Метод извлечения семантических отношений из разнородных источников текстовой информации

Александр Панченко, Аспирант МГТУ им. Н.Э. Баумана и Catholic Univeristy of Louvain

alexander.panchenko@student.uclouvain.be

23 февраля 2012 г.

В данном материале приводятся результаты исследований, выполненных при поддержке гранта РГНФ №12-04-12039в

Plan

Семантические отношения

В рамках данной работы под семантическими отношениями понимаются:

- **синонимы** (отношения эквивалентности):
  $\langle car, SYN, vehicle \rangle, \langle animal, SYN, beast \rangle$
- **гиперонимы** (иерархические отношения):
  $\langle car, HYPER, Jeep\ Cherokee \rangle, \langle animal, HYPER, crocodile \rangle$
- **ко-гиперонимы** (общий гипероним):
  $\langle Toyota\ Land\ Cruiser, COHYPER, Jeep\ Cherokee \rangle$

Формально:

- $r = \langle c_i, t, c_j \rangle$ – **семантическое отношение**, где $c_i, c_j \in C$ – **слова**, такие как *radio* или *receiver operating characteristic*, $t \in T$ – **тип семантического отношения**, такой как *синонимия* или *гипонимия*
- $R \subseteq C \times T \times C$ – множество **семантических отношений**
- $R \subseteq C \times C$ – множество **нетипизированных отношений**

Источники семантических отношений

Тезаурус: граф $G = (C, R)$

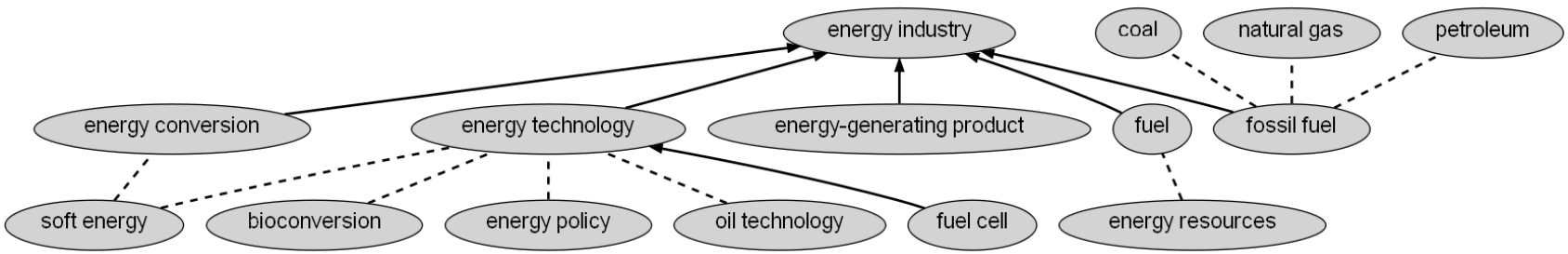


Рис.: Часть информационно-поискового тезуаруса EuroVoc.

$T = \{NT, RT, USE\}$
$R =$

- $\langle$energy-generating product, NT, energy industry$\rangle$
- $\langle$energy technology, NT, energy industry$\rangle$
- $\langle$petrolium, RT, fossil fuel$\rangle$

Другие источники: онтологии, семантические сети, словари синонимов, терминологические классификаторы и т.п.

Применение семантических отношений

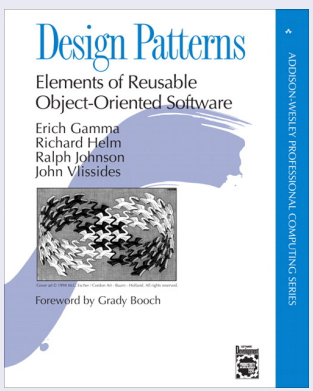
Семантические отношения представляют **знание о языке** полезное для различных приложений **автоматической обработки текста** (АОТ):

- Расширение и рекомендация поискового запроса в ИПС (Hsu et al., 2006)
- Построение вопросно-ответных систем (Sun et al., 2005)
- Категоризация текстовых документов (Tikk et al, 2003)
- Разрешение омонимии (Patwardhan et al., 2003)

Проблема

- Существующие ресурсы часто **недоступны или недостаточны** для
  - конкретного приложения
  - предметной области
  - языка

Пример: магазин продающий книги



"Design Patterns: Elements of Reusable Object-Oriented Software" $\Leftrightarrow$ "Gang of Four Book" $\Leftrightarrow$ GOF

- Как выдать в результате поиска книгу по запросу "GOF"?

Проблема

- **Ручное создание требуемых семантических ресурсов**:
  - (+) Точный результат
  - (–) Крайне дорогостоящий и трудоемкий процесс
  - (–) Неприменимо в большом количестве случаев
- Существующие методы извлечение отношений:
  - (–) Не обеспечивают достаточной точности
- Поэтому,актуальной задачей является разработка методов автоматического извлечения семантических отношений:

Метод извлечения семантических отношений

Вход: слова C, типы семантических отношений T
Выход: семантические отношения $\hat{R} \sim R$

Проблема: существующие методы извлечения отношений

Основанны на ...

- **лексико-синтаксических шаблонах** (Snow, 2004)
- **корпусе текстов** (Филиппович и Прохоров, 2002; Grefenstette, 1994; Curran and Moens, 2002)
- **определениях из словарей/энциклопедий** (Zesch, 2006)
- **семантических сетях** (Pedersen, 2004)
- **фолксономиях** (Markovich and Gabrilovich, 2006)
- **подобии формы слов** (Левеншнейн)
- **структуре гиперссылок документов** (Nakayama et al., 2007)
- **логах поисковых запросов** (Baezo-Yates, 2007)

Проблема: существующие методы

Состояние исследований и разработок:

- Существует **множество разнородных** методов извлечения.
- Предоставляющих **взаимодополняющую информацию**.
- Мы предлагаем **комбинацию** методов для улучшения результатов.

Цели исследования:

- Какой базовый метод извлечения является наилучшим?
- Как **эффективно комбинировать** методы для улучшения точности извлечения?

