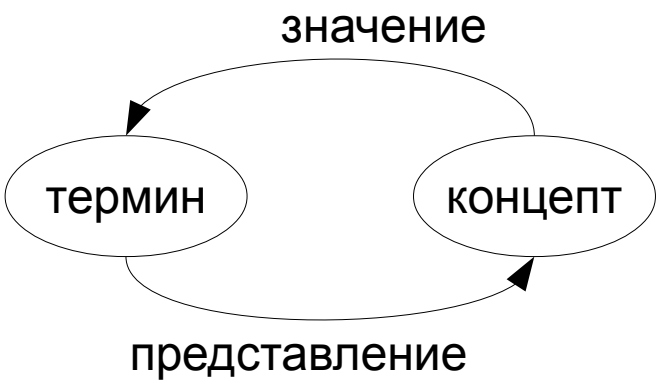
Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии

Варламов М.И., Коршунов А.В.

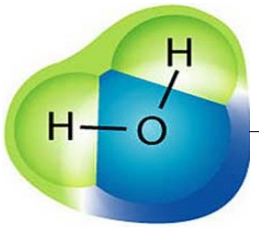
Институт системного программирования РАН

Семантическая близость концептов

- **Концепты** – различные понятия окружающего мира вообще или конкретной предметной области
- **Термины – слова** или словосочетания, являющиеся текстовыми представлениями концептов



- Некоторые пары концептов сильнее связаны по значению, чем другие
  - Сильнее ассоциируются друг с другом в сознании человека
  - Чаще встречаются в текстах в рамках одного контекста
- Мера **семантической близости** концептов – числовая оценка степени их смысловой связанности



сильно связаны

0.784

слабо связаны

0.094

вода *море* *компьютер*

Разрешение лексической многозначности

- **Разрешение лексической многозначности** – задача выбора правильного значения многозначного термина в зависимости от контекста
- Влияет на качество машинного перевода, информационного поиска и других задач
- Расчет семантической близости к терминам контекста может помочь определить верное значение многозначного термина

Система Текстерра

- Система анализа текстов **Текстерра (ИСП РАН)**
  - Позволяет находить в текстах термины и сопоставлять им концепты-значения
    - В частности, решает задачу разрешения лексической многозначности
  - База знаний строится на основе Википедии:
    - Концепты – статьи Википедии
    - Термины – названия статьи и тексты гиперссылок на статьи
    - Связи между концептами – гипертекстовые ссылки между статьями
  - Общее число концептов - более 5 миллионов, ссылок – более 250 миллионов
  - Граф ссылок Википедии имеет свойства безмасштабных (scale-free) сетей
    - Степени вершин распределены по степенному закону (число вершин степени k пропорционально величине $k^{-\gamma}$)
    - Малый диаметр – порядка O(loglogn), где n – число вершин в графе
  - Имеется открытый API на https://api.at.ispras.ru/

Система Текстерра

- Для оценки семантической близости в системе Текстерра используется взвешенная **мера Дайса**:

$$\mathrm{sim}_{\text{Дайс}}(x, y) = \frac{\sum_{z \in N(x) \cap N(y)} [w(x, z) + w(y, z)]}{\sum_{z \in N(x)} w(x, z) + \sum_{z \in N(y)} w(y, z)}$$

  - N(x) – множество соседей вершины (концепта) x в графе ссылок Википедии (включая x)
  - w(x,y) – вес ссылки между x и y (w(x,x) = 0)
- Данная мера равна 0, если пара концепций не имеет общих соседей в графе ссылок Википедии – а таких пар очень много!
- Возможно, семантическую близость концептов точнее оценивать на основе **расстояния (длины кратчайшего пути)** между ними в графе ссылок Википедии

Почему кратчайшие пути?

- Интуитивное предположение гласит: чем дальше концепты друг от друга, тем менее они семантически связаны и наоборот
- Скорее всего, такое предположение уже возникало, и не раз – *верно*
- *Так в чем же отличие нашего подхода?*
- Во-первых, мы используем *индексацию* графа ссылок Википедии для расчета расстояний между концептами
  - Расчет длин кратчайших путей влечет высокие расходы по памяти/времени
    - предварительный расчет и хранение растояний для всех пар – дорого по памяти (O(n^2))
    - обход графа при выполнении приложения – дорого по времени (O(n))
  - Часто такие меры локализуют, ограничивая глубину обхода графа, что ухудшает практические результаты
  - Использование специальных структур данных – индексов – позволяет снизить расходы на вычисление расстояний, не локализуя меру
- Во-вторых, Текстерра различает несколько типов ссылок Википедии, и мы разделяем их для расчета семантической близости концептов

