Künstl Intell
DOI 10.1007/s13218-015-0417-5

TECHNICAL CONTRIBUTION

Extracting Semantics from Unconstrained Navigation on Wikipedia

Thomas Niebler[1] · **Daniel Schlör**[1] · **Martin Becker**[1] · **Andreas Hotho**[1]

Received: 15 May 2015 / Accepted: 7 November 2015

Abstract Semantic relatedness between words has been successfully extracted from navigation on Wikipedia - pages. However, the navigational data used in the corresponding works are sparse and expected to be biased since they have been collected in the context of games. In this paper, we raise this limitation and explore if semantic relatedness can also be extracted from unconstrained navigation. To this end, we first highlight structural differences between unconstrained navigation and game data. Then, we adapt a state of the art approach to extract semantic relatedness on Wikipedia paths. We apply this approach to transitions derived from two unconstrained navigation datasets as well as transitions from WikiGame and compare the results based on two common gold standards. We confirm expected structural differences when comparing unconstrained navigation with the paths collected by WikiGame. In line with this result, the mentioned state of the art approach for semantic extraction on navigation data does not yield good results for unconstrained navigation. Yet, we are able to derive a relatedness measure that performs well on both unconstrained navigation data as well as game data. Overall, we show that unconstrained navigation data on Wikipedia is suited for extracting semantics.

✉ Thomas Niebler
niebler@informatik.uni-wuerzburg.de

Daniel Schlör
schloer@informatik.uni-wuerzburg.de

Martin Becker
becker@informatik.uni-wuerzburg.de

Andreas Hotho
hotho@informatik.uni-wuerzburg.de

[1] University of Wuerzburg, Würzburg, Germany

Keywords Usage analysis · Semantic web

1 Introduction

The ever increasing amount of digital content, for example on the World Wide Web, requires intelligent access in order to extract useful information. To this end, semantic relatedness of natural language, can help to learn ontologies used to build knowledge graphs for the Semantic Web. Semantic relatedness between words has been extracted from a variety of sources [11]. Especially, Wikipedia has been in the focus of corresponding approaches because it "is not just a regular website but a rich network representing human knowledge as well as the connections between single pieces of knowledge, by means of hyperlinks" [8]. Since such links are an indicator of relatedness, navigation data on Wikipedia are interesting for studying emerging semantics and have been exploited before [6, 9].

Problem setting The navigation data used for extracting semantic relatedness in the mentioned work are not directly extracted from Wikipedia. Instead, they are recorded in the context of games, like WikiGame[1] and WikiSpeedia[2]. Both games provide environments where users have to navigate on Wikipedia from a given source page to a given destination page using existing links only. Yet, this setting induces unnatural navigation patterns: Fig. 1 shows a possible path taken by a user who has to navigate from the source concept "Asteroid" to the destination concept "Franks". Such a navigational pattern might not be observed without explicit incentives. This is, because even if a user is interested in both "Asteroids" and "Franks",

[1] http://www.thewikigame.com.

[2] http://www.wikispeedia.net.

Published online: 16 December 2015

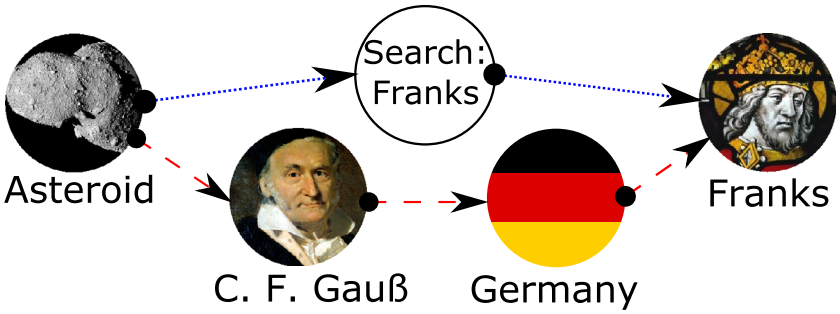

Fig. 1 Schematic example of real navigation (*blue*, *dotted line*) using e.g. a search page to reach the goal vs. game navigation (*red*, *dashed path*) only following hyperlinks

she is more likely to use the search function (as indicated by the blue arrows) instead of finding a path from "Asteroids" to "Franks" via links. Thus, datasets from a game setting can considerably differ from navigation data generated by unconstrained navigation behavior, which possibly biases studies based on such datasets. At the same time, datasets from game environments are sparse compared to unconstrained navigation data. That is, in contrast to game data, which is dependent on the amount of players interested in the game, unconstrained navigation data on Wikipedia is collected every day and can be extracted from request logs. To address these limitations regarding structural bias and to increase the amount of data available to study, we investigate if semantic relatedness can also be extracted from unconstrained navigation data.