Извлечение отношений на основе мер подобия

Решение:

1. Представить результаты каждого k-го метода извлечения как метрику подобия $sim_k : C \times C \to [0; 1]$.
   - $sim_k(c_i, c_j)$ – семантическая близость c_i и c_j.
   - $sim_k(x, y) = 1$ тогда и только тогда, когда $x = y = c_i$
   - $sim_k(c_i, c_j) = sim_k(c_j, c_i)$
   - $sim_k(c_i, c_j) \leq sim_k(c_i, c_k) + sim_k(c_k, c_j)$
2. Комбинировать N метрик в sim_{cmb} с помощью $combination\_method(sim_1, sim_2, \ldots, sim_N) \to [0; 1]$
3. Вычислить отношения на основе sim_{cmb}.

Извлечение отношений на основе мер подобия

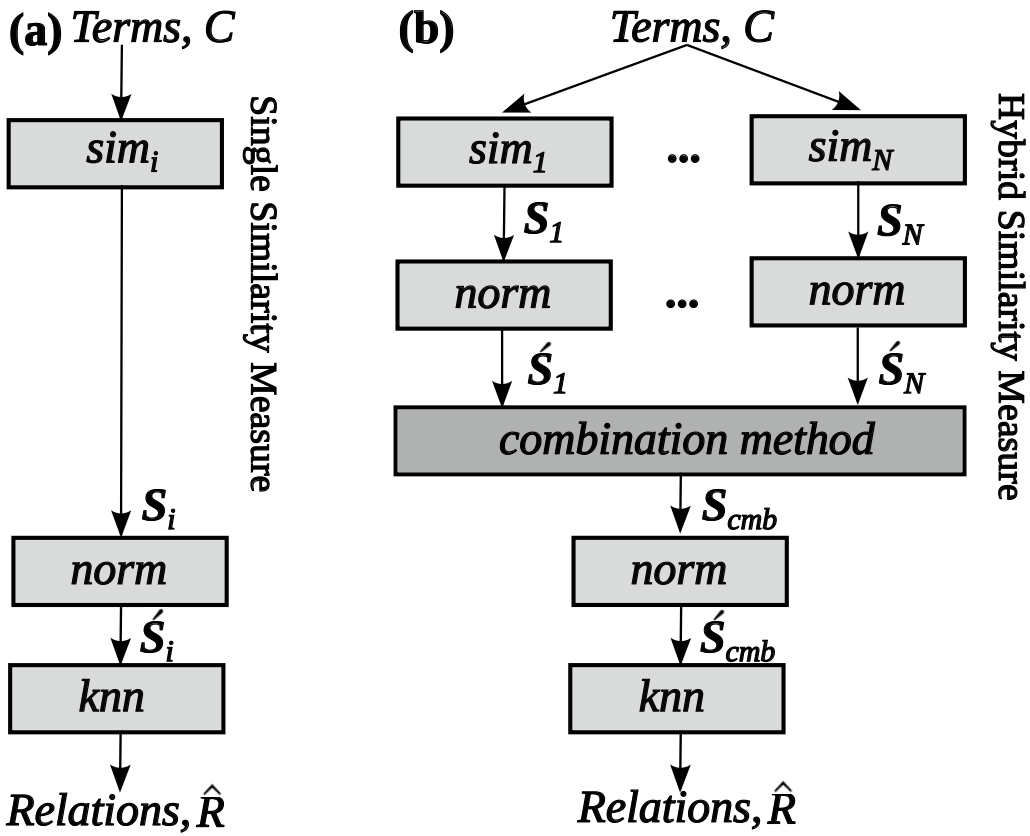


Рис.: Схема метода извлечения семантических отношений с помощью (a) отдельной метрики (b) комбинированной метрики.

Извлечение отношений на основе мер подобия

Метод

Input: Слова C, Количество ближайших соседей k
Output: Семантические отношения $\hat{R} \subseteq C \times C$

for i=1,N **do**
  $\mathbf{S}_i \leftarrow sim_i(C)$;
  $\hat{\mathbf{S}}_i \leftarrow norm(\mathbf{S}_i)$;
end
$\mathbf{S}_{cmb} \leftarrow combination\_method(\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_N)$;
$\hat{\mathbf{S}}_{cmb} \leftarrow norm(\mathbf{S}_{cmb})$;
$\hat{R} \leftarrow knn(\hat{\mathbf{S}}_{cmb}, k)$;
return $\hat{R}$;

Извлечение отношений на основе мер подобия

- sim_i – один из N комбинируемых методов представленных в виде метрик.
- $norm$ – нормализация $\hat{\mathbf{S}} = \frac{\mathbf{S} - min(\mathbf{S})}{max(\mathbf{S})}$
- knn – алгоритм k ближайших соседей:
  $\hat{R} = \bigcup_{i=1}^{|C|} \{\langle c_i, c_j \rangle : (c_j \in \text{ top } k\% \text{ of } c_i) \wedge (s_{ij} > 0)\}$.
- 34 отдельные метрики sim
  - 13 основанные на **корпусе текстов** (дистрибутивный анализ, лексико-синтаксические шаблоны)
  - 9 основанные на **Веб-корпусе текстов**
  - 6 основанные на **семантической сети**
  - 6 основанные на **определениях** из словарей/энциклопедий
- 16 отдельные метрики были выбраны из 34 для комбинирования

Метрики основанные на семантической сети

Данные: семантическая сеть WordNet 3.0, корпус SemCor.
Переменные:

- $len(c_i, c_j)$ – длина **кратчайшего пути** между c_i и c_j
- $len(c_i, lcs(c_i, c_j))$ – длина кратчайшего пути от c_i до **ближайшего общего предка (БОП)** слов c_i и c_j
- $len(c_{root}, lcs(c_i, c_j))$ – длина кратчайшего пути от **корня** c_{root} до БОП слов c_i и c_j (глубина БОП)
- $P(c)$ – **вероятность слова** c, оцененная из корпуса
- $P(lcs(c_i, c_j))$ – **вероятность БОП** слов c_i и c_j

Метрики: Инвертированная длина пути (Jurafsky and Martin, 2009), Leacock-Chodorow (1998), Wu-Palmer (1994), Resnik (1995), Jiang-Conrath (1997), Lin (1998).

