"Serelex: Search and Visualization of Semantically Related Words"

Panchenko, Alexander ; Fairon, Cédrick ; Naets, Hubert ; Morozova, Olga

Abstract

We present a system which provides, given a query, a list of semantically related terms. The terms are ranked accordingly to an original semantic similarity measure learned from a huge corpus. The system performs comparably to dictionary-based baselines with no need of any semantic resource such as WordNet. The further study shows that users are completely satisfied with 70% of query results.

Serelex: Search and Visualization of Semantically Related Words

Alexander Panchenko[1,2], Pavel Romanov[2], Olga Morozova[1], Hubert Naets[1], Andrey Philippovich[2], Alexey Romanov[2], and Cédrick Fairon[1]

[1] Université catholique de Louvain, Louvain-la-Neuve, Belgium
{Firstname.Lastname}@uclouvain.be
[2] Bauman Moscow State Technical University, Moscow, Russia
{promanov,aphilippovich,aromanov}@it-claim.ru

Abstract. We present Serelex, a system that provides, given a query in English, a list of semantically related words. The terms are ranked according to an original semantic similarity measure learnt from a huge corpus. The system performs comparably to dictionary-based baselines, but does not require any semantic resource such as WordNet. Our study shows that users are completely satisfied with 70% of the query results.

Keywords: semantic similarity measure, visualization, extraction.

1 Introduction

We present *Serelex*, a system that, given a query in English, returns a list of related terms ranked according to a *semantic similarity measure*. The system helps to learn the meaning of a query term and to discover semantically similar words in an interactive way. Unlike dictionaries and thesauri (e.g., `Thesaurus.com` or `VisualSynonyms.com`), *Serelex* relies on information extracted from text corpora. In comparison to other similar systems (e.g., BabelNet [3], ConceptNet [4], UBY [5]), *Serelex* does not depend on a semantic resource like WordNet. Instead, we build upon an original pattern-based similarity measure [1]. The proposed system has a precision rate comparable to those of the 9 baselines. Furthermore, it has a larger lexical coverage than the dictionary-based systems, provides list-, graph-, and image-based GUIs, and is open source.

2 The System

Serelex is freely available online [6]. Figure 1 presents its structure, which consists of an extractor, a server and a user interface. The extractor gathers semantic

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

relations between words from a raw text corpus. The extraction process occurs offline. The extracted relations are stored in the database. The server provides fast access to the extracted relations over HTTP. A user interacts with the system through a web interface or an API. The system as well as the data and the evaluation scripts are open source [7].

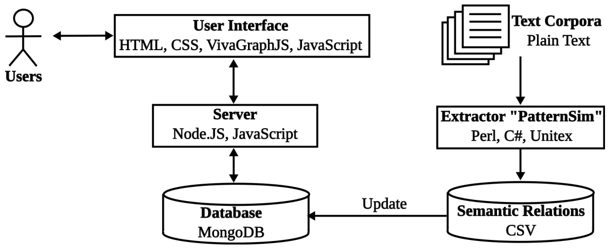


Fig. 1. Structure of the system *Serelex*.

Extractor. The extractor is based on the semantic similarity measure *PatternSim* and *Efreq-Rnum-Cfreq-Pnum* re-ranking formula [1]. This corpus-based measure relies on handcrafted lexico-syntactic patterns which extract concordances. Similarity score is proportional to the number of term co-occurrences within those concordances, e.g.: `such` {`non-alcoholic [sodas]`} `as` {`[root beer]`} `and` {`[cream soda]`}. The score is normalized with term frequencies and other extraction statistics [1]. We used as a corpus a combination of Wikipedia abstracts and ukWaC [2] (5,387,431 documents, $2.915 \cdot 10^9$ tokens, 7,585,989 lemmas, 17.64 Gb). Processing of the corpus took around 70 hours on a standard machine (Intel i5, 4Gb RAM, HDD 5400rpm). The result of the extraction is 11,251,240 untyped semantic relations (e.g., $\langle Canon, Nikon, 0.62\rangle$) between 419,751 terms.

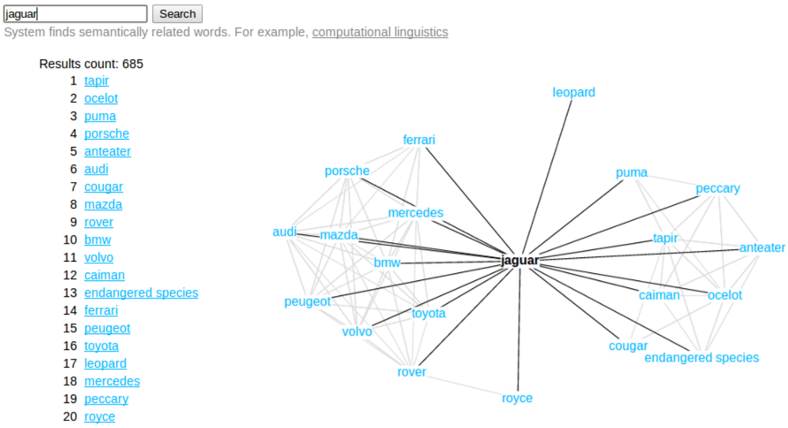
Server. The server returns a list of related words for each query, ranked according to their semantic similarity measure stored in the database. The queries are lemmatized with the DELA dictionary [8]. An approximate search is performed for queries with no results. The system can import networks in CSV format created by other similarity metrics and extractors.

