

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., Морозова О.Д.¹

¹ Université catholique de Louvain, Лувен, Бельгия

² МГТУ им. Н. Э. Баумана, Москва, Россия

Аннотация. Мы представляем Serelex, систему, которая по запросу на английском языке, предоставляет список семантически связанных слов. Слова ранжируются в соответствии с оригинальной семантической мерой сходства, полученной из большого корпуса. Система выполняет сравнение, основанное на словаре, но не требует никаких семантических ресурсов, таких как WordNet. Наше исследование показывает, что пользователи полностью удовлетворены результатами в 70% случаев

Ключевые слова: семантическая мера сходства; визуализация; извлечение.

1 Вступление

Мы представляем Serelex, систему, которая, по данному английскому слову, возвращает список связанных слов, отсортированных согласно семантической близости. Система помогает изучать значение термина запроса и интерактивно исследовать связанные слова. В отличие от словарей и тезаурусов, например Thesaurus.com или

VisualSynonyms.com, Serelex полагается на информацию, извлечённую из корпусов текстов. По сравнению с другими подобными системами (например BabelNet ¹, ConceptNet ², UBY ³), Serelex не зависит от семантических ресурсов, таких как WordNet. Вместо этого мы полагаемся на оригинальную, основанную на образцах, меру семантической близости [1]. Система имеет точность, сопоставимую с теми из 9 базовых линий. Кроме того, имеет большее лексическое покрытие, чем системы, основанные на словаре, обеспечивает интерфейс пользователя в виде списка, графа и основанного на изображениях, и является системой с открытым исходным кодом.

2 Система

Serelex полностью бесплатно доступна в сети интернет ⁴. На рис. 1 изображена архитектура системы, которая состоит из экстрактора, сервера и пользовательского интерфейса. Экстрактор извлекает семантические отношения между словами из необработанного корпуса текстов. Извлечённые отношения сохраняются в базе данных. Сервер обеспечивает быстрый доступ к извлечённым отношениям по HTTP. Пользователь взаимодействует с системой через веб-интерфейс или API. Система, а так же данные и скрипты оценки имеют открытый исходный код ⁵, доступный под лицензией LGPLv3.

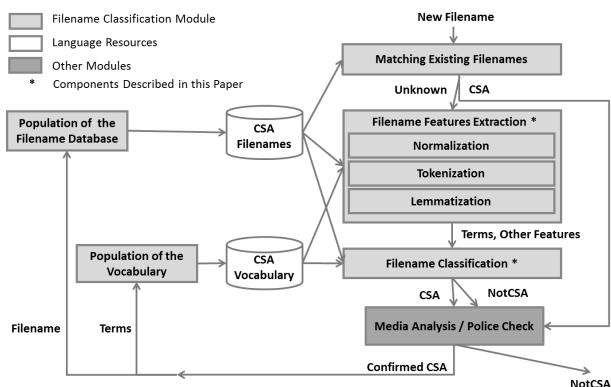


Рис. 1. Архитектура системы

¹ <http://lcl.uniroma1.it/bnexplorer/>

² <http://conceptnet5.media.mit.edu/>

³ <https://uby.ukp.informatik.tu-darmstadt.de/webui/tryuby/>

⁴ <http://serelex.cental.be>

⁵ <http://serelex.cental.be/page/about>

Экстрактор. Экстрактор основан на семантической мере подобия *PatternSim* и формуле ранжирования *Efreq-Rnum-Cfreq-Pnum* [1]. Эта, основанная на корпусе мера, полагается на лексико-синтаксические шаблоны, которые извлекают конкордансы. Сходство пропорционально числу со-возникновений термина в конкордансах, например: `such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}`. Оценка нормализуется с частотой термина и другими извлечёнными статистиками [1]. В качестве корпуса мы использовали комбинацию кратких обзоров Википедии и ukWaC [2] (5 387 431 документов, $2.915 \cdot 10^9$ токенов, 7,585,989 лемм, 17.64 Гб). Обработка корпуса заняла около 70 часов на обычной машине (Intel i5, 4Gb RAM, HDD 5400rpm). В результате выделено 11 251 240 нетипизированных семантических отношений (например, *<Canon, Nikon, 0.62>*) между 419,751 словами.

Сервер. Сервер возвращает список связанных слов для каждого запроса, отсортированных согласно их семантическому сходству, сохранённому в базе данных. Запросы лемматизируются при помощи словаря DELA ⁶. Для запросов без результатов выполняется приблизительный поиск. Система может импортировать сети в формате CSV, созданные другими метриками сходства и экстракторами.

Пользовательский интерфейс. К системе можно получить доступ через веб-интерфейс или RESTfull API. Графический интерфейс пользователя состоит из трёх основных элементов: поле поиска, список с результатами и граф результатов (см рис. 2). Пользователь взаимодействует с системой, вводя запрос – одно слово, например “mathematics”, или несколько слов, например “computational linguistics”.