Метрики основанные на Веб корпусе

Данные: количество документов возвращенных ИПС (Google, Yahoo, Yahoo BOSS, Bing).
Переменные:

- h_i – **количество документов** возвращенных по запросу слово "c_i"
- h_{ij} – **количество документов** возвращенных по запросу "c_i AND c_j"

Measures:

- Normalized Google Distance NGD (Cilibrasi and Vitanyi, 2007)
- Pointwise Mutual Information PMI-IR (Turney, 2001)

Метрики основанные на корпусе текстов

Данные: корпус WaCkypedia (800M слов) и ukWaC (2000M токенов)

Переменные:

- $\mathbf{f}_i$– вектор признаков представляющий слово c_i, основанный на **контекстном окне**
- $\mathbf{f}_i^s$ – вектор признаков представляющий слово c_i, основанный на **синтаксическом контекстном окне**

Метрики:

- Bag-of-word Distributional Analysis BDA (Sahlgren, 2006)
- Syntactic Distributional Analysis SDA (Curran, 2003)

Другие метрики:

- Latent Semantic Analysis (LSA) на корпусе TASA (Landauer and Dumais, 1997)
- NGD и PMI-IR на корпусе Factiva (Veksler et al., 2008)

Метрики основанные на корпусе текстов

Метрика основанная на **лексико-синтаксических шаблонах**

- **Данные** – корпус WaCkypedia (800M слов)
- 10 паттернов извлечения гиперонимов, ко-гиперонимов и синонимов:
- `such diverse {[occupations]} as {[doctors]}, {[engineers]} and {[scientists]}[PATTERN=1]`
- Сематническая близость между c_i и c_j пропорциональна количеству извлечений n_{ij} с помощью паттернов.

$$sim(c_i, c_j) = \frac{n_{ij}}{max_{ij}(n_{ij})}.$$

Метрики основанные на корпусе текстов

Метрика основанная на **лексико-синтаксических шаблонах**

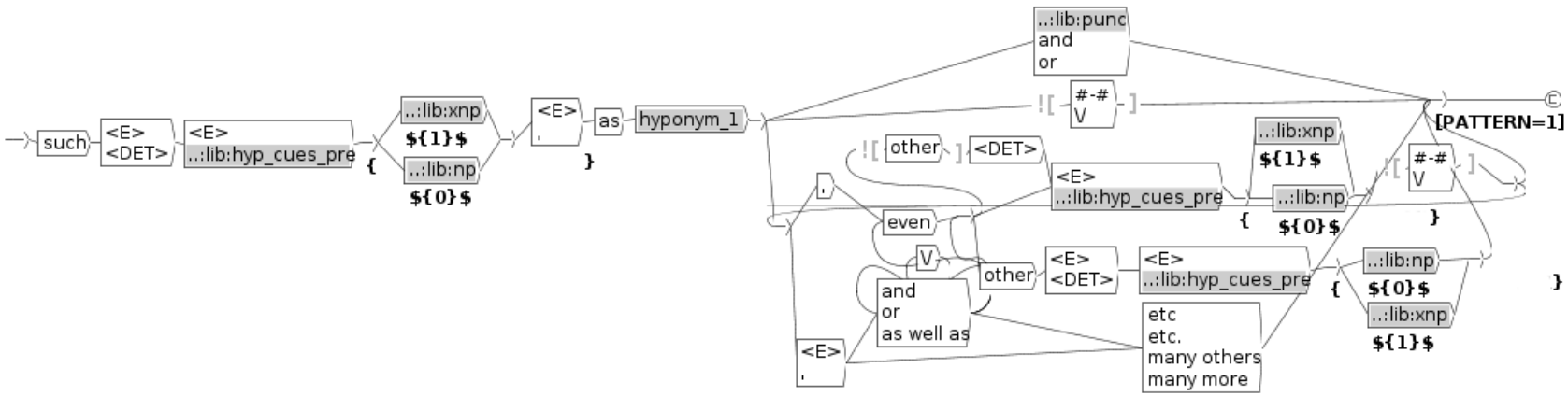


Рис.: Конечный автомат в формате UNITEX для извлечения гиперонимов (под-автоматы обозначены серым цветом; <E> обозначает пустой символ; <DET> обозначает предлоги; полужирные символы – аннотирующие тэги)

Метрики основанные на определениях

Данные: определения WordNet, Wikipedia, и Wiktionary.
Переменные:

- $gloss(c)$ – **определение** слова
- $sim(gloss(c_i), gloss(c_j))$ – **подобие слов** на основании их определений
- $\mathbf{f}_i$ – **вектор признаков** слова c_i, вычисленный на корпусе из всех определений методов контекстного окна
- $\mathbf{f}_i$ **вектор признаков**, представляющий собой определение слова c_i
- $exist(c_i, c_j)$ связь между c_i и c_j в словаре

Метрики:

- WktWiki – BDA на определениях Wiktionary и Wikipedia
- ExtendedLesk на WordNet (Banerjee and Pedersen, 2003)
- GlossVectors на WordNet (Patwardhan and Pedersen, 2006)

Метрики основанные на определениях

Метрика WktWiki

Input: Слова C, Количество признаков β
Output: Матрица подобия, $\mathbf{S}$ $[C \times C]$

$D \leftarrow get\_wiktionary\_definitions(C)$;
$D \leftarrow D \cup get\_wikipedia\_definitions(C)$;
$\mathbf{F} \leftarrow construct\_fmatrix(C, D, \beta)$;
$\mathbf{F} \leftarrow pmi(\mathbf{F})$;
$\mathbf{S} \leftarrow cos(\mathbf{F})$;
$\mathbf{S} \leftarrow update\_similarity(\mathbf{S})$;
return $\mathbf{S}$;

Методы комбинирования метрик подобия

- Цель метода комбинирования метрик *combiation_method* – более точно оценить семантическое подобие слов, на основе **информации из нескольких источников**.
- Метод комбинирования получает на вход множество матриц подобия $\{\mathbf{S}_1, \ldots, \mathbf{S}_K\}$ сгенерированных K метриками и возвращает комбинированную матрицу подобия $\mathbf{S}_{cmb}$.
- Здесь s_{ij}^k – **семантическое подобие слов** c_i и c_j согласно k-ой метрике sim_k.
- Мы используем 8 методов комбинирования метрик.