Approach In order to explore if semantic relatedness can be extracted from unconstrained navigational data, we first highlight structural differences between unconstrained navigation and game data. We then adopt a state of the art approach to extract semantic relatedness on Wikipedia paths by Singer et al. [6]. We apply this approach to transitions derived from two unconstrained navigation datasets as well as transitions from WikiGame and compare the results based on two common gold standards, namely WS-353 and MEN.

Contribution and Findings We confirm expected structural differences when comparing unconstrained navigation with the paths collected by WikiGame. In line with this result, the mentioned state of the art approach for semantic extraction on navigation data does not yield good results for unconstrained navigation. Yet, we are able to derive a relatedness measure that performs well on both unconstrained navigation data as well as game data. In the case of only considering transitions, our adopted approach even outperforms the state of the art. Overall, we show that unconstrained navigation data on Wikipedia is suited for extracting semantics.

Structure. The paper is structured as follows: In Sect. 2, we present an overview on related work. We then give a quick description of the used data (Sect. 3) and our methodology (Sect. 4), followed up by our results in Sect. 5.

2 Related Work

In this section, we give a short overview of related work connected to research on semantic relatedness and Wikipedia navigation analysis. In general, semantic relatedness between words has been successfully extracted from a variety of sources [11]. Yet, in the following, we focus on work regarding semantic extraction from Wikipedia and summarize several articles specifically addressing semantics in the context of navigation data as well as closely related work on navigation analysis.

Extraction of Semantic Information Since Wikipedia is a big collection of carefully written articles on specific subjects, it is widely used as a corpus for research on semantic relatedness. One of the most well-known works concerning semantic relatedness on Wikipedia was done by Gabrilovich and Markovitch [3]. They propose the ESA measure, which calculates semantic relatedness between Wikipedia concepts based on TF-IDF vectors and correlate their resulting relatedness ranking to the WS-353 dataset. The potential of Wikipedia's category taxonomy for calculating semantic relatedness is shown by Strube and Ponzetto [7] and compared with several baseline approaches using WordNet. They show that methods using the category structure of Wikipedia outperform Google count based methods and a WordNet baseline. However, they obtain the best result using Wikipedia, Google and WordNet in combination. Omitting the Wikipedia category-taxonomy, Milne and Witten make use of the Wikipedia hyperlink structure in order to calculate semantic relatedness introducing the Wikipedia Link-based Measure (WLM) [5]. They use a measure similar to a TF-IDF based measure and the Google distance measure, and evaluate a combination of both measures on article-link sets obtained from Wikipedia in comparison to WikiRelate and ESA. All of these methods work only on the static link structure of Wikipedia and do not take user navigation into account.

User and navigation analysis West et al. created a game called Wikispeedia to supervise user navigational behaviour [9]. They used the data collected from this game to examine the taken paths when provided with a navigation task. It was

Table 1 Base statistics for all experiment datasets

	WikiGame	WikiClickIU	WikiStream	WikiLink
#requests	62.5M	4.0M	1090.2M	n/a
#links	2.3M	2.8M	14.4M	494.2M
#un. trgts	357,068	1.3M	2.2M	18.9M
#un. srcs	330,693	789,734	1.4M	29.0M

WikiLink does not contain any requests by design

shown that most navigation followed a "zoning-out-homing-in" pattern. They also proposed a new semantic relatedness measure for word pairs encountered in paths, but it only can measure relatedness between concepts if they appeared on a path together. In [8], the authors examine a bigger dataset from Wikispeedia than in [9] and develop a method to predict the next target in a path. Their findings in human navigation analysis support their previous results, which they successfully apply to target prediction, what in turn could be used to improve website navigation design. Singer et al. [6] calculate semantic relatedness on a path dataset from the WikiGame. They also try out different path subsets to find out which factors exert the greatest influence on the performance of their method. Lastly, in [10], West et al. extend their previous method of target prediction to additionally predict the source of a possible navigation to ultimately enhance linkage between Wikipedia articles. For their analyses, they use a dump of the WikiGame as well as Wikispeedia for evaluation, which was carried out automatically as well as through human judgment. In all of these works, only game data is examined, without taking the potential bias into account, that a game setting might induce.

