Методы разрешения лексической многозначности, основанные на семантической близости в сетях документов

Денис Турдаков
turdakov@gmail.com

Московская секция ACM SIGMOD,
29 октября 2009

План

- **Введение**
  - Многозначность
  - Обзор методов разрешения лексической многозначности
  - Семантическая близость
- **Сети документов**
  - Википедия
  - Вычисление семантической близости
- **Предложенные алгоритмы**
  - Метод, использующий однозначный контекст
  - Метод на основе Марковской модели
  - Метод на основе Марковской модели, обобщенной на случай нескольких независимых цепей

Многозначность

- Морфологическая (грамматическая)
- Синтаксическая
- **Лексическая**
- Семантическая
- Прагматическая

Устранение лексической многозначности

- Наиболее частое значение (MCS)
- Алгоритм Леска (1986): “PINE CONE”
  - PINE
    1. Kinds of **evergreen tree** with needle-sheped leaves
    2. Waste away through sorrow or illness
  - CONE
    1. Solid body which narrows to a point
    2. Something of this shpe whether solid or hollow
    3. Fruit of certain **evergreen tree**
  - PINE #1 ∩ CONE #3 = 2

Основные вопросы

- Что такое значение?
  - гранулярность словарей
  - использование слов в тексте
- Определение контекста
  - микро-контекст
  - тематический контекст
  - контекст, определяемый областью знаний
- Как оценивать и сравнивать алгоритмы?
  - Точность и полнота
  - Тестовые коллекции (SensEval)

Обзор работ

- 50-е - 80-е годы XX века
  - Машинный перевод
- 90-е годы
  - Методы, основанные на внешних источниках знаний (WordNet)
  - Методы, основанные на обучении по размеченным корпусам
  - Методы, основанные на обучении по неразмеченным корпусам XXI век
- XXI век

Семантическая близость

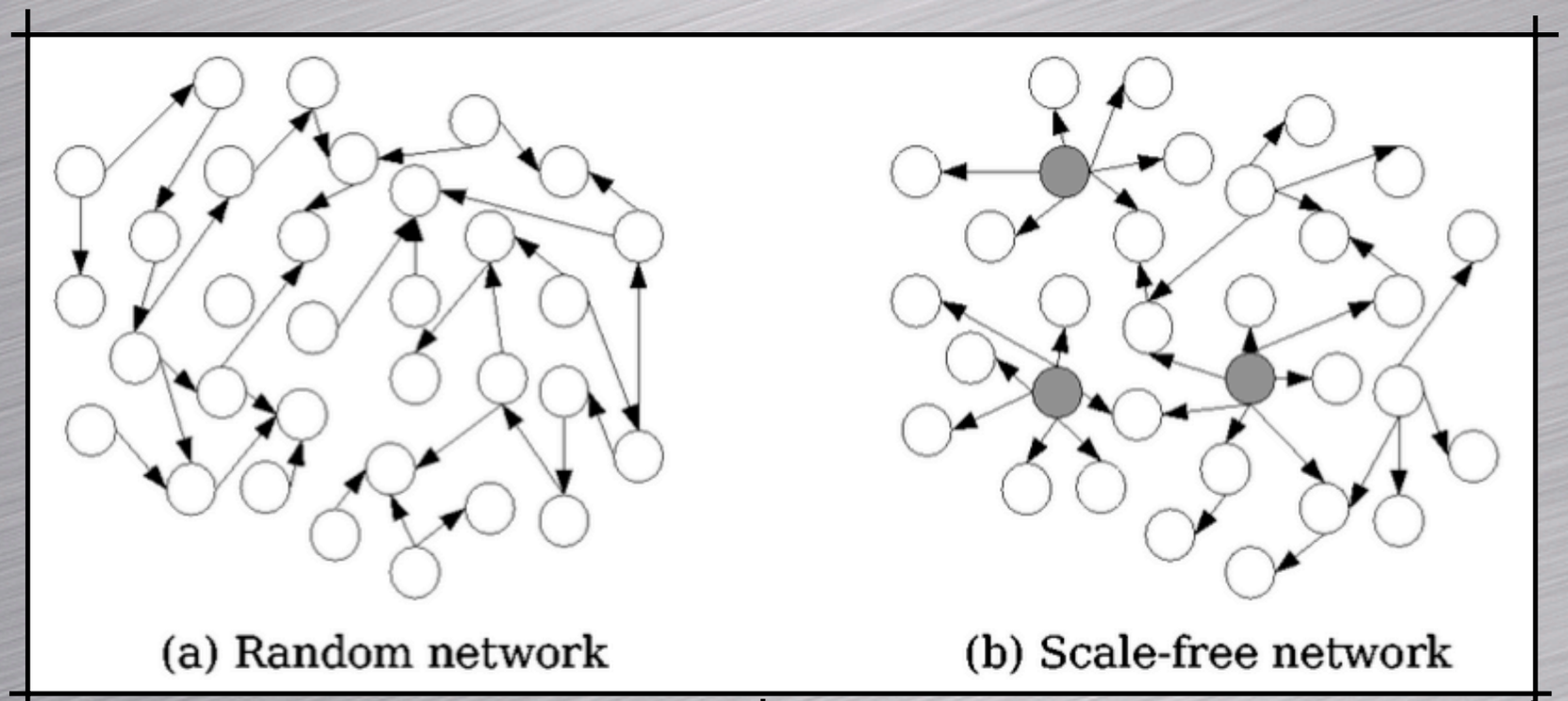
- **Определение**. Семантической близостью называется отображение $f : X \times X \to \mathcal{R}$, ставящее в соответствие паре объектов действительное число
- Свойства
  - $0 \leq f(x, y) \leq 1$
  - $f(x, y) = 1 \Leftrightarrow x = y$
  - $f(x, y) = f(y, x)$