Методы комбинирования метрик подобия

1. **Mean**. Среднее между значениями подобия метрик:

$$\mathbf{S}_{cmb} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{S}_k \Leftrightarrow s_{ij}^{cmb} = \frac{1}{K}\sum_{k=1,K} s_{ij}^{k}.$$

2. **Mean-Nnz**. Среднее между ненулевыми значениями подобия:

$$s_{ij}^{cmb} = \frac{1}{|k : s_{ij}^{k} > 0, k = 1, K|}\sum_{k=1,K} s_{ij}^{k}.$$

3. **Mean-Zscore**. Среднее между стандартизированными значениями подобия метрик:

$$\mathbf{S}_{cmb} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathbf{S}_k - \mu_k}{\sigma_k},$$

где μ_k и σ_k – среднее и стандартное отклонение значений подобия k-ой метрики ($\mathbf{S}_k$)

Методы комбинирования метрик подобия

4. **Median**. Медиана значений подобия метрик:

$$s_{ij}^{cmb} = median(s_{ij}^{1}, \ldots, s_{ij}^{K}).$$

5. **Max**. Максимальное значение из значений подобия метрик:

$$s_{ij}^{cmb} = max(s_{ij}^{1}, \ldots, s_{ij}^{K}).$$

6. **RankFusion**.Среднее значение ранга пары слов:

$$s_{ij}^{cmb} = \frac{1}{K} \sum_{k=1,K} r_{ij}^{k},$$

где r_{ij}^{k} – ранг, соответствующий значению подобия s_{ij}^{k}.

Методы комбинирования метрик подобия

7. **RelationFusion**. Идея – объединить лучшие отношения найденные каждым методом. При этом отношения извлеченные несколькими методами – лучше.

Input: Sim.matrices produced by N measures $\{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$, kNN threshold k

Output: Combined similarity matrix, $\mathbf{S}_{cmb}$

1 **for** i=1,N **do**
2 $\quad R_i \leftarrow knn(\mathbf{S}_i, k)$;
3 $\quad \mathbf{R}_i \leftarrow relation\_matrix(R_i)$
4 **end**
5 $\mathbf{S}_{cmb} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{R}_i$;
6 **return** $\mathbf{S}_{cmb}$;

$$r_{ij} = \begin{cases} 1 & \text{if } \langle c_i, t, c_j \rangle \in R_k \\ 0 & \text{else} \end{cases}$$

Методы комбинирования метрик подобия

8. **Logit**. Метод основан на обучении с учителем и использует **Логистическую регрессию** (Agresti, 2002).
   - 8.1 Обучение бинарного классификатора на множестве пар семантически связных и несвязных слов BLESS and SN
   - 8.2 Отношение $\langle c_i, t, c_j \rangle$ представлено в виде K-мерного вектора из значений подобия $(s_{ij}^1, \ldots, s_{ij}^K)$
   - 8.3 Целевая переменная r_{ij} (категория):

$$r_{ij} = \begin{cases} 0 & \text{тип отношения } t = random \\ 1 & \text{иначе} \end{cases}$$

   - 8.4 Применение модели для комбинирования метрик:

$$\mathbf{S}_{cmb} = \frac{1}{1 + e^{-z}}, z = w_0 + \sum_{k=1}^{K} w_k \mathbf{S}_k,$$

где $K + 1$ коэффециента $(w_0, w_1, \ldots, w_K)$ – веса регрессии полученные в результате обучения.

Какие из отдельных метрик следует комбинировать?

- Количество **возможных комбинаций**:
  $\sum_{m=2}^{34} C_{34}^{m} = \sum_{m=2}^{34} \frac{34!}{m!(34-m)!} = 1.718 \cdot 10^{10}$
  $\sum_{m=2}^{16} C_{16}^{m} = \sum_{m=2}^{16} \frac{16!}{m!(16-m)!} = 65535$
- **Экспертный выбор** – 5, 9 и 15 метрик из 16
- **Forward Stepwise Procedure** – 7,8a,8b,10 метрик из 16
- Анализ коэффициентов **Логистической регрессии** – 12 из 16

Критерии основанные на суждениях субъектов о семантической связанности

слово, c_i	слово, c_j	субъект, **s**	sim, **s**	субъект (ранг), **r**	sim (ранг), $\hat{\mathbf{r}}$
tiger	cat	7.35	0.85	1	3
book	paper	7.46	0.95	2	2
computer	keyboard	7.62	0.81	3	1
...	...	...	...	...	...
possibility	girl	1.94	0.25	64	65
sugar	approach	0.88	0.05	65	23

Данные:

- WordSim353 – 353 пар слов (Finkelstein, 2002)
- MC – 30 пар слов (Miller Charles, 1991)
- RG – 65 пар слов (Rubenstein Goodenough, 1965)

Коэффициент корреляции Пирсона: $\rho = \frac{cov(\mathbf{s},\hat{\mathbf{s}})}{\sigma(\mathbf{s})\sigma(\hat{\mathbf{s}})}$

Коэффициент корреляции Спирмена:: $r = \frac{cov(\mathbf{r},\hat{\mathbf{r}})}{\sigma(\mathbf{r})\sigma(\hat{\mathbf{r}})}$

Критерии точности извлечения отношений

слово, c_i	**слово,** c_j	**тип отношения,** t
judge	adjudicate	syn
judge	arbitrate	syn
judge	asessor	syn
judge	chancellor	syn
judge	gendarmerie	syn
judge	sheriff	syn
...	...	...
judge	pc	random
judge	fare	random
judge	lemon	random

Данные:

- BLESS (Baroni and Lenci, 2011) – 26554 отношений (hyper, coord, mero, event, attri, random)
- SN (Panchenko, 2012) – 14682 отношений (syn, random)

Критерии точности извлечения отношений

- Основаны на количестве правильно извлеченных отношений.
- R – все семантические отношения, не являющиеся случайными ($\langle animal, random, bishop \rangle$ и т.п.)
- $\hat{R}(k)$ множество извлеченных отношений при количестве ближайших соседей k

Критерии

- Точность $P(k) = \frac{|R \cap \hat{R}(k)|}{|\hat{R}(k)|}$
- Полнота $R(k) = \frac{|R \cap \hat{R}(k)|}{|R|}$.
- F1-мера $F(k) = 2 \cdot \frac{Precision(k) \cdot Recall(k)}{Precision(k) + Recall(k)}$
- MAP $M(M) = \frac{1}{M} \sum_{k=1}^{M} Precision(k)$

- Мы используем $P(10)$, $P(20)$, $P(50)$, $R(50)$, $MAP(20)$, $MAP(50)$.