Кроме графового интерфейса пользователя, реализован интерфейс, основанный на изображениях. При это, всю рабочую область занимают графическое представление слов, связанных с данным. Изображения получаются с помощью сервиса jpg.to ⁷. По клику на изображение происходит переход к словам, семантически связанным с нажатым.

Так же, были разработаны приложения для Windows 8 ⁸ и Windows Phone ⁹. Данные приложения используют RESTfull API для получения результатов запроса пользователя и также используют сервис jpg.to для получения изображений и выполнены с учетом ре-

⁶<http://infolingua.univ-mlv.fr/>, доступно под лицензией LGPLLR.

⁷<http://jpg.to/about.php>

⁸<http://apps.microsoft.com/windows/app/1s8e/48dc239a-e116-4234-87fd-ac90f030d72>

⁹<http://www.windowsphone.com/s?appid=dbc7d458-a3da-42bf-8da1-de49915e0318>

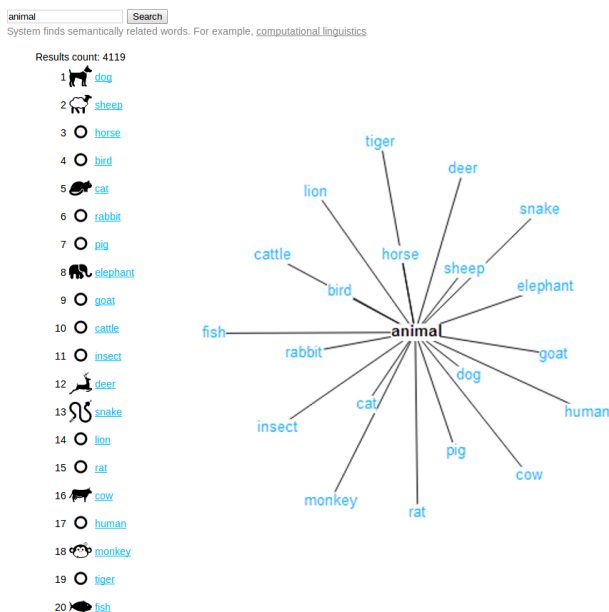


Рис. 2. Графический интерфейс пользователя

комендаций по построению пользовательского интерфейса приложений для Windows и Windows Phone. В рамках создания приложений для Windows 8 была создана переносимая библиотека классов (Portable Class Library), а, так как исходные коды приложения для Windows 8 доступны на Github ¹⁰, это предоставляет возможность сторонним программистам создавать свои приложения на платформе .NET 4.5, использующие сервис Serelex, без написания своего кода для доступа к RESTful API. Отличительной особенностью приложения для Windows Phone является то, что оно позволяет сразу же выполнить поиск в Google по результатам запроса благодаря наличию специальной кнопки.

Оценки и результаты. Мы оценивали систему, исходя из четырех задач (см. [1] для подробностей): 1. **Корреляция с человеческим суждением.** Мы использовали стандартные наборы данных для измерения корреляции Спирмена с человеческим суждением. Наша система сравнивается с базовыми результатами, включающими 3 метрики, основанные на WordNet (WuPalmer [3],