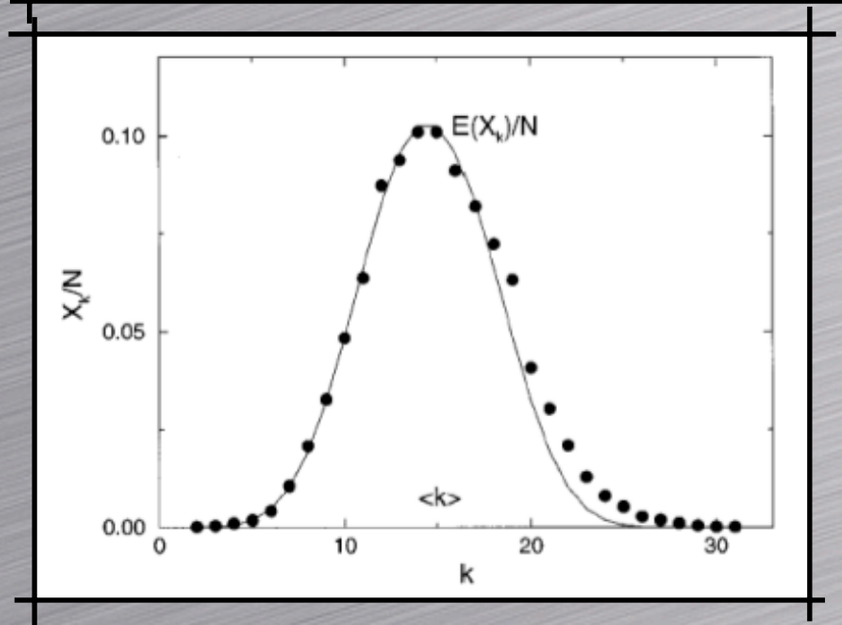
Сеть документов

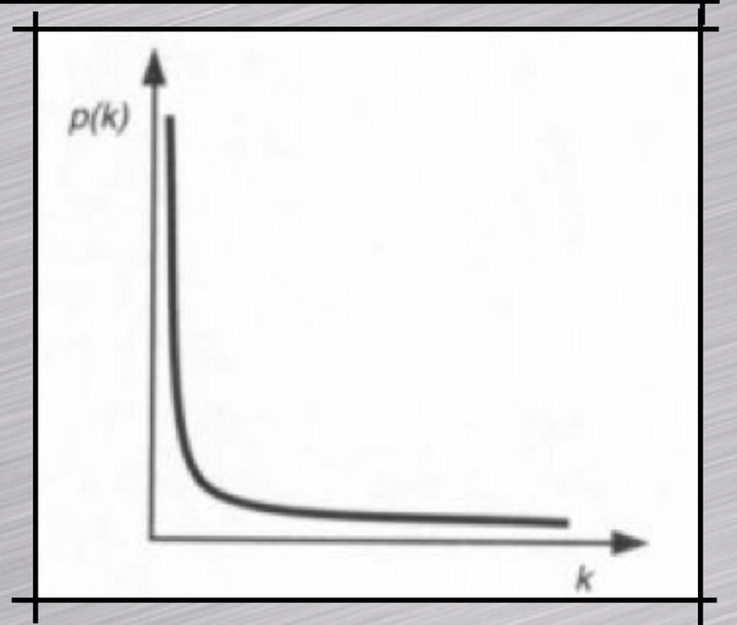
- Filippo Mencer 2004
- **Определение.** Сеть документов - это случайный граф, вершинами которого являются текстовые документы, а ребрами - гипертекстовые ссылки между ними
- Примеры: WWW, Wikipedia, сети цитирования и соавторства в научной литературе и т. д.
- Являются частным случаем безмасштабных графов

Безмасштабные графы

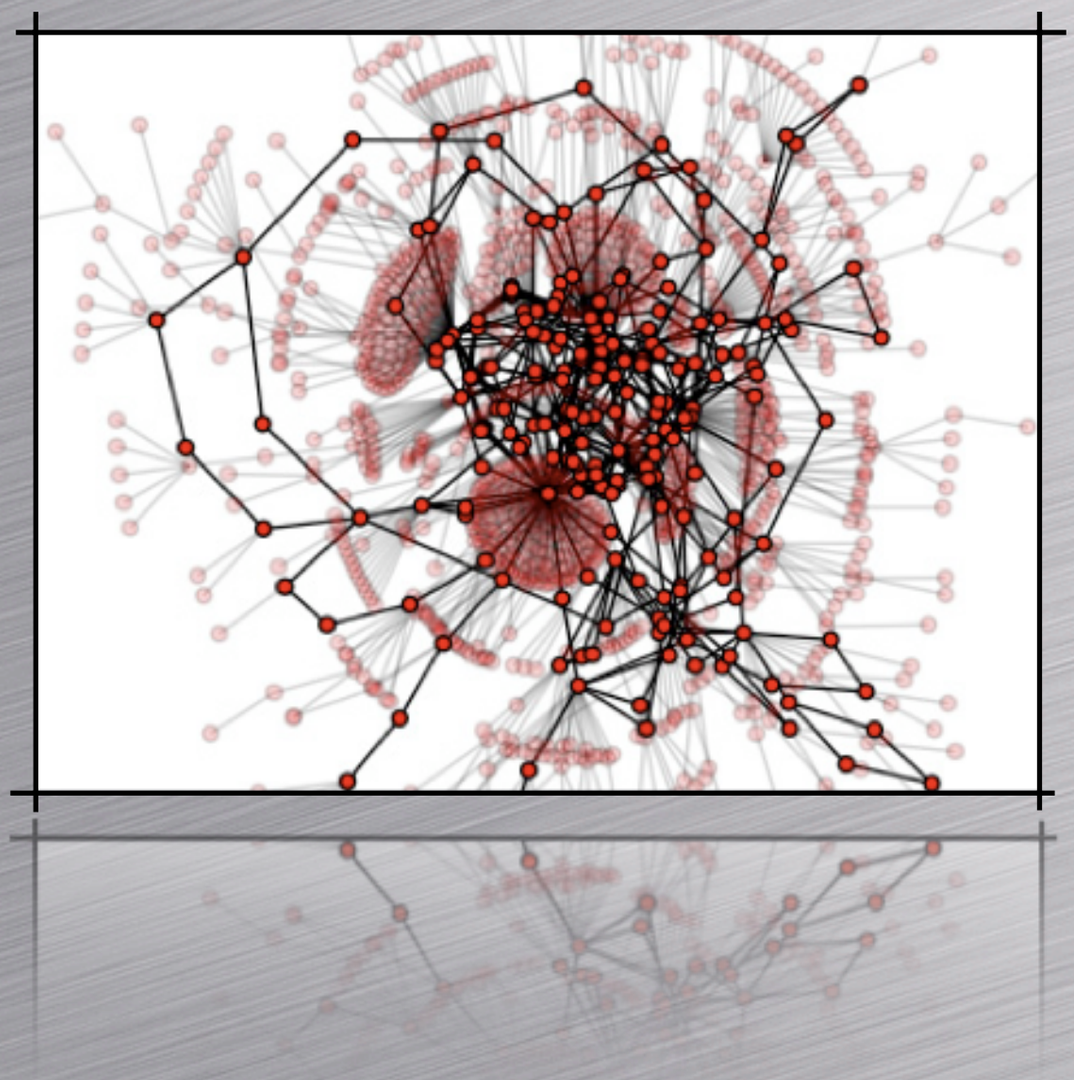
- A. L. Barabasi, R Albert, 1999







Безмасштабные графы



- Хабы, которые связывают сеть
- Большое ядро
- Малый диаметр
- Само-похожесть
- Эффект суммирования шумов

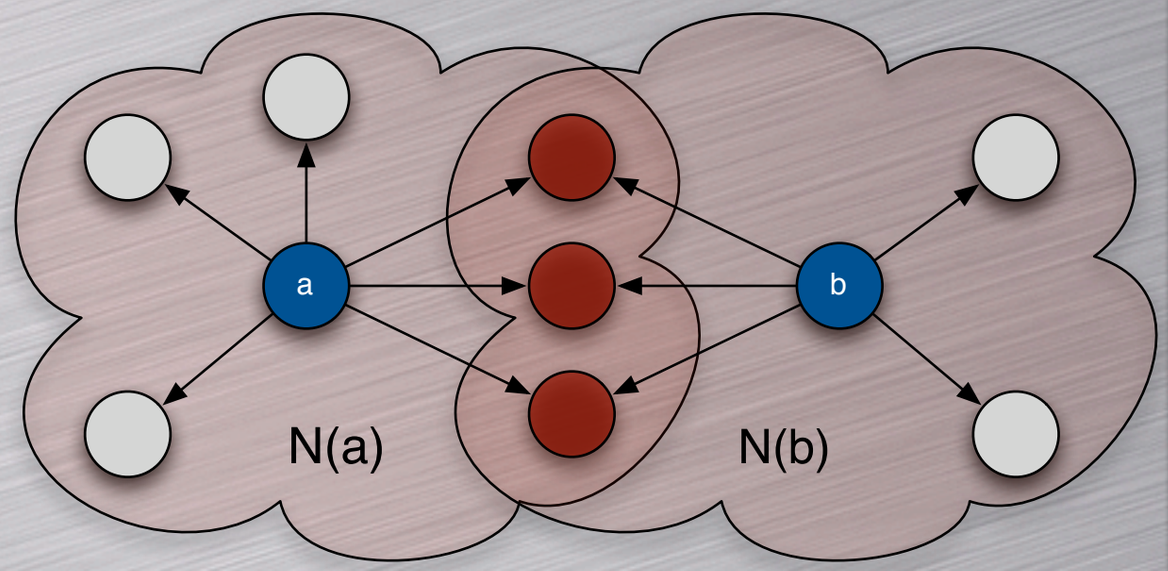
Семантическая близость в сетях документов