Индексация графа ссылок Википедии

- Пусть $G=\langle V, E \rangle$ – граф
- Назовем меткой вершины u множество пар

$$L(u)=\{(w, \mathrm{dist}_G(u, w))\}_{w \in C(u)}, C(u) \subset V$$

- где $\mathrm{dist}_G(u, w)$ – расстояние между u и w в графе G

- Индекс графа – набор меток вершин, обеспечивающий *двухшаговое покрытие* (2-hop cover) графа, т.е. $\forall u, v \in V \quad \exists w \in C(u) \cap C(v)$: w принадлежит кратчайшему пути между u и v
- Тогда расстояние между вершинами u и w можно посчитать как

$$min\{\delta_{uw}+\delta_{wv} | (w, \delta_{uw}) \in L(u), (w, \delta_{wv}) \in L(v)\}$$

  - При упорядоченном хранении записей в метках вершин сложность вычисления расстояния – O(s), где s – средний размер метки одной вершины

Индексация графа ссылок Википедии

- Построение индекса можно реализовать методом *разметки вершин с отсечением* (Pruned Landmark Labeling) [1]
- "Полная" разметка:
  - обходим граф в ширину из каждой вершины
  - если из u достигли v за d шагов, добавляем (u,d) к метке вершины v
- Разметка с отсечением вершин
  - обходим граф в ширину из вершины u_n и за d шагов встречаем вершину v
  - если d можно посчитать с помощью текущего индекса:

$$d = min\{\delta_{uw} + \delta_{wv} | (w, \delta_{uw}) \in L(u_n), (w, \delta_{wv}) \in L(v)\}$$

    то *отсекаем* v (не продолжаем обход из нее)
- Существует открытая реализация, применимая на данный момент только к *неориентированным* графам

---

[1] Akiba, T., Iwata, Y., & Yoshida, Y. (2013, June). Fast exact shortest-path distance queries on large networks by pruned landmark labeling. In Proceedings of the 2013 international conference on Management of data (pp. 349-360). ACM.

Типы ссылок Википедии

- Внутритекстовые
- В инфобокс-секции
- Вида "Основная статья"
- В секции "См. также"
- Категорийные

Lomonosov Moscow State University (Russian: Московский государственный университет имени М. В. Ломоносова, *Moskóvskiy gosudárstvennyy universitét ímeni M. V. Lomonósova*), previously known as **Lomonosov University** or **MSU** (Russian: университет Ломоносова, *Universitét Lomonósova*; Russian: МГУ, *MGU*), is one of the oldest and largest universities in Russia. Founded in 1755, the university was renamed in honor of its founder, Mikhail Lomonosov, in 1940. It also claims to have the tallest educational building in the world. Its current rector is Viktor Sadovnichiy.

Contents [hide]

Staff and students [edit]

Currently the university employs more than 4,000 academics and 15,000 support staff. Approximately 5,000 scholars work at the university's research institutes and related facilities. More than 40,000 undergraduates and 7,000 advanced degree candidates are enrolled. More than 5,000 specialists participate in refresher courses for career enhancement. Annually, the university hosts approximately 2,000 students, graduate students, and researchers from around the world.

Lomonosov Moscow State University

Московский государственный университет имени М. В. Ломоносова

Coat of arms of the Lomonosov State University of Moscow

Latin: *Universitas Publica Moscuensis Lomonosoviana*

Motto	Наука есть ясное познание истины, просвещение разума (*Science is clear learning of truth and enlightenment of the mind*)
Established	1755
Type	Public
Rector	Viktor Sadovnichiy
Admin. staff	15,000
Students	47,000
Undergraduates	40,000
Postgraduates	7,000
Location	Moscow, Russia
Campus	urban
Affiliations	UNICA IFPU
Website	www.msu.ru

Типы ссылок Википедии

- Внутритекстовые
- В инфобокс-секции
- Вида "Основная статья"
- В секции "См. также"
- Категорийные

Famous alumni and faculty [edit]

Main article: List of Moscow State University people

Famous alumni of the Moscow State University [show]

11 Nobel laureates and 5 Fields Medal winners are affiliated with the university. It is the alma mater of many famous writers such as Anton Chekhov and Ivan Turgenev, politicians such as Mikhail Gorbachev or Mikhail Suslov, as well as renowned mathematicians and physicists such as Boris Demidovich, Vladimir Arnold, and Andrey Kolmogorov.