¹⁰<https://github.com/jgc128/Serelex4Win>

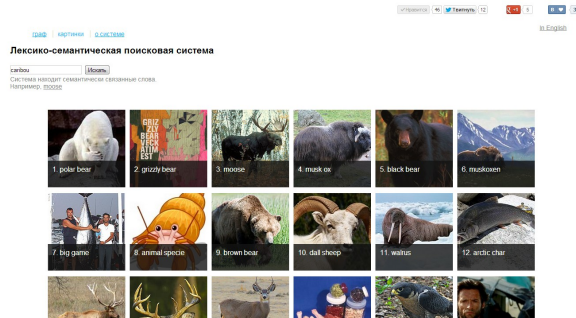


Рис. 3. Интерфейс, основанный на изображениях

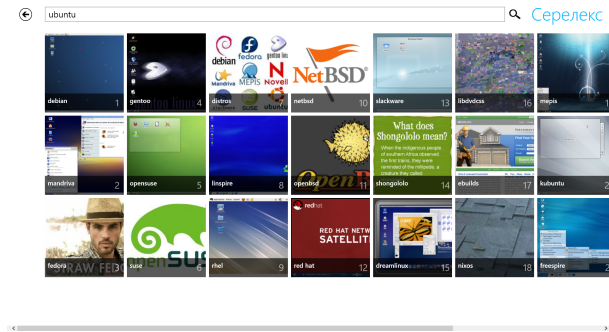


Рис. 4. Скриншот приложения для Windows 8

LeacockChodorow [4], *Resnik* [5]), 3 основанные на словарях (*ExtendedLesk* [6], *GlossVectors* [7], *WiktioaryOverlap* [8]), и 3 метрики, основанные на корпусах (*ContextWindow* [9], *SyntacticContext* [9], *LSA* [10]). 2. **Ранжирование семантических отношений.** Эта задача опирается на набор семантических отношений (BLESS, SN) для оценивания *относительной* точности и recall каждой метрики. Точность *Serelex* сопоставима с 9 базовыми метриками, но её recall серьезно ниже, в связи с разреженностью подхода, основанного на шаблонах (см. Рис 6 (а)). 3. **Извлечение семантических отношений.** Мы оценивали точность извлеченных отношений для 49 слов (словарь надора данных RG). Три аннотатора указывали, связаны ли термины семантическими отношениями, или нет. Каждому из них было предложено отметить, релевантны ли первые 50 результатов, или нет. Мы вычислили точность извлечения как $k = \{1, 5, 10, 20, 50\}$.

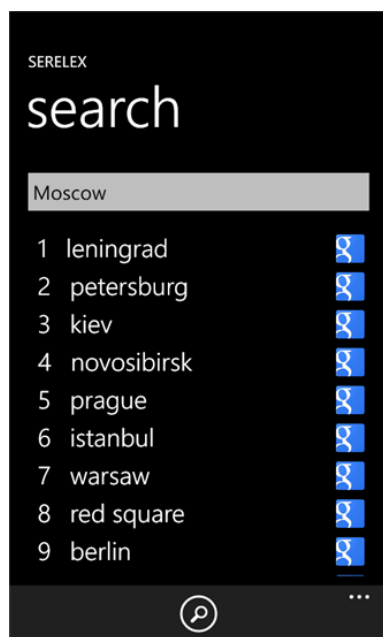


Рис. 5. Скриншот приложения для Windows Phone

Средняя точность варьируется между 0,736 для первых результатов, и 0,599 для первых 50 результатов (см. Рис 6 (b)). The inter-raters agreement в терминах каппы Флейса значительное (0.61-0.80). 4. **Удовлетворение пользователей.** Мы также измеряли удовлетворение пользователей от наших результатах. 23 экспертам было предложено выбрать 20 запросов по своему усмотрению и ранжировать первые 20 результатов как релевантные, нерелевантные и частично релевантные для каждого из запросов. Мы собрали 460 решений экспертов и 233 решения анонимных пользователей (см. Рис 6 (с)). Пользователи и эксперты (пользователей просили воспользоваться системой) вместе создали 594 уникальных запросов. В соответствии с этим экспериментов, результаты были релевантны в 70% случаях, и нерелевантны в 10% случаях. Наконец, в 20% запросов (recall) были релевантные и нерелевантные результаты.

3 Выводы

Мы представили систему, которая находит семантически связанные слов. Наши результаты показывают точность, сопоставимую с подходом, основанным на словарях и обладают лучшим покрытием, так как отношения извлекаются непосредственно из текста. Система достигает точности@1 в около 74%, и удовлетворения пользователей в 70% результатов запросов без необходимости какого-либо ручного составления словаря.

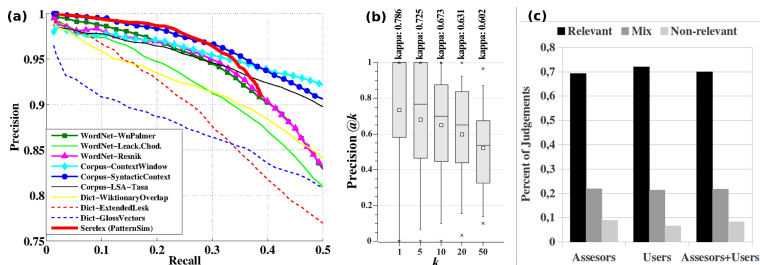


Рис. 6. Оценка: (a) precision-recall граф задачи ранжирования семантических отношений в BLESS; (b) задача извлечения семантических отношений; (c) удовлетворение пользователей от первых 20 результатах

Список источников

1. Panchenko, A., Morozova, O., Naets, H.: A semantic similarity measure based on lexico-syntactic patterns. In: Proceedings of KONVENS 2012. (2012) 174–178
2. 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
3. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: ACL'1994. (1994) 133–138
4. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet (1998) 265–283
5. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: IJCAI. Volume 1. (1995) 448–453
6. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: IJCAI. Volume 18. (2003) 805–810
7. 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

8. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from wikipedia and wiktionary. In: LREC'08. (2008) 1646–1652
9. Van de Cruys, T.: Mining for Meaning: The Extraction of Lexico-Semantic Knowledge from Text. PhD thesis, University of Groningen (2010)
10. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. *Discourse processes* **25**(2-3) (1998) 259–284