- Mencer 2004: Вероятность возникновения ссылки между документами коррелирует с близостью документов, вычисленной на основе их текстового содержимого (векторная модель)
- Методы, основанные на ссылочной структуре
  - Локальные
  - Глобальные

Локальные методы

- $sim_{cos}(a,b) = \dfrac{|N(a) \cap N(b)|}{\sqrt{|N(a)|^2 + |N(b)|^2}}$

- $sim_{Jaccard}(a,b) = \dfrac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|}$



- $sim_{Dice}(a,b) = \dfrac{2\,|N(a) \cap N(b)|}{|N(a)| + |N(b)|}$

- $sim_{GD}(a,b) = \dfrac{\log(\max(|N(a)|, |N(b)|)) - \log(|N(a) \cap N(b)|)}{\log(|W|) - \log(\min(|N(a)|, |N(b)|))}$

Глобальные методы

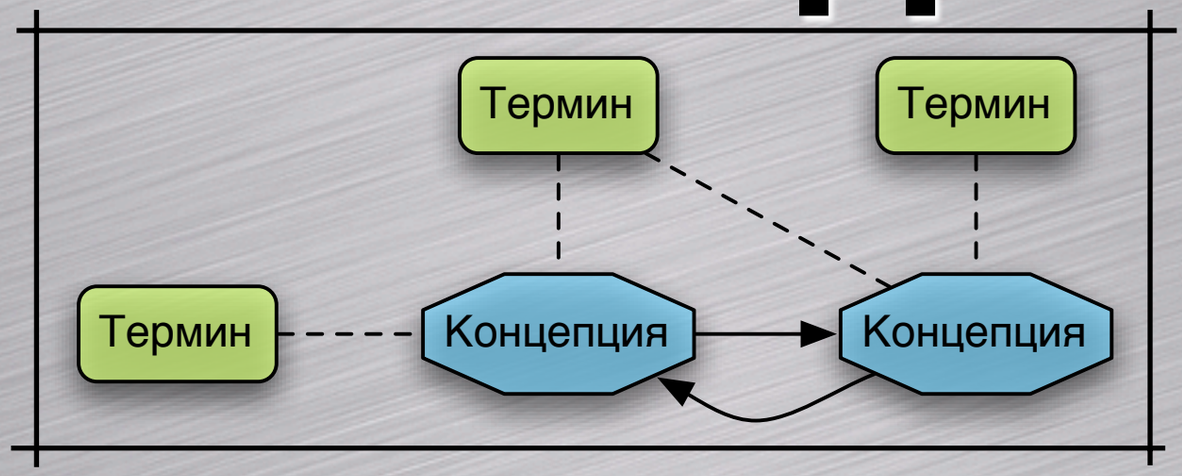
- Основаны на модели случайного блуждания (random walk)
- Методы адаптирующие PageRank для вычисления семантической близости
- SimRank (Jeh, Widom 2002)

SimRank

- $s(a,a) = 1$
- $s(a,b) = \frac{C}{|I(a)|\,|I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$
- $0 < C < 1$
- Lizorkin, Grinev, Velikhov, Turdakov. VLDB'08:

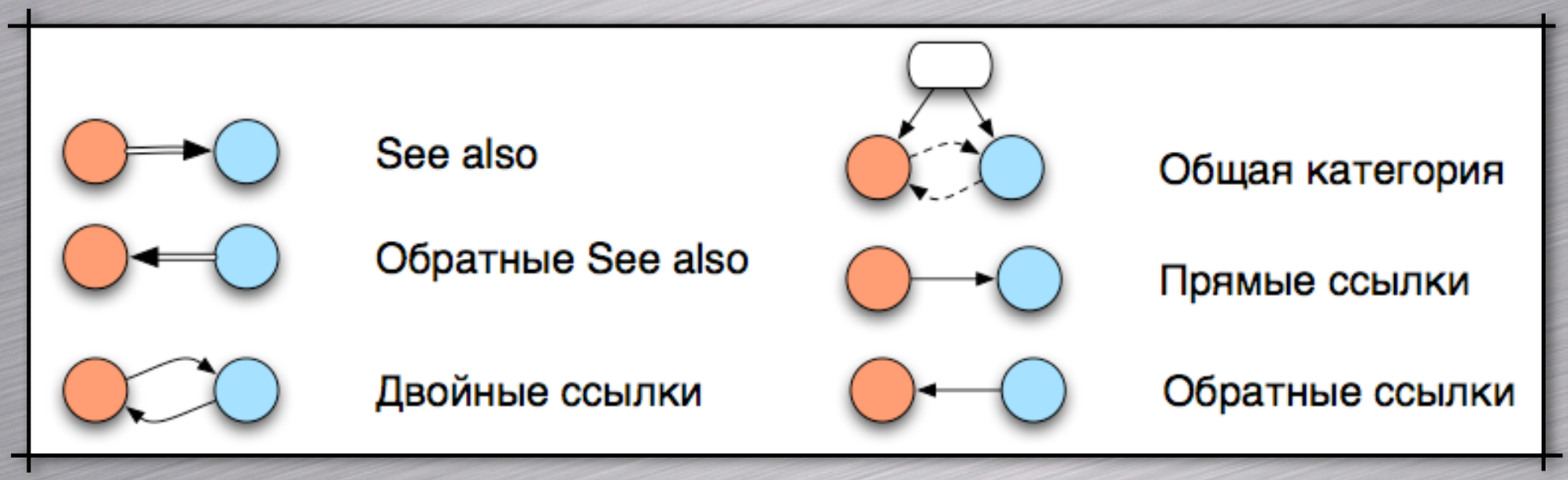
  $O(\min(NL, \frac{N^3}{\log_2 N}))$

Википедия



- Более 3 000 000 концепций
- Заголовки - составные термины
- Списки значений многозначных терминов
- Синонимы
- Ссылки: [[Концепция|Термин]]

Типы ссылок



See Also	5	Двойные ссылки	2
Обратные See Also	2	Общая категория	1.5
Обычные ссылки	1	Обратные обычные ссылки	0.5
Даты	0	Ссылки в инфобоксах	1

+ нормализация весов

Семантическая близость во взвешенных сетях

- Необходимо определить теоретико-множественные операции для взвешенных ребер

- Теория нечетких множеств
  - T-норма $\mu_{A\cap B}(x) = T(\mu_A(x), \mu_B(x))$
  - S-норма $\mu_{A\cup B}(x) = S(\mu_A(x), \mu_B(x))$
- Пример

$$\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x) = \min(\mu_A(x), \mu_B(x))$$
$$\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x) = \max(\mu_A(x), \mu_B(x))$$

Выбор лучшей меры

$$sim(A,B) = \frac{\sum\limits_{N \in \{n(A) \cap n(B)\}} [w(A,N) + w(B,N)]}{\sum\limits_{N \in n(A)} w(A,N) + \sum\limits_{N \in n(B)} w(B,N)}$$