See also [edit]

- Seven Sisters (Moscow)
- Education in Russia
- List of early modern universities in Europe
- List of universities in Russia
- List of rectors of the Moscow State University

Moscow portal
University portal

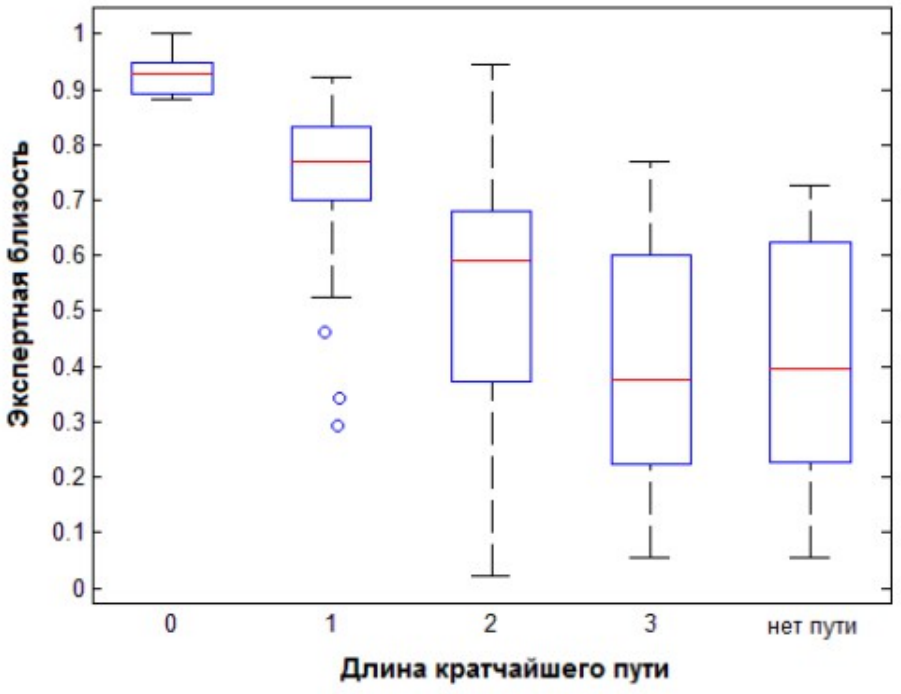
Categories: Schools of international relations | Moscow State University alumni | 1755 establishments | 1780s architecture | Buildings and structures built in the Soviet Union | Moscow State University | Educational institutions established in the 1750s | Education in Russia | Education in the Soviet Union | Skyscrapers between 200 and 249 meters | Stalinist architecture | Visitor attractions in Moscow | Seven Sisters (Moscow) | Skyscrapers in Moscow | Towers in Moscow | Universities and colleges in Moscow

- В мере Дайса, используемой в системе Текстерра, наибольшие веса имеют ссылки в секции "См. также" и ссылки вида "Основная статья"

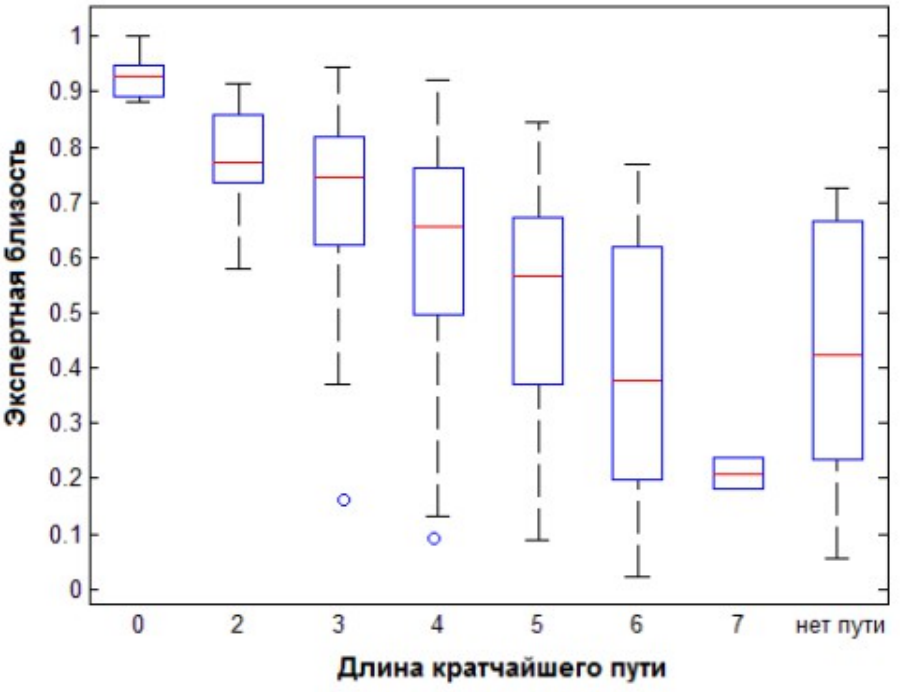
Зависимость экспертной близости от расстояния по различным типам ссылок

