Серелекс: поиск и визуализация семантически связанных слов

Панченко А.И.[1,2], Романов П.В.[2], Романов А.В[1],
Филиппович А.Ю.[2], Филиппович Ю.Н.[2], Морозова О.Д.[1]

[1] Université catholique de Louvain, Лувен, Бельгия
[2] МГТУ им. Н. Э. Баумана, Москва, Россия

Аннотация. Данная статья представляет Серелекс, систему, которая по запросу на английском языке предоставляет список семантически сязанных слов. Слова ранжируются в соответствии с оригинальной метрикой семантической близости, обученной на корпусе естественно-языковых текстов. Точность работы системы сравнима с аналогами основанными на WordNet и других словарях. При этом, наша система использует только информацию извлеченную непосредственно из текстов. Наше исследование показывает, что пользователи полностью удовлетворены результатами поиска семантически связанных слов в 70% случаев.

Ключевые слова: метрика семантической близости; визуализация семантических отношений.

1 Введение

Мы представляем Серелекс, систему, которая по заданному английскому слову, возвращает список слов отсортированных по семантической близости с запросом [1]. Программа помогает изучить

[1] Данная статья является расширенной версией [1].

значение иностранных слов и интерактивно исследовать связанные лексические единицы. В отличие от аналогичных систем основанных на словарях и тезаурусах, таких как `Thesaurus.com` или `VisualSynonyms.com`, Серелекс использует информацию извлечённую из корпуса естественно-языковых текстов. По сравнению с другими подобными системами, такими как BabelNet [2], ConceptNet [3] или UBY [4], Серелекс не зависит от семантических ресурсов, таких как WordNet. Система использует оригинальную метрику семантической близости между словами основанную на лексико-синтаксических шаблонах [2]. Согласно нашим экспериментам, точность использованного подхода сопоставима с метриками предложенными в предыдущих исследованиях. Кроме этого, представляемая система обеспечивает большее лексическое покрытие, чем аналоги основанные на словарях, предоставляет три альтернативных способа визуализации результатов запросов (в виде списка, графа или набора изображений) и имеет открытый исходный код.

2 Система

Серелекс находится в открытом доступе в Интернет [5]. Система состоит из экстрактора, сервера и пользовательсокого интерфейса (см. Рис. 1). Задача экстрактора заключается в извлечении семантических отношений между словами из корпуса естестенно-языковых текстов на английском языке. Извлечённые отношения сохраняются в базе данных. Сервер обеспечивает быстрый доступ к извлечённым отношениями по HTTP. Пользователь взаимодействует с системой через веб-интерфейс или API. Исходный код системы, данные и скрипты оценки качества работы доступны на условиях лицензии LGPLv3 [6].

Экстрактор. Подсистема извлечения семантических отношений основана на метрике семантической близости *PatternSim* и формуле ранжирования *Efreq-Rnum-Cfreq-Pnum* [2]. Данная метрика основана на лексико-синтактических шаблонах подобных тем, которые были предложены в [3]. Данные паттерны извлекают из корпуса текстов множество конкордансов, таких как:

- `such diverse {[occupations]} as {[doctors]}, {[engineers]} and {[scientists]}`
- `such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}`

[2] `http://lcl.uniroma1.it/bnxplorer/`
[3] `http://conceptnet5.media.mit.edu/`
[4] `https://uby.ukp.informatik.tu-darmstadt.de/webui/tryuby/`
[5] `http://serelex.cental.be` или `http://serelex.it-claim.ru`
[6] `http://serelex.cental.be/page/about`

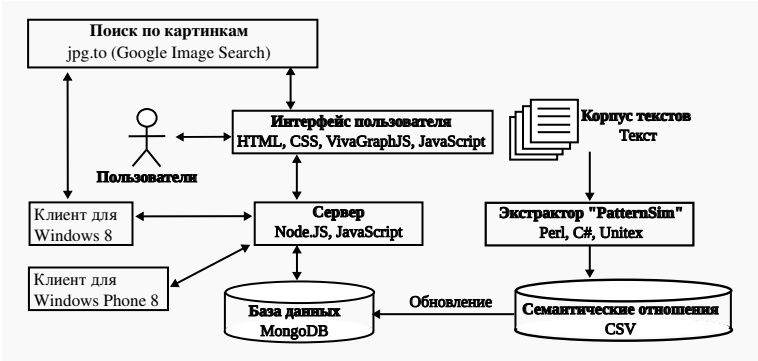

Рис. 1. Архитектура системы.