3 Datasets

We use two types of datasets in this study. That is, experiment datasets containing navigation data from which we extract semantic information and evaluation datasets comprised of word pairs with a human-assigned similarity score which we use to evaluate our findings.

3.1 Experiment Datasets

We compare three request datasets on Wikipedia (*WikiGame*, *WikiClickIU*, *WikiStream*), and use the link network of Wikipedia (*WikiLink*) to derive serveral baseline datasets. All three request datasets contain transitions from a source page to a target page. *WikiGame* contains game navigation, while the other two datasets consist of unconstrained navigation. *WikiClickIU* was extracted from university network traffic and *WikiStream* is taken directly from Wikipedia. Table 1 lists the basic statistics of our experiment datasets.

WikiGame The *WikiGame*[3] is a competitive navigation game on the articles of Wikipedia. A game instance is given by two randomly drawn Wikipedia pages, the *source* and the *target* page. The dataset at hand contains all played games from Feb, 17th 2009 till Sept, 12th 2011. We extracted all possible transitions between two successively clicked pages.

WikiClickIU The authors of [4] made available a large dataset of about 53.5 billion HTTP requests by users at Indiana University between 2008 and 2010 to study structure and dynamics of Web traffic networks[4]. We extracted all requests originating from and targeting the Wikipedia domain, thus retaining about 4 million requests with 1.3 million distinct target Wikipedia content pages. We name this *WikiClickIU*.

WikiStream Wulczyn published a clickstream dataset directly from Wikipedia, which we call *WikiStream*[5]. This dataset does not contain single requests, but only lists request counts from or to pages on Wikipedia in February 2015. Observations with less than 10 occurrences have been removed by Wulczyn. We only use the requests with source and target inside of Wikipedia.

WikiLink *WikiLink*[6] is a snapshot of the Wikipedia link network from January 2015, which is freely available. We include this dataset to evaluate if the semantic information we extract from unconstrained navigation data actually results from the observed user requests or if it is merely a result from the underlying link structure. To this end we build four baseline datasets based on WikiLink and WikiStream: (i) *binary*, (ii) *restricted*, (iii) *random sampling*, and (iv) *distribution sampling*. (i) The most simple baseline (*binary*) interprets each link in the network as a single request. This represents the information provided solely by the link network. (ii) Now, we want to know if the selection users make by (not) visiting certain Wikipedia pages contains semantic information. Thus, we build the *restricted* baseline by removing all requests from the binary dataset whose source page is not observed as a source page in WikiStream. (iii) + (iv) Additionally, we assume that (not) using certain outlinks from a source page also contains semantic information. Thus, we build two sampled datasets based on the *restricted* baseline. For the *random sampling* approach, we randomly sample outlinks (not individual requests) without replacement from the *restricted* dataset set until we have the same number of overall observed outlinks as in WikiStream. For the *distribution sampling*

[3] http://www.thewikigame.com.

[4] http://cnets.indiana.edu/groups/nan/webtraffic/click-dataset/.

[5] http://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/.

[6] https://dumps.wikimedia.org/enwiki/20150112/.