- Экспертные данные – набор WordSim-353
  - Стандартный набор данных для оценки качества мер семантической близости
  - Содержит 353 пары *слов* с экспертными значениями близости
  - Слова в наборе мы заменили идентификаторами соответствующих статей Википедии

Внутритекстовые ссылки

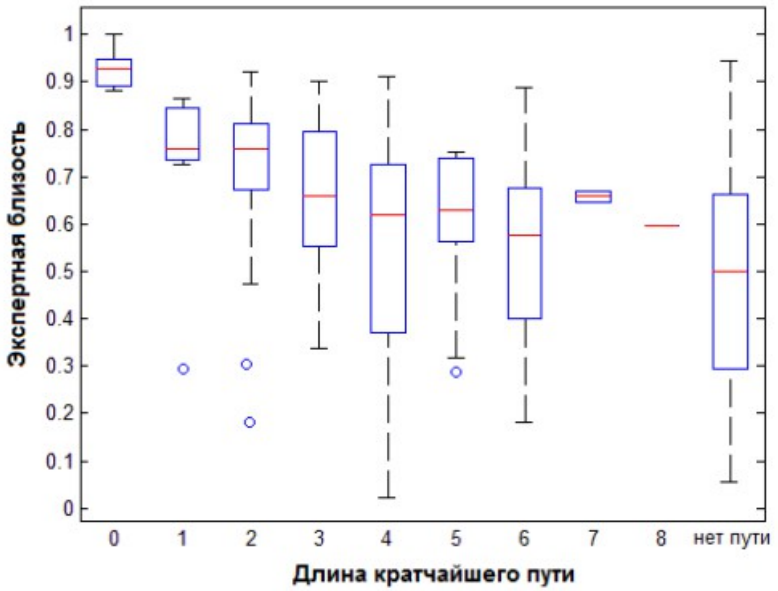


Категорийные ссылки

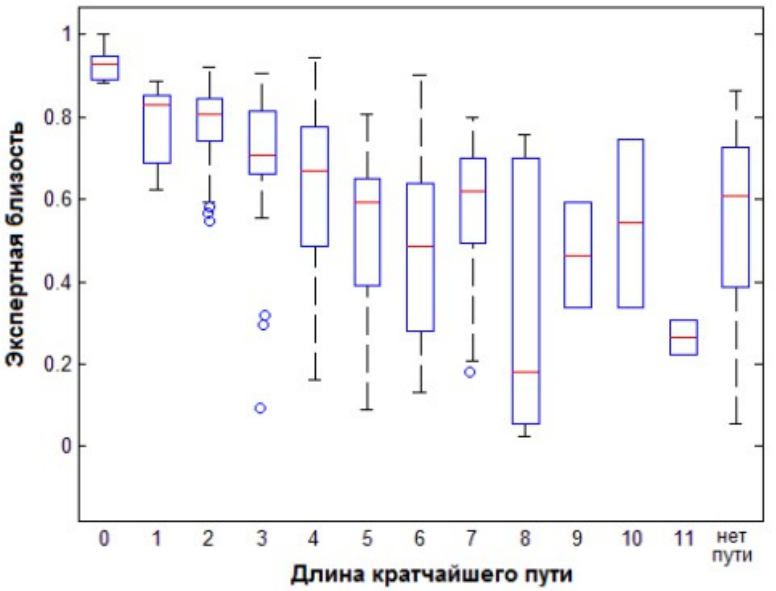


Зависимость экспертной близости от расстояния по различным типам ссылок

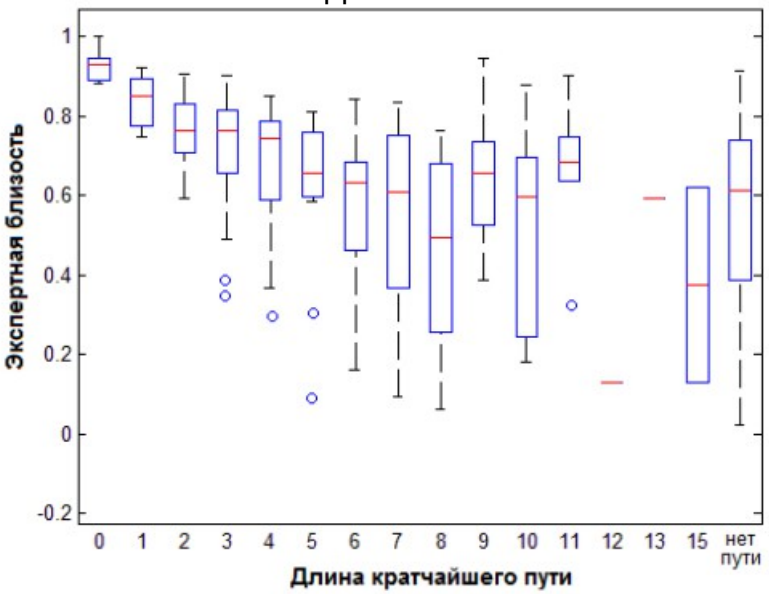
Ссылки в инфобокс-секции



Ссылки в секции "См. также"