- Наилучший результат разрешения лексической многозначности

$$n(B_1 B_2 \ldots B_m) = \bigcup_{i-1}^{m} n(B_i)$$

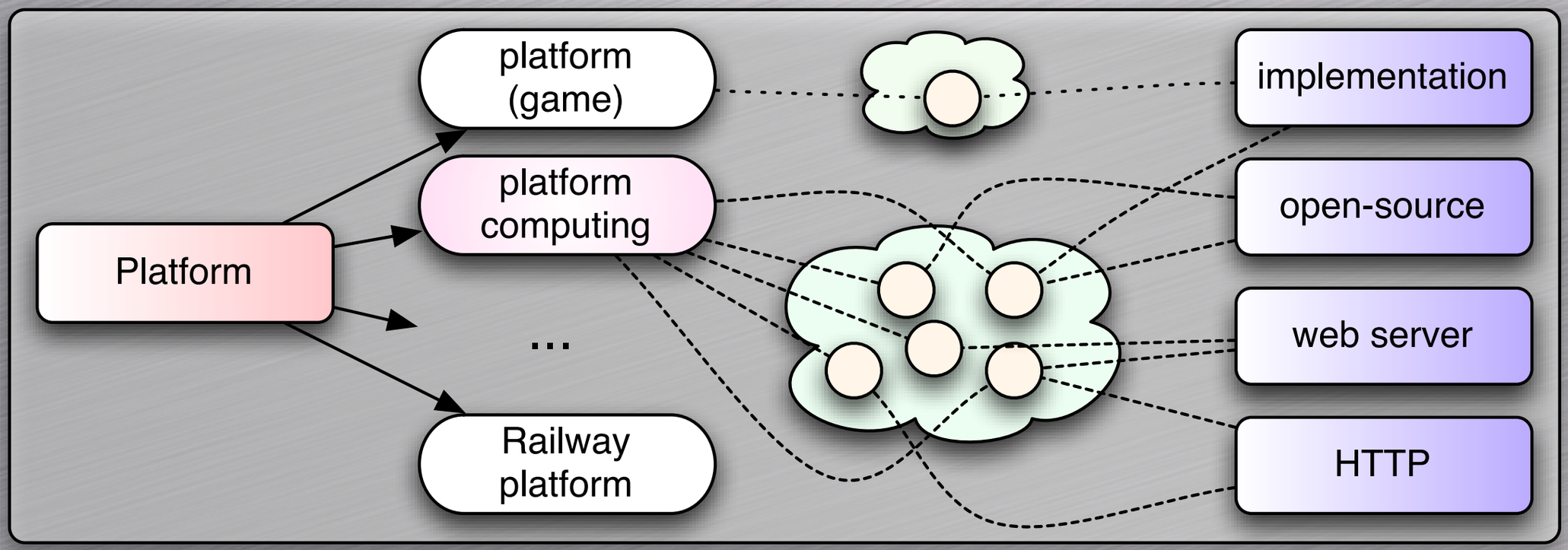
WSD и Википедия

- R. Bunescu, M. Pasca (2007)
- S. Cucerzan (2007)
- R. Mihalcea, A. Csomai. Wikify! (2007)
- O. Medelyan. Topic indexing with Wikipedia (2008)
- D. Milne, I. Witten. Learning to link with Wikipedia (2008)

Метод, использующий однозначный контекст

D.Turdakov, P.Velikhov (SYRCoDIS 2008)

Jigsaw is W3C's open-source project that started in May 1996. It is a web server platform that provides a sample HTTP 1.1 implementation and …



Тестовые коллекции

	статьи Википедии	Milne and Witten [89]	Новости и научные статьи
#документов	500	50	131
#терминов	50947	727	8236
#многозначных терминов	39332	479	6952
#среднее кол-во значений	35.34	29.94	22.34

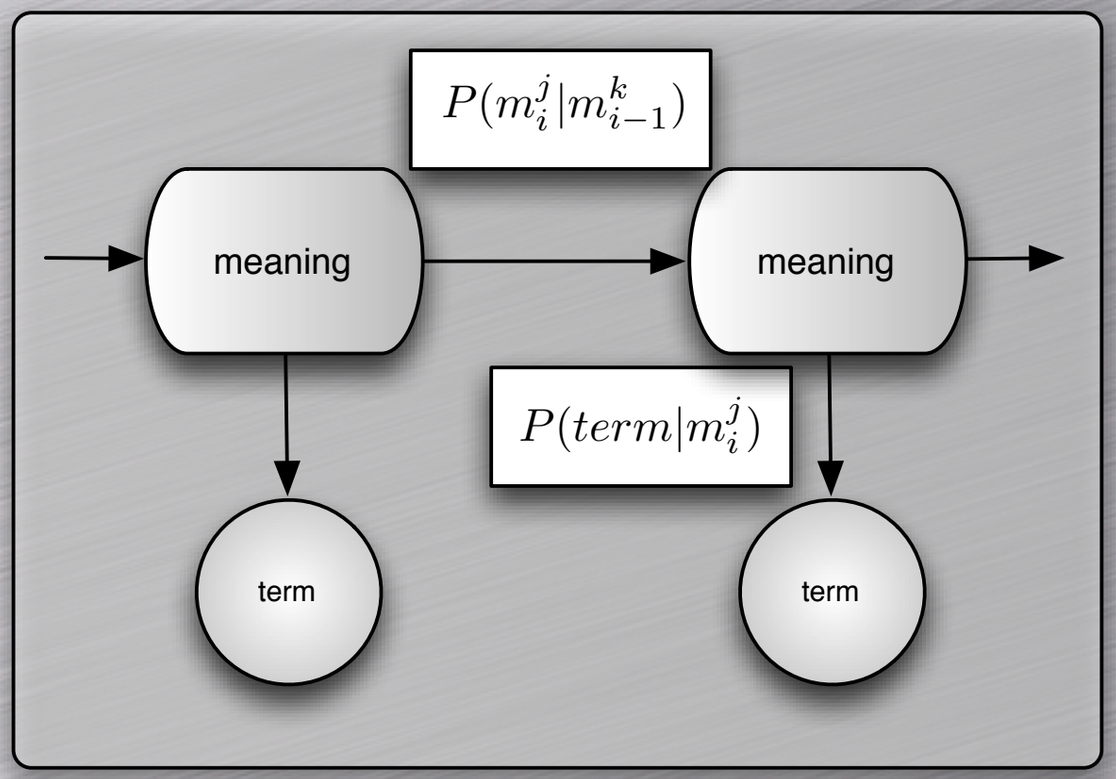
Метод, использующий однозначный контекст

	статьи Википедии	Milne and Witten	Новости и научные статьи
Точность (Июль'08)	87,12	78,81	64,34
Точность (Март'09)	80,55	74,28	40,25

Метод на основе Марковской модели

$$\hat{\mu} = \arg\max_{\mu} P(\mu \mid \tau)$$

$$\hat{\mu} = \arg\max_{\mu} \left(\frac{P(\mu)P(\tau \mid \mu)}{P(\tau)} \right)$$



$$\hat{\mu} = \arg\max_{\mu} \left(\prod_{i=1}^{n} P(m_i|m_{i-k:i-1}) \cdot P(t_i|m_i) \right)$$

- C. Loupy, et. al. (1998): 72.3% (71.5% SemCor)
- A. Molina, et. al. (02, 04): 60.2% (58.0% SensEval-2)

Оценка параметров

RCDL'09:

- Модель перехода:

$$P(m_i|m_{i-k:i-1}) = \alpha \cdot [sim(m_i; m_{i-k:i-1}) + \beta \cdot P(m_i)]$$