Пример: оценка точности извлечения отношений

- Точность $P(50\%) = \frac{1}{7} \approx 0.86$

слово, c_i	**слово,** c_j	**тип отношения**	s_{ij}
aficionado	enthusiast	syn	0.07197
aficionado	fan	syn	0.05195
aficionado	admirer	syn	0.01964
aficionado	addict	syn	0.01326
aficionado	devotee	syn	0.01163
aficionado	foundling	random	0.00777
aficionado	fanatic	syn	0.00414
aficionado	adherent	syn	0.00353
aficionado	capital	random	0.00232
aficionado	statute	random	0.00029
aficionado	blot	random	0.00025
aficionado	meddler	random	0.00005
aficionado	enlargement	random	0.00003
aficionado	bawdyhouse	random	0.00000

Отдельные метрики: корреляция с суждениями людей

Sim.Measure	MC Dataset		RG Dataset		WordSim353 Dataset	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Random	0.172 ***	0.056 ***	-0.060 ***	-0.047 ***	-0.158 ***	-0.122 ***
WN-Resnik	0.823	**0.784**	0.823	0.757	0.350	**0.330**
WN-Short.Path	0.755	0.724	0.782	0.788	0.366	0.290
WN-Leack.Chod.	0.779	0.724	0.841	**0.789**	0.313	0.295
WN-WuPalmer	0.768	0.742	0.800	0.775	0.270	0.330
WN-Lin	0.769	0.754	0.737	0.619	0.287	0.203
WN-JiangConrath	0.473 *	0.719	0.575	0.587	0.227	0.175
NGD-Bing	0.035 ***	0.063 ***	0.174 ***	0.181 ***	0.042 ***	0.058 ***
NGD-Yahoo	0.387 **	0.330 ***	0.448	0.445	0.290	0.254
NGD-Google	0.085 ***	0.019 ***	-0.013 ***	-0.012 ***	0.120 **	0.150 *
NGD-GoogleWiki	0.306 ***	0.334 ***	0.452	**0.501**	0.205	0.250
PMI-IR-Bing	0.079 ***	0.120 ***	0.116 ***	0.149 ***	0.000 ***	0.003 ***
PMI-IR-Google	0.046 ***	-0.107 ***	-0.061 ***	-0.039 ***	0.097 ***	0.113 **
PMI-IR-GoogleWiki	0.508 *	**0.498 ***	0.401	0.411	0.254	**0.279**
BDA-sent-10000	0.642	0.638	0.694	0.703	0.383	0.362
BDA-1-5000	0.658	0.676	0.704	0.758	0.448	0.438
BDA-2-5000	0.667	0.638	0.698	0.734	0.441	0.439
BDA-3-5000	0.722	0.692	0.752	0.782	0.467	**0.465**
BDA-5-5000	0.710	0.683	0.755	**0.787**	0.467	0.455
BDA-8-5000	0.707	0.697	0.746	0.764	0.455	0.440
BDA-10-5000	0.710	0.718	0.746	0.764	0.443	0.425
SDA-6-100000	0.759	0.790	0.741	0.792	0.380	0.496
SDA-9-100000	0.756	0.790	0.732	0.7873	0.384	0.491
SDA-21-100000	0.756	**0.790**	0.731	0.785	0.384	0.490
LSA-Tasa	0.737	0.694	0.645	0.604	0.527	0.565
NGD-Factiva	0.602	0.602	0.618	0.599	0.565	0.599
PMI-Factiva	0.312 ***	0.442 **	0.436	0.517	0.314	0.559
Def-WN-GlossVec.	0.566	0.653	0.647	0.738	0.383	0.322
Def-WN-Ext.Lesk	0.355 ***	0.792	0.340 *	0.717	0.209	0.409
Def-Wkt-1000	0.625	0.687	0.655	0.760	0.416	0.492
Def-Wkt-2500	0.625	0.687	0.655	0.760	0.382	0.527
Def-WktWiki-1000	0.704	**0.759**	0.701	**0.754**	0.453	**0.545**
Def-WktWiki-2500	0.704	0.759	0.701	0.754	0.416	0.520
Comb-Avg-4	0.847	0.859	0.867	0.887	0.500	0.508
Comb-Avg-8	0.858	0.858	0.867	0.883	0.537	**0.555**
Comb-Avg-14	0.847	**0.859**	0.867	**0.887**	0.500	0.508

Рис.: Pearson – корреляция Пирсона, Spearman – корреляция Спирмена.