Ссылки вида "Основная статья"



- Для обычных и категорийных ссылок близость монотонно убывает с увеличением расстояния
- Подграфы по различным типам ссылок имеют разные диаметры
- Малые диаметры могут указывать на то, что подграф Википедии по данному типу ссылок – безмасштабный

Исследование корреляции с экспертной близостью

- Для каждого типа ссылок T была рассмотрена мера близости:

$$\mathrm{sim}_T(x, y) = \frac{1}{1 + \mathrm{dist}_T(x, y)}$$

где $\mathrm{dist}_T(x, y)$ - расстояние между x,y по ссылкам типа T

- Для каждой меры была рассчитана корреляция с экспертными оценками на адаптированном наборе данных WordSim-353:

Мера сем. близости	**Корреляция с экспертными оценками**		
	по Пирсону	**по Спирмену**	**по расстоянию**
Мера Дайса	0.3376	**0.6804**	0.4901
Внутритекстовые ссылки	**0.5072**	0.6059	**0.5409**
Категорийные ссылки	0.4252	0.5695	0.4805
Ссылки в инфобоксах	0.3633	0.3986	0.3684
Ссылки в секции „См. также“	0.3850	0.4593	0.4273
Ссылки вида „Основная статья“	0.3388	0.3013	0.3209

- Мера на основе внутритекстовых ссылок лучше коррелирует с экспертными оценками, чем мера Дайса, по коэффициентам Пирсона и расстояния

Построение комбинации мер близости по отдельным типам ссылок

- Можно ли улучшить корреляцию мер по отдельным типам ссылок путем их комбинации?
- Рассмотрим линейную комбинацию отдельных мер близости, включая меру Дайса:

$$\mathrm{sim}(x, y) = \sum_{\mathrm{p} \in \{\text{типы ссылок}\}} w_{\mathrm{p}} \mathrm{sim}_{\mathrm{p}}(x, y) + w_{\text{Дайс}} \mathrm{sim}_{\text{Дайс}}(x, y)$$

- Для поиска весов w по набору данных WordSim-353 построили множество относительных ограничений R:

$$R = \{(x, y, z) : \mathrm{sim}_{\text{эксперт}}(x, y) > \mathrm{sim}_{\text{эксперт}}(x, z)\}$$

Построение комбинации мер близости по отдельным типам ссылок

- Поиск весов w: пассивно-агрессивный алгоритм
  - Инициализация: $w^0 = (0, \ldots, 0)$
  - На шаге t
    - Сэмплируем $(x_i, x_j, x_k) \in R$
    - Обновляем w согласно

$$w^t = \operatorname{argmin} \frac{1}{2} \|w - w^{t-1}\|^2 + C\,\xi$$

$$1 - (\operatorname{sim}_w(x_i, x_j) - \operatorname{sim}_w(x_i, x_k)) \leq \xi$$

  - Проводили по 10^6 итераций с C=0.01
- Похожий подход использовался для построения меры OASIS [2]

---

[2] Chechik, G., Shalit, U., Sharma, V., & Bengio, S. (2009). An Online Algorithm for Large Scale Image Similarity Learning. In NIPS (pp. 306-314).

Построение комбинации мер близости по отдельным типам ссылок

- Корреляция с экспертными оценками (со скользящим контролем по 10 блокам)

Мера сем. близости	**Корреляция с экспертными оценками**		
	по Пирсону	**по Спирмену**	**по расстоянию**
Мера Дайса	0.3376	0.6804	0.4901
Лин. комб. без меры Дайса	**0.5388**	0.6390	**0.6183**
Лин комб с мерой Дайса	0.5193	**0.6853**	0.5856

  - Использование кратчайших путей позволило улучшить показатели меры Дайса относительно всех коэффициентов корреляции

- Полученные веса для различных типов ссылок

Тип ссылок	**Внутрит.**	**Категор.**	**Инфобокс**	**См. также**	**Осн. статья**
Вес	**0.3307**	**0.5843**	0.0317	0.0406	0.0126