Название	# Документов	# Словоформ	# Лемм	Размер
Википедия	2,694,815	2,026 · 10^9	3,368,147	5.88 Гб
ukWaC	2,694,643	0.889 · 10^9	5,469,313	11.76 Гб
Википедия + ukWaC	5,387,431	2.915 · 10^9	7,585,989	17.64 Гб

Таблица 1. Корпуса текстов использованные системой.

- `{traditional[food]}, such as {[sandwich]}, {[burger]}, and {[fry]}`
- `{[mango]},{[pineapple]}, {[jackfruit]} and other{[fruit]}`
- `{primitive [snake]}, such as {[boa]} and {[python]}`
- `{[France]},{[ Belgium]} and other {European [country]}`

Семантическое сходство двух слов пропорционально количеству конкордансов в которых они совместно встретились. Однако окончательное значение семантической близости вычисляется с учетом и других факторов, таких как частота слов в корпусе и количество извлеченных отношений для каждого из слов [2]. Мы произвели извлечение из коллекции текстовых документов состоящей из заголовков статей Википедии и корпуса ukWaC [4] (см. Таблицу 1). Обработка данного корпуса заняла около 72 часов на стандартной рабочей станции (Intel i5, 4Гб ОЗУ, HDD 5400 об/мин). В результате извлечения было выявлено 11,251,240 нетипизированных семантических отношений, таких как $\langle Canon, Nikon, 0.62\rangle$, между 419,751 леммами.

Сервер. Сервер возвращает множество связанных слов для каждого запроса, отсортированных согласно их семантической близости, сохранённой в базе данных. Запросы лемматизируются при помощи

словаря DELA [7]. Для запросов на которые не нашлось ни одного результата выполняется приблизительный поиск с помощью расстояния Левенштейна. Система позволяет импортировать семантические отношения сохраненные в формате CSV, которые были извлечены альтернативными экстракторами.

Пользовательский интерфейс. Система предоставляет доступ через веб-интерфейс, с помощью приложений для платформы Windows, либо через RESTful веб-сервис. Веб интерфейс состоит из трёх основных элементов: строки поиска, списка результатов и графа результатов (см. Рис. 2). Пользователь взаимодействует с системой формулируя поисковый запрос - слово значение которого требуется узнать. Такой запрос может быть выражен односложным словом, таким как “mathematics”, либо словосочетанием, таким как “computational linguistics”.

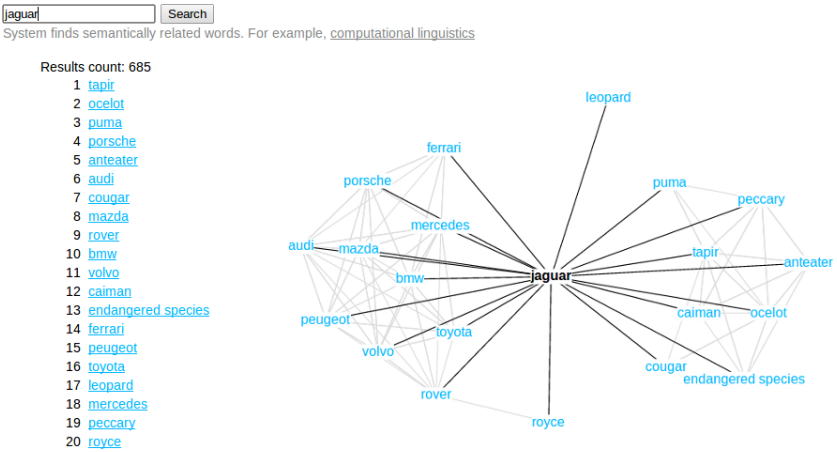

Рис. 2. Визуализация результатов поиска.