Отдельные метрики: извлечение отношений

Sim.Measure	BLESS Dataset						SN Dataset					
	P(10)	*P (20)*	*M(20)*	*P(50)*	*F(50)*	*M(50)*	*P(10)*	*P(20)*	*M(20)*	*P(50)*	*F(50)*	*M(50)*
Random	0.546	0.541	0.549	0.543	0.522	0.545	0.504	0.501	0.506	0.498	0.498	0.501
WN-Resnik	0.977	0.958	**0.979**	0.718	0.690	**0.900**	0.948	0.908	0.949	0.725	0.725	0.874
WN-Short.Path	0.967	0.925	0.969	0.722	0.693	0.885	0.981	0.947	0.977	0.752	0.752	0.906
WN-Leack.Chod.	0.967	0.925	0.969	0.722	0.693	0.885	0.982	0.951	0.978	0.756	0.756	0.911
WN-WuPalmer	0.978	0.938	0.976	0.706	0.678	0.885	0.979	0.959	**0.979**	0.764	0.764	**0.916**
WN-Lin	0.975	0.919	0.969	0.776	0.745	0.880	0.924	0.853	0.906	0.637	0.637	0.808
WN-JiangConrath	0.981	0.909	0.970	0.732	0.703	0.875	0.916	0.835	0.897	0.615	0.615	0.792
NGD-Bing	0.725	0.692	0.713	0.695	0.670	0.695	0.676	0.682	0.696	0.639	0.639	0.681
NGD-Yahoo	0.940	0.907	0.941	0.782	0.751	**0.885**	—	—	—	—	—	—
NGD-YahooBoss	0.847	0.843	0.819	0.747	0.718	0.808	—	—	—	—	—	—
NGD-Google	0.991	0.934	**0.980**	0.651	0.625	0.865	—	—	—	—	—	—
NGD-GoogleWiki	0.874	0.836	0.875	0.702	0.674	0.815	—	—	—	—	—	—
PMI-IR-Bing	0.675	0.650	0.664	0.692	0.667	0.664	0.610	0.608	0.622	0.647	0.647	0.635
PMI-IR-YahooBOSS	0.823	0.822	0.797	0.724	0.696	0.787	—	—	—	—	—	—
PMI-IR-Google	0.822	0.749	0.839	0.660	0.634	0.758	—	—	—	—	—	—
PMI-IR-GoogleWiki	0.791	0.761	0.808	0.676	0.649	0.755	—	—	—	—	—	—
BDA-sent-10000	0.962	0.920	0.960	0.799	0.767	0.901	0.941	0.898	0.944	0.724	0.724	0.866
BDA-1-5000	0.971	0.94	0.967	0.826	0.793	0.921	0.969	0.926	0.965	0.737	0.737	0.891
BDA-2-5000	0.966	0.939	0.962	0.829	0.796	0.920	0.970	0.929	0.966	0.738	0.738	0.894
BDA-3-5000	0.97	0.947	0.969	0.835	0.802	**0.927**	0.974	0.932	0.969	0.743	0.743	0.897
BDA-5-5000	0.975	0.946	0.973	0.833	0.800	0.927	0.971	0.929	0.965	0.744	0.744	0.893
BDA-8-5000	0.974	0.943	0.972	0.827	0.794	0.923	0.968	0.924	0.964	0.741	0.741	0.889
BDA-10-5000	0.972	0.941	0.971	0.821	0.789	0.92	0.962	0.922	0.959	0.737	0.737	0.886
SDA-6-100000	0.984	0.948	0.982	0.810	0.778	0.923	0.978	0.945	0.974	0.749	0.749	0.905
SDA-9-100000	0.984	0.951	0.982	0.809	0.777	0.923	0.977	0.945	0.973	0.753	0.753	0.907
SDA-21-100000	0.985	0.953	**0.983**	0.810	0.778	0.924	0.978	0.946	**0.974**	0.753	0.753	**0.907**
LSA-Tasa	0.967	0.936	0.966	0.801	0.769	0.912	0.901	0.839	0.893	0.637	0.637	0.798
NGD-Factiva	0.959	0.916	0.958	0.800	0.768	0.896	0.900	0.832	0.895	0.651	0.651	0.804
PMI-Factiva	0.903	0.860	0.912	0.816	0.784	0.863	0.826	0.768	0.825	0.606	0.606	0.737
Def-WN-GlossVec.	0.894	0.860	0.902	0.742	0.712	0.843	0.930	0.895	0.932	0.719	0.719	0.862
Def-WN-Ext.Lesk	0.940	0.870	0.941	0.716	0.687	0.847	0.950	0.872	**0.942**	0.653	0.653	0.830
Def-Wkt-1000	0.926	0.885	0.926	0.783	0.752	0.866	0.907	0.868	0.902	0.678	0.678	0.825
Def-Wkt-2500	0.915	0.882	0.923	0.785	0.754	0.867	0.928	0.898	0.924	0.704	0.704	0.855
Def-WktWiki-1000	0.942	0.905	**0.945**	0.754	0.725	**0.876**	0.917	0.878	0.911	0.696	0.696	0.841
Def-WktWiki-2500	0.931	0.891	0.939	0.765	0.734	0.875	0.937	0.912	0.936	0.726	0.726	**0.871**
Comb-Avg-4	0.992	0.969	0.990	0.787	0.756	0.930	0.980	0.952	0.979	0.768	0.768	0.914
Comb-Rel-4	0.989	0.970	0.988	0.737	0.708	0.915	0.975	0.943	0.973	0.696	0.696	0.891
Comb-Avg-8	0.994	0.974	0.990	0.774	0.743	0.932	0.955	0.875	0.945	0.660	0.660	0.835
Comb-Rel-8	0.994	0.975	**0.992**	0.802	0.770	**0.941**	0.989	0.971	**0.986**	0.760	0.760	**0.925**
Comb-Avg-14	0.994	0.979	0.992	0.792	0.760	0.938	0.957	0.880	0.947	0.663	0.663	0.839
Comb-Rel-14	0.994	0.973	0.992	0.811	0.779	0.939	0.987	0.966	0.985	0.759	0.759	0.920

Рис.: Здесь P – точность, R – полнота, F – F1-мера, M – MAP.