  - Наиболее ценную информацию для расчета семантической близости предоставляют внутритекстовые и категорийные ссылки

Сравнение с существующими мерами на основе кратчайших путей

- Мера Цеша [3]

$$\mathrm{sim}_{\text{Цеш}}(x, y) = \mathrm{D}_{\text{произв.ссылки}} - \mathrm{dist}_{\text{произв.ссылки}}(x, y)$$

- Мера Ликока-Чодороу [4]

$$\mathrm{sim}_{\text{Л-Ч}}(x, y) = -\log \frac{\mathrm{dist}_{\text{катег.ссылки}}(x, y)}{2\,\mathrm{D}_{\text{катег.ссылки}}}$$

Мера сем. близости	**Корреляция с экспертными оценками**		
	по Пирсону	**по Спирмену**	**по расстоянию**
Мера Цеша	0.5192	0.5893	0.5406
Мера Ликока-Чодороу	**0.5473**	0.5655	0.5580
Лин комб без меры Дайса	0.5388	**0.6390**	**0.6183**

[3] Zesch T., Müller C., Gurevych I. Using Wiktionary for Computing Semantic Relatedness //AAAI. – 2008. – Т. 8. – С. 861-866.

[4] Strube M., Ponzetto S. P. WikiRelate! Computing semantic relatedness using Wikipedia //AAAI. – 2006. – Т. 6. – С. 1419-1424.

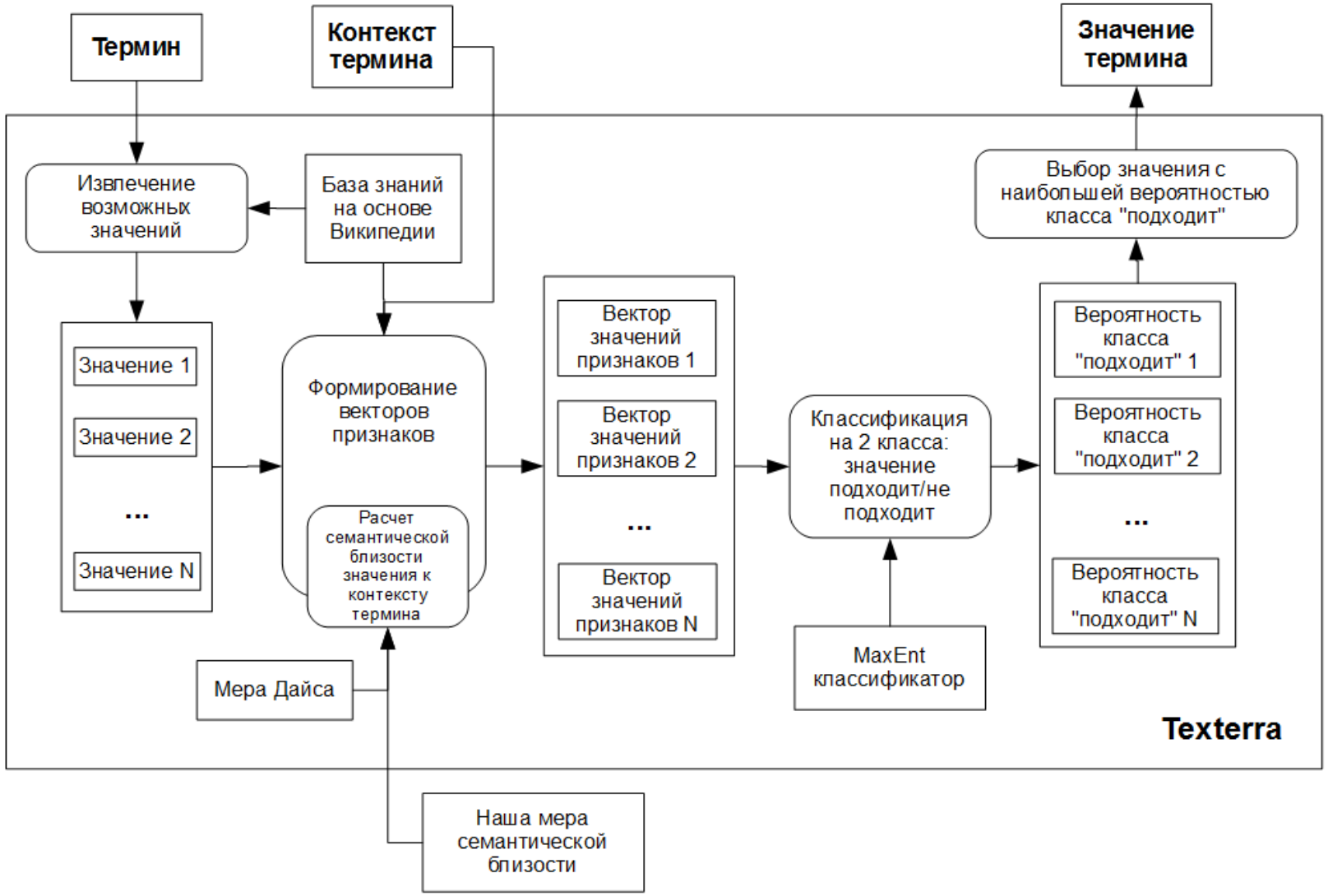
Тестирование разрешения лексической многозначности

