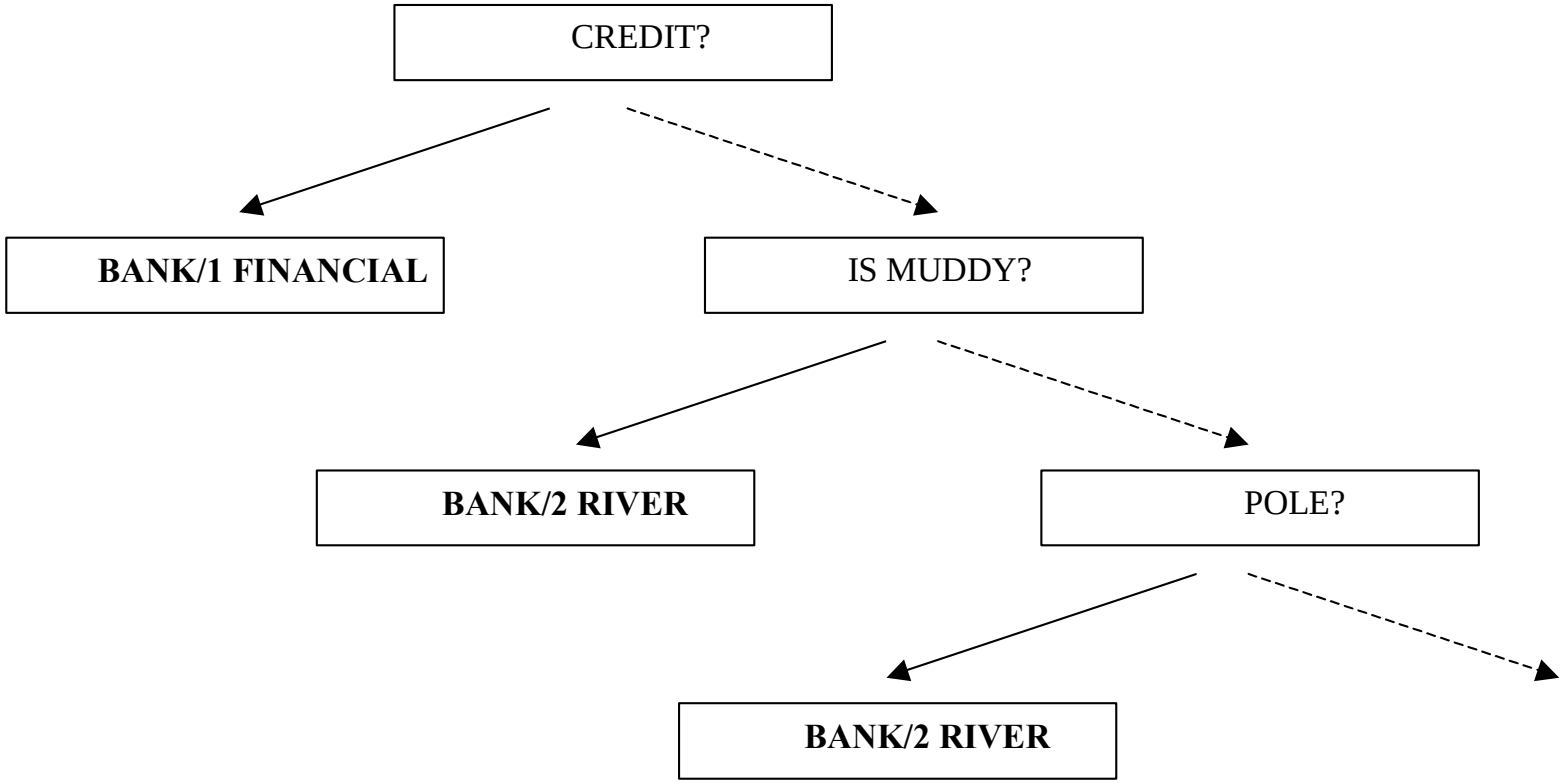
- 2,000 примеров “bank”, 1,500 - bank/1 (финансы) и 500 r bank/2 (река)
  - P(S=1) = 1,500/2,000 = .75
  - P(S=2) = 500/2,000 = .25
- “credit” встречается 200 раз с bank/1 and 4 раза с bank/2.
  - P(F1=“credit”) = 204/2,000 = .102
  - P(F1=“credit”|S=1) = 200/1,500 = .133
  - P(F1=“credit”|S=2) = 4/500 = .008
- Правило Байеса…
  - P(S=1|F1=“credit”) = .133*.75/.102 = .978
  - P(S=2|F1=“credit”) = .008*.25/.102 = .020
- DL Score = abs (log (.978/.020)) = 3.89