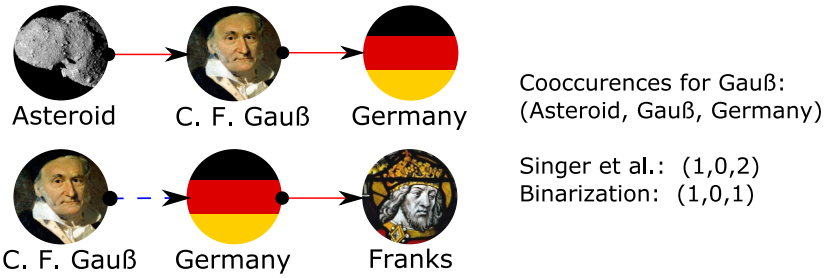

Fig. 2 Illustrative example of the semantic extraction methods. With [6] and a window size of 2, the pair (*Gauß*, *Germany*) would have been counted twice (*all arrows*), whereas we ignore the second occurrence of this pair (*blue dashed arrow*) and only count the unique pairs (*red solid arrows*)

approach, we do the same but also keep the number of outlinks for each source page the same as in WikiStream. We generated 10 samplings for each variant, calculated the semantic relatedness performance for each and took the mean of the correlation values.

3.2 Evaluation Datasets

We used two datasets to semantically evaluate our findings. Both of these consist of a set of word pairs with a human-assigned similarity score. This captures human intuition of semantic relatedness. The datasets serve as ground truths for our evaluation.

WS-353. WS-353[7] (WordSimilarity-353) [2] consists of 353 pairs of English words and names and an assigned relatedness score between 0 and 10 for each pair.

MEN. The *MEN*[8] Test Collection [1] contains 3000 word pairs together with human-assigned similarity judgments, obtained by crowdsourcing using Amazon Mechanical Turk.

4 Methodology

In the following, we give a quick overview of the features we use for structurally comparing game and unconstrained navigation data. Then, we review the method used in [6] to extract semantic information from game navigation on Wikipedia and describe our adaptation of this method to achieve competitive results on unconstrained navigation.

Structural features To compare game and unconstrained navigation on a structural level, we particularly investigate two features: the (i) page overlap and the (ii) category overlap. (i) For the page overlap, we investigate how the set of the top 1000 most visited pages overlaps between the different datasets. (ii) For the category overlap, we decided on three classes which characterize the pages the best: *Person*, *Movie* and *Other*. Then, we classify each page of the top 1000 most frequent pages according to certain keywords in the category strings and in the page title. Afterwards we compare the class sizes of the different datasets.

Weighted Navigational Semantics We use the method proposed in [6], which is based on counting cooccurrences of concepts on paths from WikiGame, given a window of size k, resulting in cooccurrence vectors for each concept. Because WikiStream only contains transitions, we set $k := 2$ throughout the remainder of this work. These vectors are compared pairwise using the cosine similarity measure, giving a semantic similarity value of the corresponding concept pair. For evaluation, the resulting list of concept pair similarities is then correlated to a list of human similarity judgments (e.g. WS-353) using the Spearman correlation coefficient. A high absolute correlation value means that the method captures human judgment well.

Binarized Navigational Semantics We modify the approach described above. In particular we do not count cooccurrences of word pairs, but instead only keep the unique occurrence of a word pair. We call this *binarization*. This way, we reduce the impact of the limited number of very frequent links and raise the importance of less used links. An illustrative example is shown in Fig. 2.

5 Results and Discussion

Given our proposed methods from Sect. 4, in this section, we evaluate if unconstrained navigation data is suitable for extracting semantic relatedness. To this end, we first highlight structural properties and differences of unconstrained navigation data compared to data generated by gameplay. Secondly, we present and discuss the results of our experiments regarding extracting semantic relatedness.

5.1 Structural Experiments

As described in Sect. 4, we compare unconstrained and game navigation both by visited pages as well as page types. WikiStream is our baseline, since it represents actual

[7] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html.

[8] http://clic.cimec.unitn.it/~elia.bruni/MEN.

Table 2 Basic structural overlap of game navigation data and unconstrained navigation data on the top 1000 pages of each dataset. We can clearly see structural differences between unconstrained and game based navigation data

	WikiGame	WikiClickIU	WikiStream
(a) Page overlap			
WikiGame	100 %	26 %	10.1 %
WikiClickIU		100 %	20.5 %
WikiStream			100 %
(b) Category overlap			
Person	76	295	496
Movie	56	218	331
Other	858	476	172

Wikipedia traffic. Table 2(a) shows the page overlap between the three navigation datasets. WikiClickIU contains twice the amount of pages from WikiStream as WikiGame. Table 2(b) shows the distribution of page categories across the three datasets. The majority of pages in WikiClickIU ($\sim$53 %) can be assigned to persons and movies, as is the case in our baseline, WikiStream ($\sim$83 %). Opposed to that, the WikiGame - top pages mostly focus on other topics ($\sim$86 %).