- В системе Текстерра для разрешения лексической многозначности используется классификатор на основе метода максимальной энтропии
  - Один из признаков – семантическая близость предполагаемого значения термина к его однозначному контексту
    - Однозначный контекст термина – все термины текста, для которых в базе знаний системы имеется ровно один концепт-значение
  - Другие признаки
    - Вероятность того, что термин ссылается на статью Википедии
    - Частота концепта в Википедии
    - Качество однозначного контекста
- Данные – 4 стандартных набора системы Текстерра для тестирования разрешения лексической многозначности:

Название	**Число текстов**	**Число концептов**	**Тематика**
Board games	35	13410	Статьи о настольных играх
AQUAINT	50	13974	Новостные статьи
MODIS-texts	131	25061	Технические статьи
Wiki	100	104401	Статьи Википедии

Тестирование разрешения лексической многозначности

- Схема работы метода системы Текстерра:



Тестирование разрешения лексической многозначности

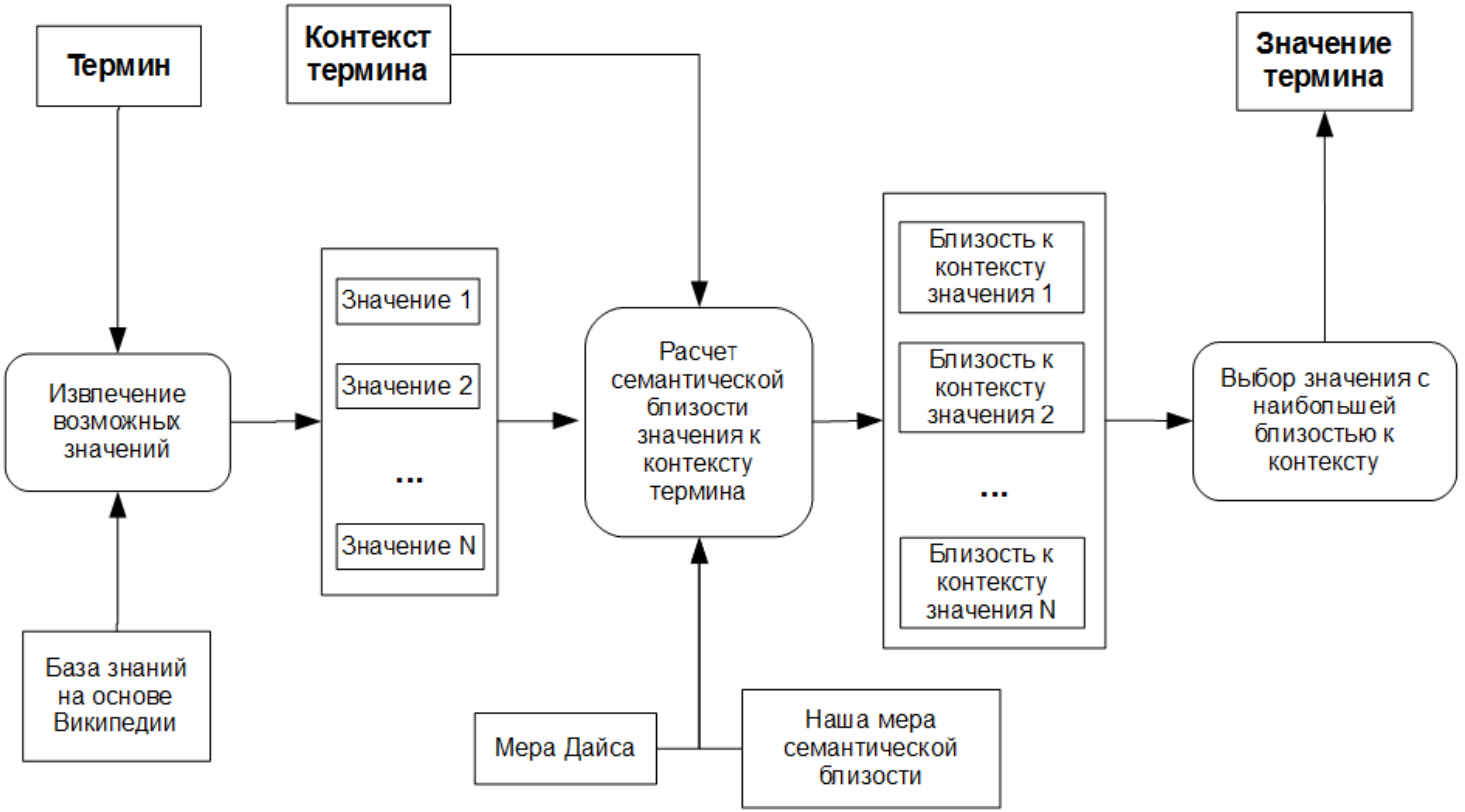
- Точность (P), полнота (R) и F-мера (F) разрешения лексической многозначности методом системы Текстерра в зависимости от используемой меры семантической близости:

		Мера Дайса	Внутритекст. ссылки	Категор. ссылки	Лин. комб. без меры Дайса	Лин. комб. с мерой Дайса
Board games	**P**	**0.8305**	0.5624	0.6265	0.6191	0.7278
	R	**0.3678**	0.2491	0.2773	0.2740	0.3221
	F	**0.5095**	0.3450	0.3842	0.3796	0.4463
AQUAINT	**P**	**0.8784**	0.8581	0.8626	0.8626	0.8649
	R	**0.5991**	0.5853	0.5883	0.5883	0.5899
	F	**0.7123**	0.6959	0.6995	0.6995	0.7014
MODIS-texts	**P**	**0.8203**	0.6166	0.6997	0.6746	0.7577
	R	**0.4407**	0.3312	0.3759	0.3624	0.4071
	F	**0.5734**	0.4309	0.4890	0.4715	0.5296
Wiki	**P**	**0.9199**	0.5390	0.6541	0.5805	0.7425
	R	**0.6017**	0.3525	0.4278	0.3796	0.4856
	F	**0.7274**	0.4262	0.5173	0.4590	0.5872

Мера Дайса показывает лучшие результаты, в сравнении с лучшей мерой на основе кратчайших путей точность выше в среднем на 13%, полнота – на 6%, F-мера – на 8%

Тестирование разрешения лексической многозначности

- Другие признаки классификатора могут ослаблять влияние признака на основе семантической близости
- Рассмотрим *"наивный"* метод разрешения лексической многозначности, использующий только семантическую близость предполагаемого значения термина к его контексту



Тестирование разрешения лексической многозначности

- Точность (P), полнота (R) и F-мера (F) разрешения лексической многозначности наивным методом в зависимости от используемой меры семантической близости:

		Мера Дайса	Внутритекст. ссылки	Категор. ссылки	Лин. комб. без меры Дайса	Лин. комб. с мерой Дайса
Board games	**P**	0.7657	**0.7676**	0.6167	0.7235	0.7282
	R	0.3391	**0.3400**	0.2731	0.3205	0.3225
	F	0.4701	**0.4712**	0.3786	0.4442	0.4471
AQUAINT	**P**	0.6689	**0.8086**	0.5766	0.7523	0.7072
	R	0.4562	**0.5515**	0.3932	0.5131	0.4823
	F	0.5425	**0.6557**	0.4676	0.6100	0.5735
MODIS-texts	**P**	0.7103	**0.7775**	0.6189	0.7054	0.7011
	R	0.3816	**0.4177**	0.3325	0.3790	0.3767
	F	0.4965	**0.5435**	0.4326	0.4931	0.4901
Wiki	**P**	0.8978	**0.9241**	0.7601	0.8791	0.9004
	R	0.5880	**0.6053**	0.4978	0.5758	0.5898
	F	0.7106	**0.7314**	0.6016	0.6958	0.7127

Лучшие результаты показывает мера по внутритекстовым ссылкам, в сравнении с мерой Дайса точность выше в среднем на 6%, полнота – на 4%, F-мера – на 5%

Заключение

- Кратчайшие пути в графе Википедии содержат полезную информацию для вычисления семантической близости
- Наиболее полезными типами ссылок для расчета семантической близости между концептами являются внутритекстовые и категорийные
- Использование меры семантической близости на основе кратчайших путей обеспечивает лучшее качество разрешения лексической многозначности по сравнению с мерой Дайса в отстутствие других признаков
- Возможные направления дальнейшей работы:
  - Исследование путей в ориентированном графе Википедии
  - Сравнение с мерами на основе случайного блуждания

Спасибо за внимание!