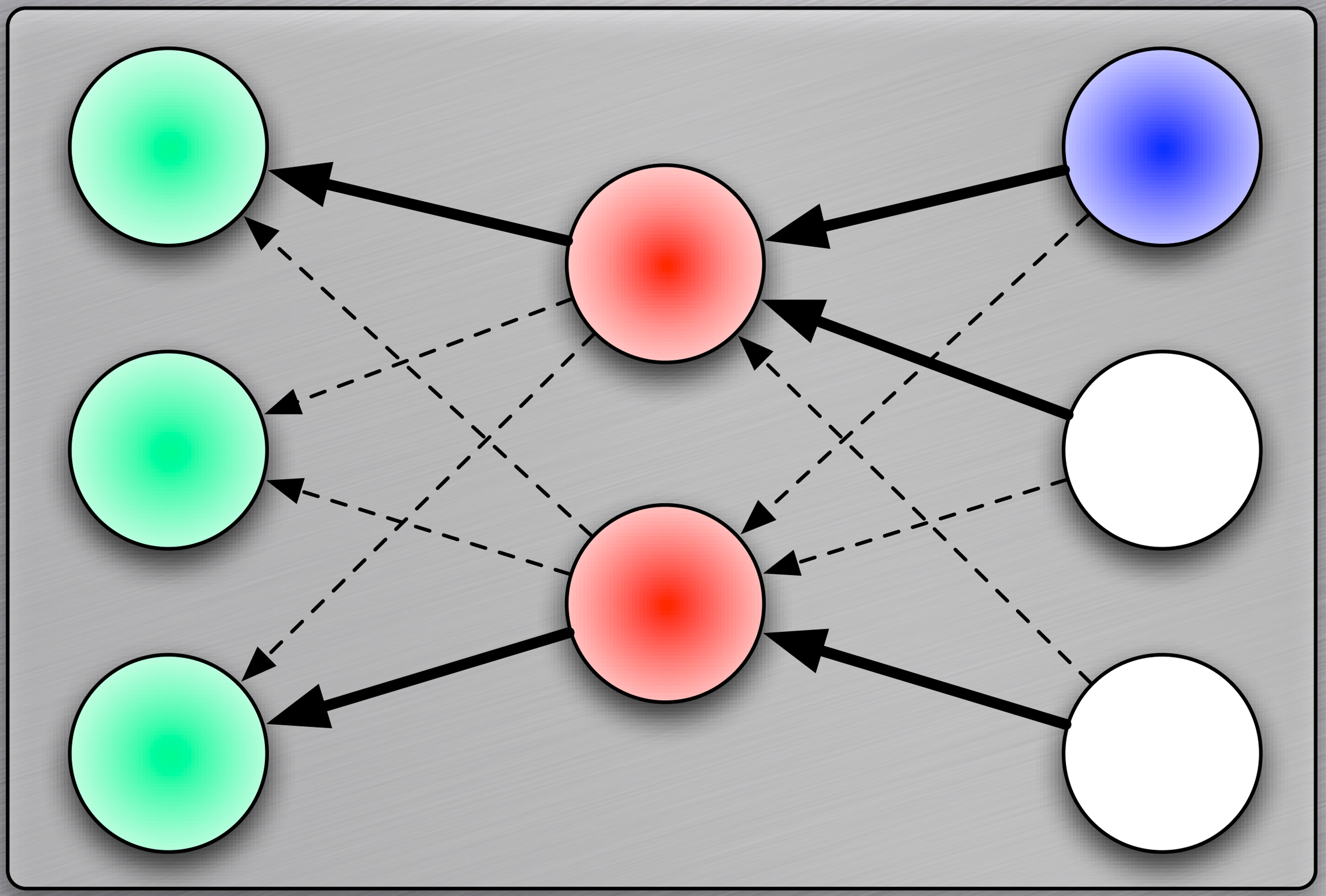
$$\alpha = 1/2 \quad \beta = 1$$

$$P(m_i) = \frac{C(m_i)}{\sum_i C(m_i)}$$

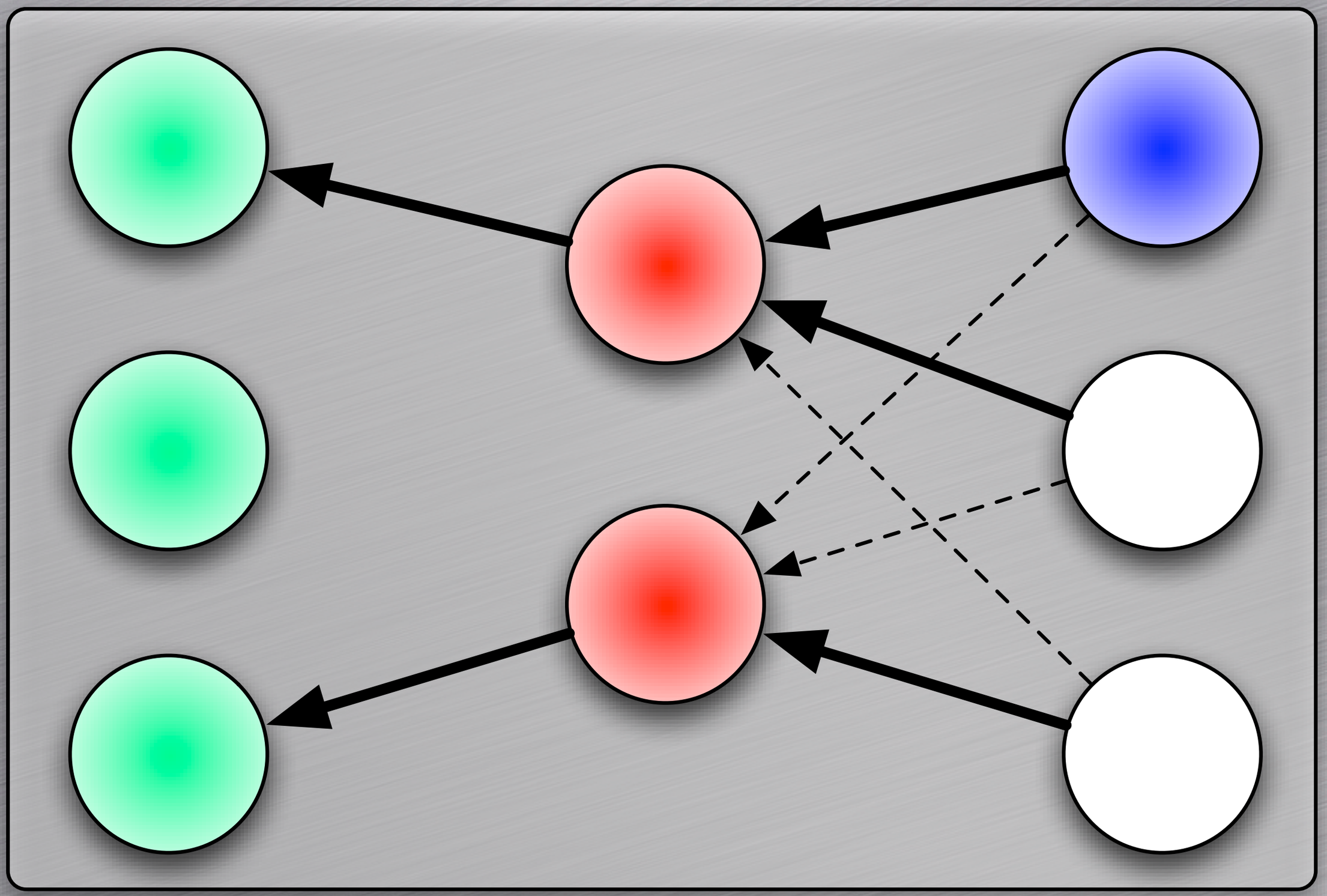
- Модель наблюдения:

$$P(t_i^j|m_i) = \frac{C(t_i^j, m_i)}{C(m_i)}$$

Алгоритм Витерби



Эвристика 1



Оценка: результаты

Новости

Порядок	HMM	Эвристика
0	53,12	53,12
1	54,00	54,00
2	54,50	54,49
3	**54,76**	54,72