Кроме графового интерфейса пользователя, реализован интерфейс, основанный на изображениях. При этом, всю рабочую область занимают графическое представление слов, связанных с результатами поиска. Мы используем изображения предоставляемые веб сервисом jpg.to [8]. По клику на изображение происходит переход к словам, семантически связанным с нажатым.

[7] `http://infolingu.univ-mlv.fr/`, доступно на условиях лицензии LGPLLR.

[8] `http://jpg.to/about.php`. Данный сервис использует Google Image Search: `http://images.google.ru/`.

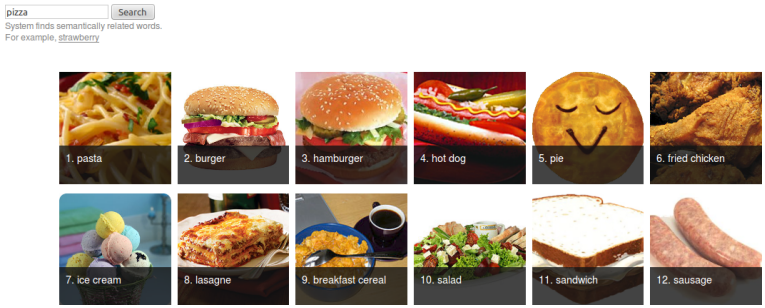

Рис. 3. Интерфейс основанный на изображениях.

В дополнение к веб-интерфейсу были разработаны приложения для Windows 8 [9] и Windows Phone 8 [10]. Данные клиенты используют веб-сервис Серелекса для получения результатов запросов и сервис jpg.to для получения изображений (см. Рис. 1). Приложения выполненны с учетом рекомендаций по построеноию пользовательского интерфейса приложений для Windows и Windows Phone, а их исходный код является открытым [11]. В рамках создания приложений для Windows была создана переносимая бибилиотека классов (Portable Class Library), которая может быть полезна для доступа к веб-сервису Серелекса из сторонних приложений. Отличительной особеностью клиента системы для Windows Phone является то, что оно позволяет сразу же выполнить поиск в Google по результатам запроса (см. Рис. 5).

3 Результаты

Мы оценили качество работы системы на четырех задачах кратко описанных ниже. Подробное описание данных экспериментов приведено в [2].

3.1 Корреляция с суждениеми о семантической близости

Мы использовали три проверочных набора данных широко распространеных в англоязычной литературе по лексической семантике: *MC* [5], *RG* [6] и *WordSim* [7]. Данные коллекции содержат множе-

[9] `http://apps.microsoft.com/windows/app/lsse/48dc239a-e116-4234-87fd-ac90f030d72`
[10] `http://www.windowsphone.com/s?appid=dbc7d458-a3da-42bf-8da1-de49915e0318`
[11] `https://github.com/jgc128/Serelex4Win`

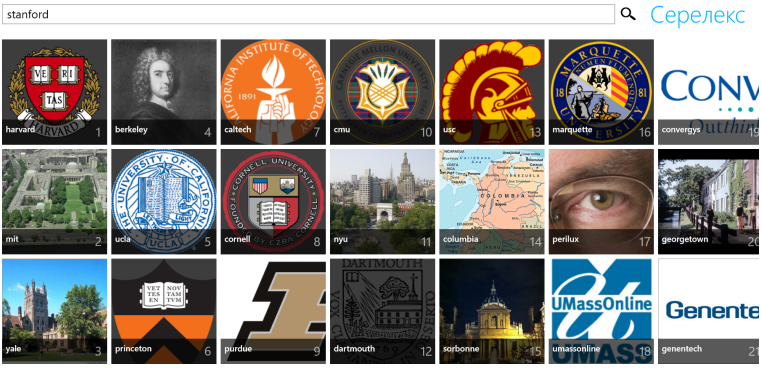

Рис. 4. Клиент системы для платформы Windows 8.