Based on these two comparisons, we can assume that unconstrained navigation significantly differs from game navigation on a simple structural level. Also, we can infer that WikiClickIU is a more realistic representation of actual web traffic patterns than WikiGame, despite a potential bias on University of Indiana related articles.

5.2 Semantic Extraction Experiments

In this section, we investigate the performance of semantic relatedness extraction on navigational data: First, we extract semantic relatedness with both methods described in Sect. 4 and compare the results, which are shown in Table 3. And second, we compare unconstrained navigation with three baselines in order to understand if the extracted semantics are based on the actual user behavior or if they are merely a result of the underlying link structure.

Weighted Navigational Semantics As a baseline, we calculated the semantic extraction performance with the unaltered method from [6] on all our request datasets with the corresponding maximally found number of evaluation word-pairs (see r^{max} in Table 3). The data from WikiGame achieve the highest correlation on both evaluation datasets, namely WS-353 and MEN. Thus, game navigation serves best for extracting semantics in this setting. The results are hardly comparable though, because the evaluation word-pairs are not the same for each dataset. To address this, we calculate the correlations on the set of word-pairs common to all datasets (columns r_{ws}^{107} and r_{MEN}^{468}). WikiGame still outperforms all other datasets, but its performance dropped strongly because of the now missing pairs.[9]

Binarized Navigational Semantics Now, we evaluate our proposed binarization method. The results presented in Table 3 show that binarization is boosting semantic extraction performance in almost every possible combination of datasets and evaluation pairs. On the WikiGame data, we even manage to outperform the reported value of 0.638 for $k = 2$ from [6] (see r_{ws}^{max}). In general, the binarization yields very competitive values when comparing unconstrained navigation with game based navigation. On the MEN dataset, our method applied to WikiStream even outperforms WikiGame for all evalution word-pair sets. Without binarization, dominant features affect the cosine calculation and the resulting similarity value extremely. Binarization changes the cooccurrence vectors in such a way that the impact of previously extremely dominant features is decreased by a large margin, while the impact of less important features is raised relatively, thus taking all features into account more fairly. Overall, we showed that with our proposed binarization method, unconstrained navigation data yields competitive results when comparing against game navigation and even outperforms the original method by [6] on transitions.

Factors of Navigational Semantics Finally, we compare the binarized version of WikiStream against the three baselines described in Sect. 3. Our goal is to show that user navigation in the form of the selection of source pages as well as corresponding links in WikiStream contains more semantic information than the underlying link network alone. The results regarding semantic relatedness extraction on our sampled datasets from Sect. 3 are shown in Table 4. We see that source selection (*restricted*) only shows a small improvement but is based on a considerably smaller dataset (240.1M vs. 494.2M links and 29M vs. 1.4M source pages) when compared to the *binary* WikiLink dataset[10]. At the same time, we see, that selecting the right amount of links at each source has a strong effect on semantic relatedness (*random* vs. *distribution sampling*). And finally, we observe that selecting the correct sources as

[9] The performance decrease on fewer evaluation pairs can be explained as follows: on one hand, with fewer data points to correlate, the correlation task becomes easier and one might thus expect that the correlation value rises. On the other hand though, data points with faulty ordering have a greater impact on the correlation score. If we remove "good" data points, i.e. with good correlated ordering, we are left with the "bad" data points. This way, it is possible to actually decrease correlation performance when leaving out data points.

[10] The restriction of WikiLink to WikiStream source pages (*restricted*) should actually contain the same number of matchable evaluation pairs. We attribute this difference to the ever changing nature of Wikipedia.