Статьи Википедии

Порядок	HMM	Эвристика
0	91,34	91,34
1	91,64	91,64
2	92,40	92,37
3	**92,51**	92,41

Выводы

- Параметры Марковской модели можно оценить с помощью семантической близости
- Эвристика 1 позволяет получить хорошие результаты
- HMM не является лучшей моделью для устранения лексической многозначности многотемных документов

Обобщение Марковской модели

Turdakov, Lizorkin. PACLIC'09

Основные предположения

- **Предположение 1**: Смысл текста можно моделировать с помощью множества независимых марковских цепей
- Мотивация: в документе может быть несколько тем, у каждой темы несколько аспектов

Примеры

1. Football, The drug, sports medicine
   - P(The drug | football) = 0
2. **Christano Ronaldo** hit the **headlines** when he crashed his **Ferrari**
   - Ferrari (sport car)
   - Matteo Ferrari (Italian Football Player)

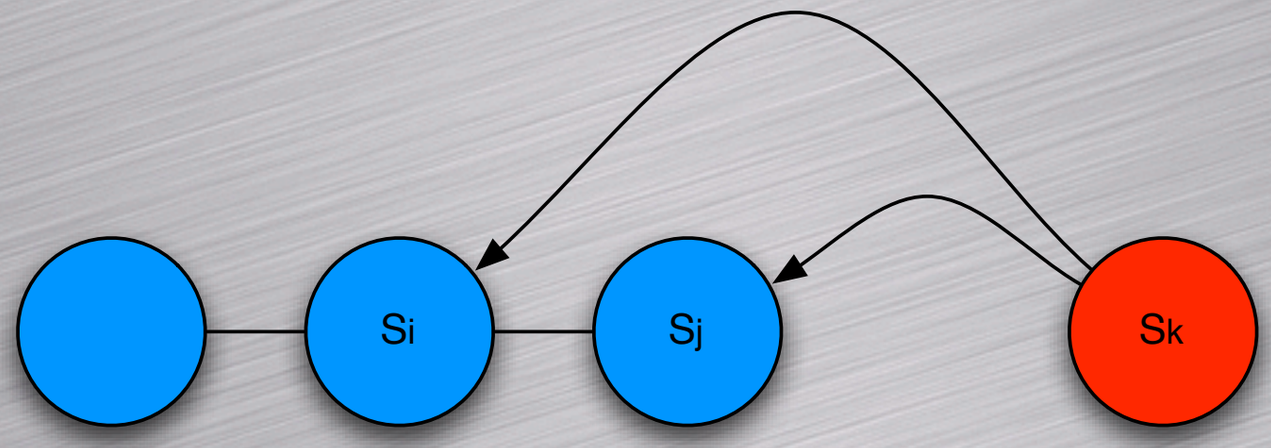
Определение модели: случай одной цепи

$$P(\overline{\mathcal{L}S_k}) = P(\mathcal{L}) \cdot P(S_k \in \mathcal{L}) \cdot P(S_k \mid \mathcal{L})$$

$$P(\mathcal{L}, \mathcal{N}) = P(\mathcal{L}) \cdot P(S_k \notin \mathcal{L}) \cdot P(S_k)$$

- **Предположение 2**. Вероятность события, что текущее состояние принадлежит цепи, зависит только от конечного числа предыдущих состояний этой цепи
- Активные состояния
- Активные цепи
- Порядок обобщенной модели

Определение модели: случай одной цепи



- **Предположение 3.**

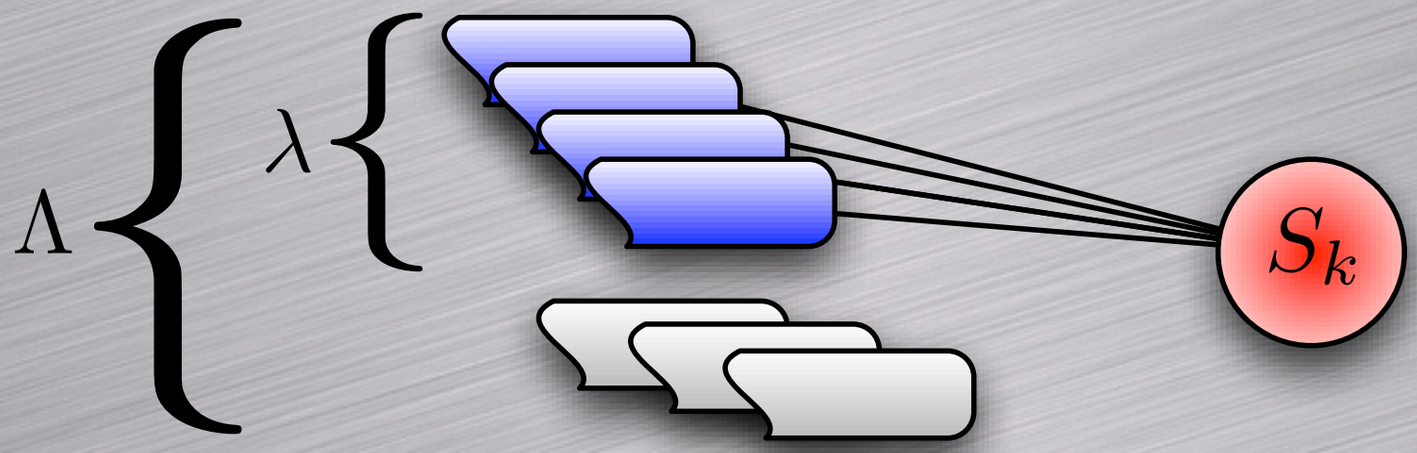
$$P(\widehat{S_i S_k} \text{ and } \widehat{S_j S_k}) = P(\widehat{S_i S_k}) \cdot P(\widehat{S_j S_k})$$

- Тогда

$$P(S_k \notin \mathcal{L}) = \prod_{S_i \in \Omega} [1 - P(\widehat{S_i S_k})] \ ,$$

$$P(S_k \in \mathcal{L}) = 1 - \prod_{S_i \in \Omega} [1 - P(\widehat{S_i S_k})]$$