Отдельные метрики

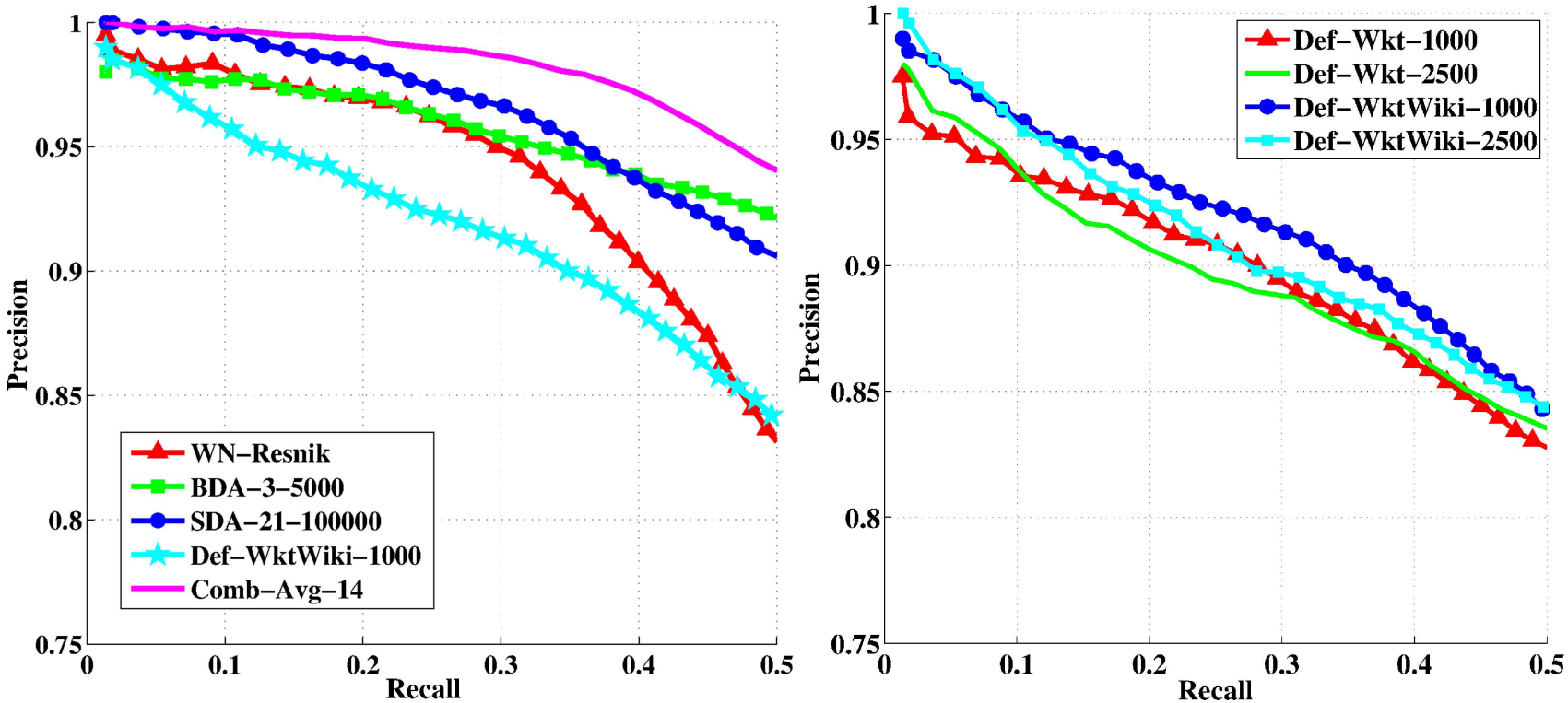


Рис.: Графики Точность-Полнота (слева) 4х лучших метрик основанных на корпусе, семантических сетях, определениях и метрика, основанная на среднем значении 14 метрик; (слева) метрики основанных на определениях Викисловаря и Википедии.

Отдельные и комбинированные метрики

Similarity Measure	MC	RG	WS	BLESS						SN					
	ρ	*ρ*	*ρ*	*P(10)*	*P (20)*	*M(20)*	*P(50)*	*M(50)*	*R(50)*	*P(10)*	*P(20)*	*M(20)*	*P(50)*	*M(50)*	*R(50)*
Random	*0.056*	*-0.047*	*-0.122*	0.546	0.542	0.549	0.544	0.546		0.504	0.502	0.507	0.499	0.502	
1) N-WuPalmer	0.742	0.775	0.331	0.974	0.929	0.972	0.702	0.879	0.674	0.982	0.959	**0.981**	0.766	0.917	**0.763**
2) N-Leack.Chod.	0.724	**0.789**	0.295	0.953	0.901	0.954	0.702	0.863	0.648	0.984	0.953	**0.981**	0.757	0.913	0.755
3) N-Resnik	0.784	0.757	0.331	0.970	0.933	0.970	0.700	0.879	0.647	0.948	0.908	0.948	0.724	0.874	0.722
4) N-JiangConrath	0.719	0.588	0.175	0.956	0.872	0.920	0.645	0.817	0.458	0.931	0.857	0.911	0.625	0.808	0.570
5) N-Lin	0.754	0.619	0.204	0.949	0.884	0.918	0.682	0.822	0.451	0.939	0.877	0.920	0.611	0.827	0.566
6) W-NGD-Yahoo	*0.330*	0.445	0.254	0.940	0.907	0.941	0.783	0.885	0.648	—	—	—	—	—	—
7) W-NGD-Bing	*0.063*	*0.181*	*0.060*	0.724	0.706	0.713	0.650	0.690	0.600	0.659	0.619	0.671	0.633	0.648	0.633
8) W-NGD-GoogleWiki	*0.334*	0.502	0.251	0.874	0.837	0.872	0.703	0.814	0.649	—	—	—	—	—	—
9) C-BowDA	0.693	0.782	0.466	0.971	0.947	0.969	0.836	0.928	**0.772**	0.974	0.932	0.968	0.742	0.896	0.740
10) C-SynDA	**0.790**	0.786	0.491	0.985	0.953	**0.984**	0.811	0.925	0.749	0.978	0.945	0.972	0.751	0.907	0.743
11) C-LSA-Tasa	0.694	0.605	0.566	0.968	0.937	0.967	0.802	0.912	0.740	0.903	0.846	0.895	0.641	0.803	0.609
12) C-NGD-Factiva	0.603	0.599	**0.600**	0.959	0.916	0.959	0.786	0.894	0.681	0.906	0.857	0.904	0.731	0.835	0.543
13) C-PatternWiki	0.461	0.542	0.357	0.972	0.951	0.976	0.944	0.957	0.287	0.920	0.904	0.907	0.891	0.900	0.295
14) D-WktWiki	0.759	0.754	0.521	0.943	0.905	0.946	0.750	0.876	0.679	0.922	0.887	0.918	0.725	0.854	0.656
15) D-GlossVectors	0.653	0.738	0.322	0.894	0.860	0.901	0.742	0.843	0.686	0.932	0.899	0.933	0.722	0.864	0.709
16) D-ExtenedLesk	0.792	0.718	0.409	0.937	0.866	0.939	0.711	0.843	0.657	0.952	0.873	0.943	0.655	0.832	0.654
H-Mean-S8a	0.834	0.864	0.734	0.994	0.980	0.994	0.870	0.960	0.804	0.985	0.965	0.985	0.788	0.928	0.787
H-MeanZscore-S8a	0.830	0.864	0.728	0.994	0.981	0.993	0.874	0.961	0.808	0.986	0.967	0.986	0.793	0.932	0.792
H-MeanNnz-S8a	**0.843**	0.847	**0.740**	0.993	0.977	0.991	0.865	0.956	0.799	0.986	0.967	0.985	0.803	0.933	0.802
H-Median-S10	0.821	0.842	0.647	0.995	0.976	0.992	0.843	0.950	0.779	0.975	0.934	0.970	0.724	0.892	0.721
H-Max-S7	0.802	0.816	0.654	0.979	0.957	0.979	0.839	0.936	0.775	0.980	0.957	0.979	0.786	0.922	0.785
H-RankFusion-S10	—	—	—	0.994	0.978	0.993	0.864	0.956	0.798	0.976	0.929	0.971	0.745	0.896	0.744
H-RelationFusion-S10	—	—	—	0.996	0.982	0.995	0.840	0.952	0.758	0.986	0.963	0.981	0.781	0.920	0.749
H-Logit-E15	0.793	**0.870**	0.690	0.995	0.987	**0.995**	0.885	0.968	**0.818**	0.995	0.984	**0.993**	0.821	0.951	**0.819**
H-MeanNnz-E5	**0.878**	0.878	0.482	0.986	0.956	0.984	0.784	0.922	0.725	0.975	0.938	0.969	0.768	0.906	0.766
H-MeanZscore-S8b	0.844	**0.890**	0.616	0.992	0.977	0.991	0.844	0.953	0.780	0.995	0.985	0.995	0.815	0.950	0.814