ство пар слов. Для каждой такой пары вручную задана мера их семантической близости:

- `automobile; car; 3.92`
- `brother; monk; 2.84`
- `glass; magician; 0.11`

Согласно результатам наших экспериментов, корреляция Спирмена между значениями семантической близости предоставляемыми системой и суждениями субъектов достигает 0.665, 0.739 и 0.520 соответственно для *MC*, *RG* и *WordSim*. Данные характеристики предложенной системы сравнимы с показателями разработанных ранее метрик семантической близости основанными на WordNet (*WuPalmer* [8], *LeacockChodorow* [9], *Resnik* [10]), словарях (*ExtendedLesk* [11], *GlossVectors* [12], *WiktionaryOverlap* [13]) и корпусах текстов (*ContextWindow* [14], *SyntacticContext* [14], *LSA* [15]).

3.2 Ранжирование семантических отношений

В данном тесте нужно отсортировать некоторое множество слов по семантической близости с заданным словом. Например, задано 50 слов, 25 из которых связано со словом “alligator”, в то время как другие 25 слов не связаны. Задача системы заключается в том чтобы так ранжировать слова чтобы семантически связанные термины имели более высокий ранг:

- `1; alligator; animal (related)`
- `...`
- `25; alligator; lizard (related)`
- `26; alligator; twin (random)`
- `...`
- `50; alligator; electronic (random)`

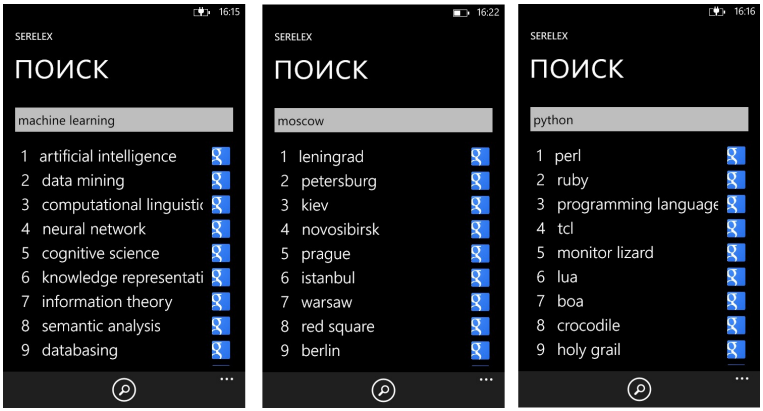

Рис. 5. Клиент системы для Windows Phone 8.

Данный тест основан на наборе семантических отношений BLESS [16] и SN [17]. Точность *Серелекса* на данной задаче сопоставима с 9 описанными выше альтернативными метриками, однако полнота серьезно ниже, в связи с разреженностью подхода основанного на шаблонах (см. Рис 6 (а)).

3.3 Извлечение семантических отношений

Кроме двух описанных выше тестов, мы также оценили точность извлечения семантических отношений для 49 слов из лексикона *RG*. В данном эксперименте трем асессорам было предложено аннотировать результаты поиска для данных 49 запросов. Каждый субъект указал для каждого из 50 первых результатов поиска является ли он релевантным или нет. Например, для запроса “fruit”:

— `1; vegetable (relevant)`
— `2; mango (relevant)`
— ...
— `50; house (non-relevant)`

На основании полученной статистики мы вычислили точность в k, где $k \in \{1, 5, 10, 20, 50\}$. Согласно результатам данного эксперимента приведенным на Рис 6 (б), средняя точность извлечения варьируется между 74% (для первого результата, $k = 1$) и 56% (50 первых результатов, $k = 50$). Мы зафиксировали значительную степень согласия асессоров для данного эксперимента в терминах каппы Флейса (0.61-0.80).