Определение модели: случай множества цепей

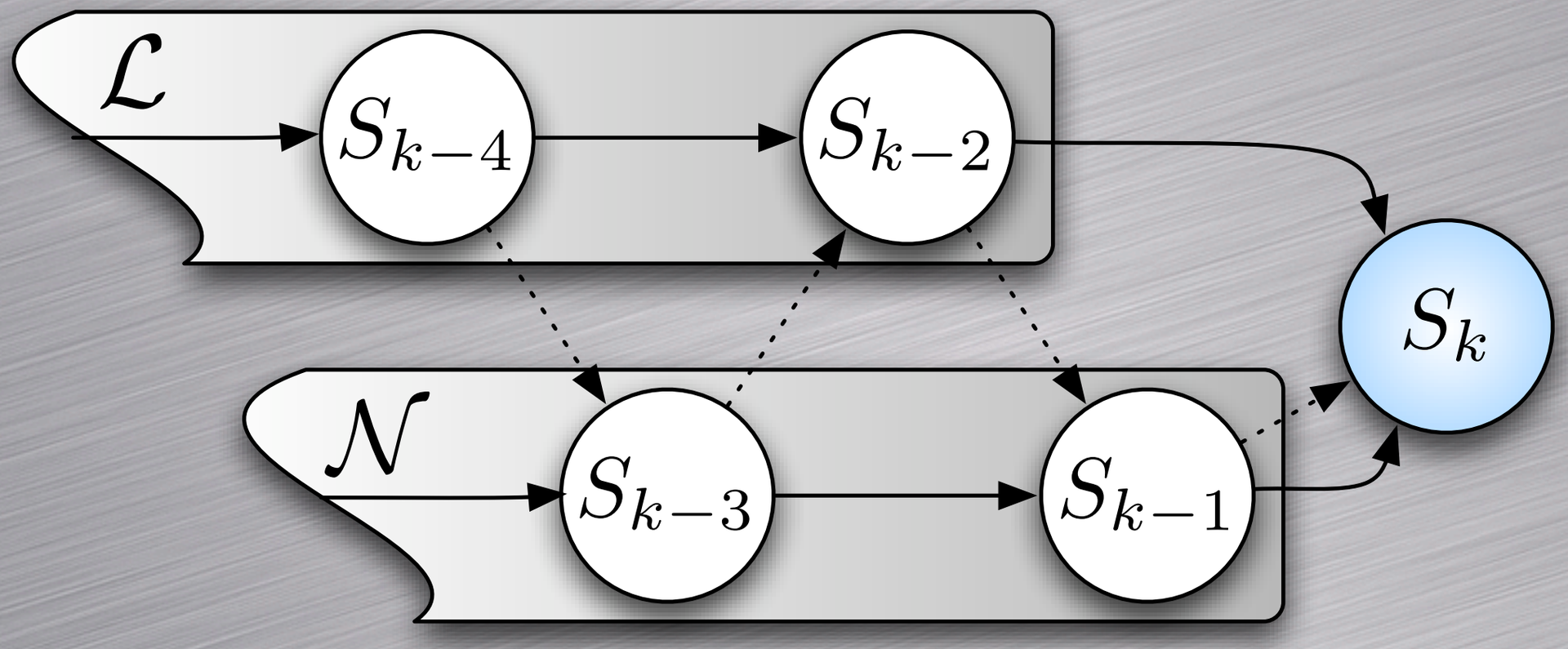


$$P(\overline{\lambda S_k}, \Lambda \setminus \lambda) = P(\Lambda) \cdot P(S_k \in \lambda, S_k \notin \Lambda \setminus \lambda) \cdot P(S_k \mid \lambda)$$

$$P(S_k \in \lambda, S_k \notin \Lambda \setminus \lambda) = \prod_{L_i \in \lambda} P(S_k \in \mathcal{L}_i) \times \prod_{L_j \in (\Lambda / \lambda)} P(S_k \notin \mathcal{L}_j)$$

$$P(S_k \mid \mathcal{L}_{i_1}, \ldots, \mathcal{L}_{i_r}) = P(S_k \mid \mathcal{L}) \ , \quad \mathcal{L} = \bigcup_{j=1}^{r} \mathcal{L}_{i_j}$$

Определение модели: случай множества цепей

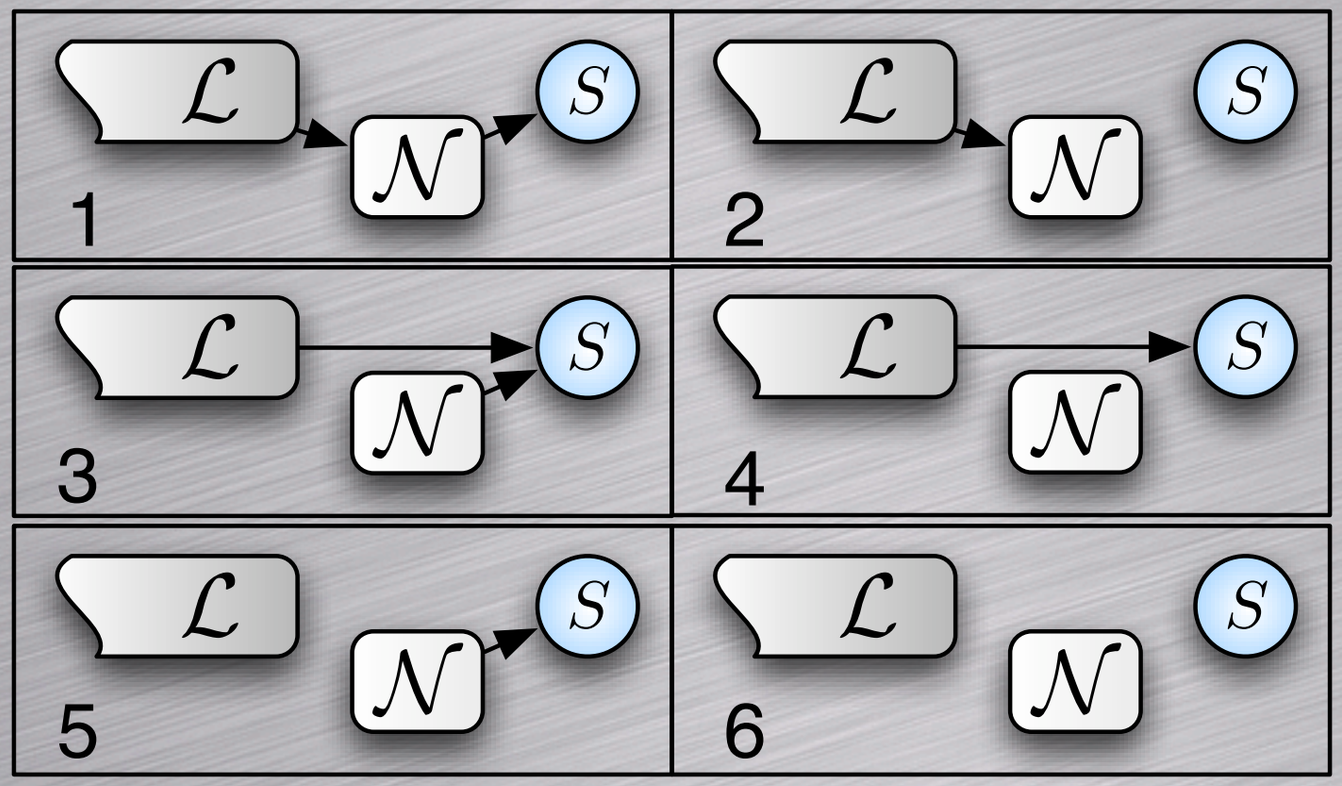


Алгоритм

$$P(\overline{\lambda S_k}, \Lambda \setminus \lambda) = P(\Lambda) \cdot P(S_k \in \lambda, S_k \notin \Lambda \setminus \lambda) \cdot P(S_k \mid \lambda)$$

- Слабая модель
- Полная модель

Слабая модель



Утверждение: Если цепь $\mathcal{N}$ является частью наиболее вероятного пути, и первое состояние S_{i_1} цепи $\mathcal{N}$ принадлежит цепи $\mathcal{L}$, с вероятностью более 0.5, тогда $\mathcal{L}$ также является частью наиболее вероятного пути, а $\mathcal{N}$ является продолжением цепи $\mathcal{L}$.