Использование списка решений

- Сортировка по DL рангу, первое совпадение проставляет значение

DL-score	Feature	Sense
3.89	*credit* within bank	Bank/1 financial
2.20	bank *is muddy*	Bank/2 river
1.09	*pole* within bank	Bank/2 river
0.00	*of the* bank	N/A

Применение списка решений



Минимизация обучающего множества

Основные принципы

- Имеется
  - Некоторый объем размеченных данных
  - Большие объемы неразмеченных данных
  - Один или больше базовых классификаторов
- Результат
  - Классификатор, улучшающий работу базового классификатора

Базовый алгоритм

- Дано:
  - Множество L – размеченных примеров
  - Множество U – неразмеченных примеров
  - Классификаторы
- Основные шаги:
  - Создать множество примеров U1 – подмножество U (P – число примеров U1)
  - Цикл I итераций
  - Натренировать классификаторы на множестве L
  - Применить для разметки примеров в U1
  - Выбрать наиболее «надежные» примеры (G)и добавить в L
  - Дополнить U1 примерами из U (P=const)
- Проблемы:
  - Непонятно, какие принципы выбора параметров: P, G, I

Эксперименты

- Тестовые данные
  - существительные Senseval-2 – 29
  - Размер размеченного корпуса: 95 тренировочных примеров,48 тестовых примеров
- Неразмеченные данные
  - Британский национальный корпус
  - Средний размер множества примеров: 7085
- Прогоны на наборах параметров
  - P = {1, 100, 500, 1000, 1500, 2000, 5000}
  - G = {1, 10, 20, 30, 40, 50, 100, 150, 200}
  - I = {1, ..., 40}

  Результаты: исходный классификатор: 53.84%, результирующий классификатор – 65.61

Интернет-как корпус

- Построить размеченный корпус, пользуясь однозначными близкими по смыслу конструкциями
- Однозначные синонимы
- Фрагмент определения: produce#5- bring onto the market
- Выражение, порожденное из определения,
- produce#6 - синсет {grow, raise, farm, produce}, толкование “cultivate by growing”

SP = cultivate NEAR growing AND
(grow OR raise OR farm OR produce)

Конференция Senseval

Тестирование систем: Senseval

- Задачи
  - Разрешение многозначности для набора слов (40)
  - Разрешение многозначности для всех слов текста
- Оценки
  - Точность: число слов, размеченных правильно, по отношению к числу слов, обработанных системой
  - Полнота: число слов, размеченных правильно, по отношению к числу слов в тестовом множестве

Пример:

В тестовом множестве –100 слов

Система работала с 75 словами

Правильно – 50 слов

Точность = 50/75 = 0.66;

Полнота = 50/100 = 0.50

Особенности тестирования: набор слов

- Решетка: часть речи, количество значений, частотность - generous, onion
- Выдаются размеченные примеры - 1, 2 предложения
- Тренировочное задание, основное задание
- Уровни гранулярности:
- Подробный, обобщенный, смешанный
- Результаты Senseval-3:
  - 72% подробный уровень
  - 79% обобщенный уровень
  - Выбор самого частотного значения: 55.2 для подробного уровня, 64,5 обобщенный уровень

Особенности тестирования: все слова текста. Senseval-3

- 2 статьи Wall Street Journal и фрагмент Брауновского корпуса
- 2081 слов - для тестирования
- Разметка по набору значений WordNet.
- Согласие между аннотаторами – 72.5%
  - Особые метки: нет значения, несколько значений
- U – нет значения.
  - Система должна также выдавать U
  - Максимальная точность – 65.2, средняя точность по системам –52.2
  - Базовый уровень выбор первого значений WordNet – 60.9

Тестирование в приложениях. Информационный поиск

- SemEval (ACL 2007)
- Заданы: система поиска, система перевода /расширения запроса
- Участники должны выбрать наилучшую стратегию разрешения многозначности
- Языки запроса: английский, испанский
- Язык документа: английский
- Подзадания:
  - разрешить многозначность в корпусе, расширить синонимами и переводами, измерить эффективность поиска
  - Разрешить многозначность запросов, расширить синонимами и переводами, измерить эффективность поиска
  - 250 запросов конференции CLEF

Задание

- Написать метод разрешения многозначности (Kilgariff, 2000) – пересечение предложения с толкованиями
  - Ввод значений слова и лемматизация
  - Ввод предложений с многозначным словом и лемматизация
  - Вычисление пересечения лемм предложения и лемм каждого из значений
  - Выдача лучшего значения,
  - Тестирование на 5 многозначных сущ., отчет