User Interface. One can access the system via a graphical user interface or a RESTful API. The GUI consists of three key elements: a search field, a list of the results and a graph of the results (see Fig. 2). A user interacts with the system by issuing a query – a single word such as "mathematics" or a multiword expression such as "computational linguistics". Query suggestions are sorted at the same time by term frequency in the corpus, by query frequency, and alphabetically. A list of results contains 20 terms which are the most semantically related to the query. The graph of results provides an alternative representation of the toplist. It enables visualization of semantic relations with a force-directed graph layout

[7] `http://serelex.cental.be/page/about`, available under conditions of LGPLv3.

[8] `http://infolingu.univ-mlv.fr/`, available under conditions of LGPLLR.

algorithm based on the Barnes-Hut simulation [3]. The layout incorporates the secondary relations: words related to the words linked to the query. This lets the layout algorithm cluster the results. A user can issue additional queries by clicking on the nodes. Alternatively, the system can visualize the search results as a set of images.



jaguar Search

System finds semantically related words. For example, computational linguistics

Results count: 685

1 tapir
2 ocelot
3 puma
4 porsche
5 anteater
6 audi
7 cougar
8 mazda
9 rover
10 bmw
11 volvo
12 caiman
13 endangered species
14 ferrari
15 peugeot
16 toyota
17 leopard
18 mercedes
19 peccary
20 royce

Fig. 2. Graphical user interface of the *Serelex* system: results of the query "jaguar".

3 Evaluation and Results

We evaluated the system against four tasks (see [1] for details):

1. **Correlation with Human Judgements.** We used standard datasets (MC, RG, WordSim) to measure Spearman's correlation with human judgements. Our system performs comparably to the baselines, that includes 3 WordNet-based measures (*WuPalmer* [4], *LeacockChodorow* [5], *Resnik* [6]), 3 dictionary-based measures (*ExtendedLesk* [7], *GlossVectors* [8], *WiktionaryOverlap* [9]), and 3 corpus-based measures (*ContextWindow* [10], *SyntacticContext* [10], *LSA* [11]).

2. **Semantic Relation Ranking.** This task relies on a set of semantic relations (BLESS, SN) to estimate *relative* precision and recall of each measure. The precision of *Serelex* is comparable to the 9 baselines, but its recall is seriously lower due to the sparsity of the pattern-based approach (see Figure 3 (a)).

3. **Semantic Relation Extraction.** We estimated the precision of the extracted relations for 49 words (the vocabulary of the RG dataset). Three annotators indicated whether the terms are semantically related or not. Each of them was asked to label each result from the top 50 as relevant or irrelevant. We calculated extraction precision at $k = \{1, 5, 10, 20, 50\}$. Average precision varies between 0.736 for the top relation and 0.599 for the top 50 (see Figure 3 (b)). The inter-raters agreement in terms of Fleiss's kappa is substantial (0.61-0.80).

4. **User Satisfaction.** We also measured user satisfaction with our results. 23 assessors were asked to issue 20 queries of their choice and, for each of them, to rank the top 20 results as relevant, irrelevant, or a mix of both. We collected 460 judgements from the 23 assessors and 233 judgements from 109 anonymous users (see Fig. 3 (c)). Users and assessors (users asked to assess the system) issued together 594 distinct queries. According to this experiment, the results are relevant in 70% of the cases and irrelevant in 10% of the cases. Finally, 20% of queries recall both relevant and irrelevant results.

4 Conclusion

We presented a system which finds semantically related words. Our results have shown that it has a precision comparable to the dictionary-based baselines and a better coverage as it extracts relations directly from texts. The system achieves a Precision@1 of around 74%, and users are satisfied with 70% of the query results without the need for any manually-crafted dictionary.

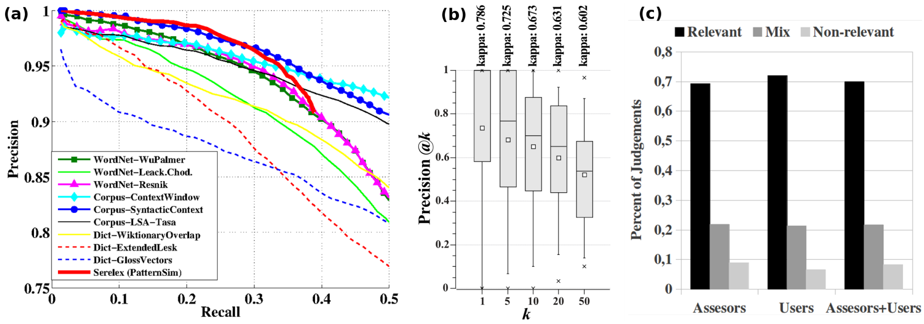


Fig. 3. Evaluation: (a) precision-recall graph of the semantic relation ranking task on BLESS; (b) semantic relation extraction task; (c) users' satisfaction of top 20 results.

References

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. Barnes, J., Hut, P.: A hierarchical 0 (n log iv) force-calculation algorithm. nature **324** (1986) 4
4. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: ACL'1994. (1994) 133–138

5. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet (1998) 265–283
6. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: IJCAI. Volume 1. (1995) 448–453
7. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: IJCAI. Volume 18. (2003) 805–810
8. 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
9. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from wikipedia and wiktionary. In: LREC'08. (2008) 1646–1652
10. Van de Cruys, T.: Mining for Meaning: The Extraction of Lexico-Semantic Knowledge from Text. PhD thesis, University of Groningen (2010)
11. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3) (1998) 259–284