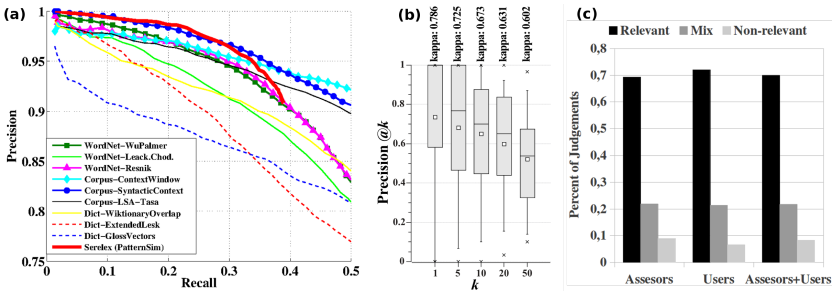

Рис. 6. Результаты: (а) задача ранжирования семантических отношений; (б) задача извлечения семантических отношений; (в) удовлетворенность пользователей первыми 20 результатами поиска.

3.4 Удовлетворенность пользователей качеством поиска

Каждому из 23-х асессоров участвующих в исследовании было предложенно выбрать 20 запросов по своему усмотрению и оценить первые 20 результатов поиска как релевантные, нерелевантные либо как частично релевантные. В результате данной кампании оценки, мы собрали 460 суждений асессоров и 233 суждения анонимных пользователей системы. Пользователи и асессоры вместе осуществили 594 уникальных запроса. В соответствии с этим экспериментом, результаты поиска являются релевантными для 70% запросов и нерелевантными для 10% запросов (см. Рис 6 (в)). Наконец, в 20% случаев первые 20 результатов содержат смесь релевантных и нерелевантных слов.

4 Выводы

Мы представили систему, которая производит поиск семантически связанных слов. Оценка качества работы системы на четырех задачах показала что точность системы сопоставима с аналогичными разработками предложенными ранее. При этом, в отличие от большинства аналогов, Серелекс не использует составленные вручную словари. За счет этого достигается лучшее лексическое покрытие, т.к. семантические отношения извлекаются непосредственно из текста. Наконец, анкетирование пользователей показало что первый результат поиска релевантен в 74% случаях, а для 70% запросов пользователи полностью удовлетворены первыми 20 результатами.

Список источников

1. Panchenko, A., Romanov, P., Morozova, O., Naets, H., Romanov, A., Philippovich, A., Fairon, C.: Serelex: Search and visualization of semantically related words. In Proceedings of the 35th European Conference in Information Retrieval (2013)
2. Panchenko, A., Morozova, O., Naets, H.: A semantic similarity measure based on lexico-syntactic patterns. In: Proceedings of KONVENS 2012. (2012) 174–178
3. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: ACL. (1992) 539–545
4. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: A collection of very large linguistically processed web-crawled corpora. LREC **43**(3) (2009) 209–226
5. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the workshop on Human Language Technology, ACL (1993) 303–308
6. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. ACM **8**(10) (1965) 627–633
7. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: WWW 2001. (2001) 406–414
8. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: ACL'1994. (1994) 133–138
9. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet (1998) 265–283
10. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: IJCAI. Volume 1. (1995) 448–453
11. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: IJCAI. Volume 18. (2003) 805–810
12. Patwardhan, S., Pedersen, T.: Using WordNet-based context vectors to estimate the semantic relatedness of concepts. Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together (2006) 1–12
13. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from wikipedia and wiktionary. In: LREC'08. (2008) 1646–1652
14. Van de Cruys, T.: Mining for Meaning: The Extraction of Lexico-Semantic Knowledge from Text. PhD thesis, University of Groningen (2010)

15. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3) (1998) 259–284
16. Baroni, M., Lenci, A.: How we blessed distributional semantic evaluation. In: GEMS (EMNLP), 2011. (2011) 1–11
17. Panchenko, A., Morozova, O.: A study of hybrid similarity measures for semantic relation extraction. EACL 2012 (2012) 10