Рис.: Характеристики 16 отдельных и 8 комбинированных метрик. MC, RG, WordSim353 – корреляция с суждениями человека. BLESS, SN – точность извлечения семантических отношений. Наилучшие значения в группе (отдельные/комбинированные) обозначены полужирным шрифтом; наилучшие значения обозначены серым цветом. Статистически незначимые корреляции ($p > 0.05$) обозначены курсивом, иначе $p \leq 0.05$.

Отдельные и комбинированные метрики

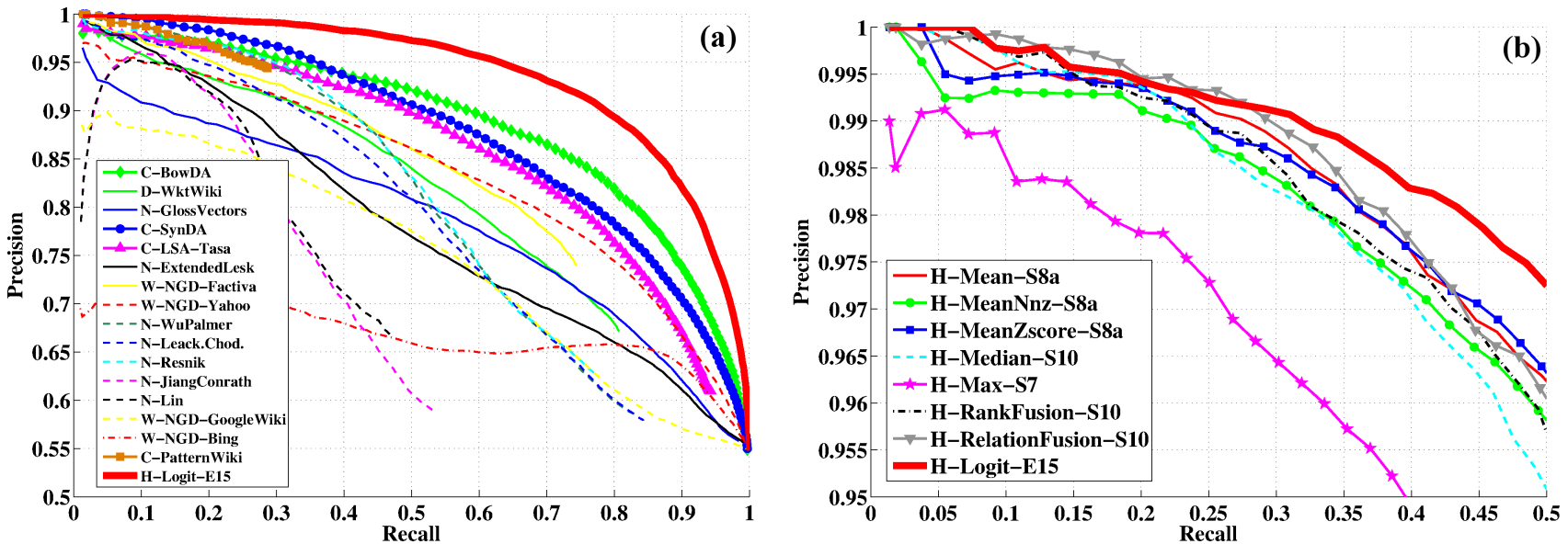


Рис.: График Точность-Полнота на множестве отношений BLESS для (a) 16 отдельных метрик и лучшей комбирированной метрике H-Logit-E15; (b) 8 комбинированных метрик.

Заключение:

Новизна работы:

1. Сравнение 34 отдельных метрик подобия на задаче извлечения семантических отношений
2. Разработка и сравнительный анализ 8 комбинированных методов извлечения отношений
3. Предложенный комбинированный метод H-Logit-E15, основанный на логистической регрессии:
   - Превосходит все отдельные и комбинированные методы
   - Достигает корреляции с суждениями человека до 0.870
   - Достигает MAP(20) при извлечения отношений до 0.995

Дальнейшие исследования:

Более сложные методы комбинирования:

- Методы комбинирования с учителем – **машины опорных векторов SVM** (Вапник, 1992)
- Методы комбинирования без учителя – **тензорные разложения PARAFAC, NTF** (Colda, 2004)

Приложения:

- Расширение поискового запроса
- Лексико-семантическая поисковая система