Вычисление пути

- Необходимо сравнить все возможные пути через активные цепи
- Пусть активные цепи представленны в виде узлов графа, соединенные узлы - соединенные цепи
- Тогда всевозможные пути представляют собой множество неупорядоченных разбиений узлов на связные компоненты

$$B_n = \sum_{k=0}^{n} S(n,k), \quad |S(n,k)| = |S(n-1,k)| \cdot k + |S(n-1,k-1)|$$

Полная модель

- Утверждение не выполняется
- Необходимо рассмотреть все возможные пути через активные состояния

Применение к WSD

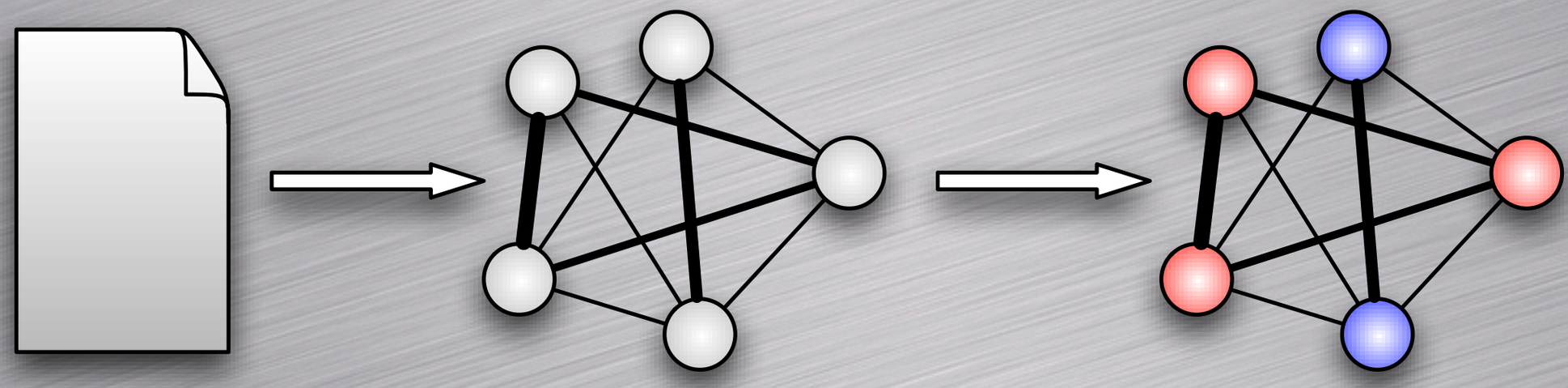
- Модель перехода
- Модель наблюдения
- Вероятность $P(\widehat{m_i m_j})$

Эвристика 2

- Вероятность события, что две концепции принадлежат одной цепи, является функцией от семантической близости:

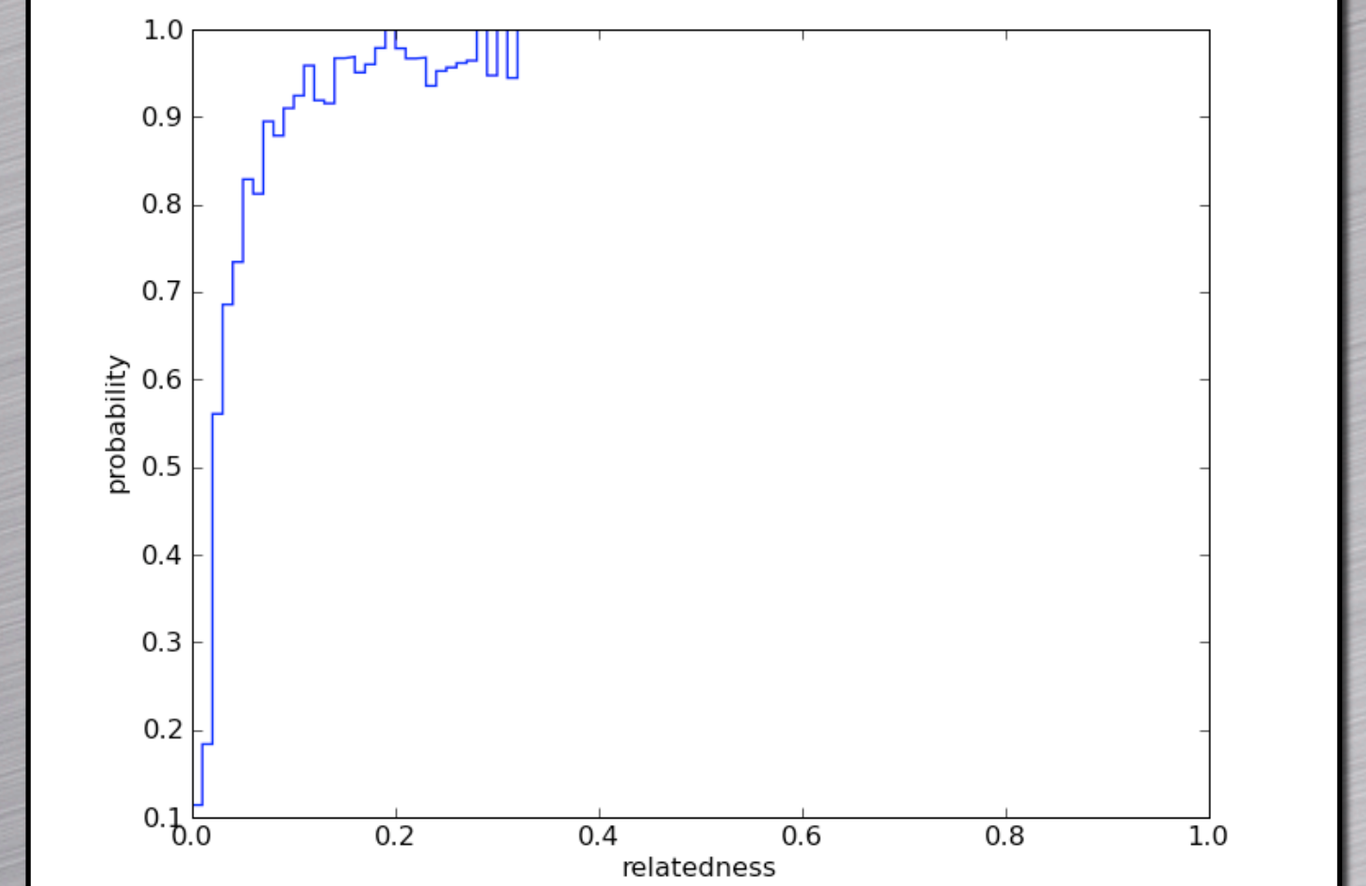
$$P(\widehat{m_1 m_2}) = \phi(sim(m_1, m_2))$$

Обучающее множество



- Множество неразмеченных текстовых документов
- Взвешенный граф концепций
- Кластеризация графа
- Концепции в одном кластере - положительные примеры

Функция ϕ


probability
relatedness
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.4
0.6
0.8
1.0

Результаты

	Wikipedia articles	Milne and Witten [89]	News and scient. papers
Turdakov and Velikhov [2]	85.12	78.81	64.34
MCS	90.10	83.10	67.61
HMM-1	90.13	83.30	67.61
HMM-2	91.51	83.69	67.72
$h = 2, m = 0$	93.36	89.00	74.75
$h = 2, m = 1$	93.68	89.18	75.10
$h = 2, m = 2$	92.87	88.80	73.93
$h = 3, m = 0$	93.67	88.41	75.60
$h = 3, m = 1$	93.72	**89.38**	75.56
$h = 3, m = 2$	**93.78**	**89.38**	**76.13**