Table 3 Performance comparison of our datasets on common evaluation pairs

Dataset	WS-353 correlations					MEN correlations				
	r_{ws}^{262}	r_{ws}^{224}	r_{ws}^{107}	r_{ws}^{max}	Pairs	r_{MEN}^{2111}	r_{MEN}^{1305}	r_{MEN}^{468}	r_{MEN}^{max}	Pairs
WikiStream (weighted)	0.528	0.506	0.388	0.527	288	0.411	0.399	0.445	0.370	2906
WikiStream (binary)	0.719	0.706	0.543	0.709	288	0.713	0.712	0.494	0.640	2906
WikiGame (weighted)	*n/a*	0.688	0.500	0.638	236	*n/a*	0.581	0.412	0.575	1396
WikiGame (binary)	*n/a*	0.722	0.563	0.728	236	*n/a*	0.644	0.461	0.639	1396
WikiClickIU (weighted)	*n/a*	*n/a*	0.398	0.419	120	*n/a*	*n/a*	0.268	0.225	568
WikiClickIU (binary)	*n/a*	*n/a*	0.454	0.458	120	*n/a*	*n/a*	0.278	0.214	568

In each column, we give the correlation values r_{ds}^{x} for the denoted evaluation dataset *ds* and the denoted number of common pairs *x*. The *n/a* values are given when we could not give a correlation value due to the lack of evaluation pairs in the dataset. "Weighted" denotes the original method by [6] and "'binary' denotes our new binarization method. Binarization allows for competitive performance for extraction of semantic relatedness on unconstrained navigation data compared to game based navigation

Table 4 Sampling results on baseline datasets based on WikiLink and WikiStream

Dataset	r_{WS-353}	Pairs	r_{MEN}	Pairs
Binary	0.625	270	0.548	2137
Restricted	0.629	269	0.553	2131
Random sampling	0.272	65.8	0.091	362.5
Distribution sampling	0.431	156.4	0.384	1011.4
WikiStream (binary)	0.709	288	0.640	2906

See Sects. 3 and 5 for details

well as the correct links (WikiStream (binary)) results in the best performance for extracting semantic relatedness. Thus, overall, we showed that, indeed, user behaviour, and not merely the underlying link structure, is an important factor for being able to extract semantic relatedness from unconstrained navigation data when using our new binarization method.

6 Conclusion

In this paper, we have investigated if unconstrained navigation data on Wikipedia can be used to extract semantics. We adopted a state of the art approach and found that, despite significant structural differences, deriving semantics from unconstrained navigation data performs similarly well as when using game data. Our adopted approach even outperformed the state of the art in some cases. Overall, we showed that unconstrained navigation data is suited for extracting semantics and that further research in this field may yield even more promising results.

Future work includes the extension of transitions to paths as basis for second order co-occurrences and investigation of other information systems (such as BibSonomy) in regard of human navigation behaviour and semantic exploitability.

Acknowledgments This work is funded by the DFG through the PoSTS II project. We also want to thank Alex Clemesha for providing us with the game data from the WikiGame website.

References

1. Bruni E, Tran NK, Baroni M (2014) Multimodal distributional semantics. J Artif Intell Res (JAIR) 49:1–47
2. Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G, Ruppin E (2001) Placing search in context: The concept revisited. In: Proc. of the 10th international conference on World Wide Web
3. Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. of the 20th international joint conference on Artifical intelligence
4. Meiss M, Menczer F, Fortunato S, Flammini A, Vespignani A (2008) Ranking web sites with real user traffic. In: Proc. First ACM International Conference on Web Search and Data Mining (WSDM), pp 65–75
5. Milne D, Witten IH (2008) An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: Proc. of the Conference on Artificial Intelligence, AAAI '08
6. Singer P, Niebler T, Strohmaier M, Hotho A (2013) Computing semantic relatedness from human navigational paths: A case study on wikipedia. IJSWIS 9(4):41–70
7. Strube M, Ponzetto SP (2006) Wikirelate! computing semantic relatedness using wikipedia. In: Proc of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI Press, 2
8. West R, Leskovec J (2012) Human wayfinding in information networks. In: Proc. of the 21st WWW Conf
9. West R, Pineau J, Precup D (2009) Wikispeedia: An online game for inferring semantic distances between concepts. In: Proc. of the 21st International Joint Conference on Artificial Intelligence (IJCAI)
10. West R, Paranjape A, Leskovec J (2015) Mining missing hyperlinks from human navigation traces: a case study of wikipedia. In: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, pp. 1242–1252
11. Zhang Z, Gentile A, Ciravegna F (2012) Recent advances in methods of lexical semantic relatedness - a survey. Nat Lang Eng 1(1):1–69