Analiza podobieństwa obiektów i pojęć w grafach wiedzy

Weronika T. Adrian

KRaKEn' Research Group
kraken@agh.edu.pl

Department of Applied Computer Science
AGH University of Science and Technology

Seminarium KIS, 25.11.2020 r.

Outline

Presentation Outline

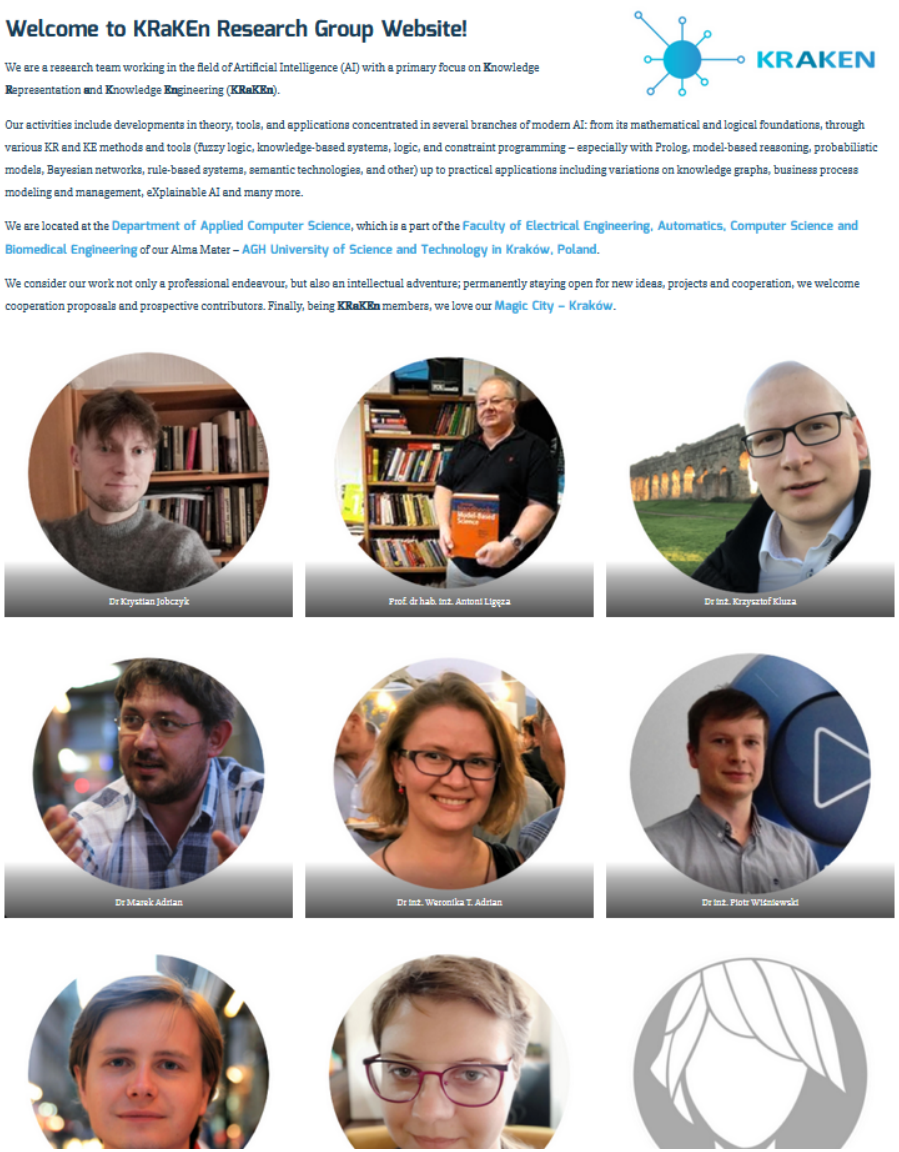
KRaKEn' Research Group

Knowledge Representation and Knowledge Engineering group:

- 1xProf. A. Ligęza
- 5xPhD: K.Kluza, K.Jobczyk, M.Adrian, W.T.Adrian, P.Wiśniewski
- 4xMSc: B.Stachura-Terlecka, M.Ślażyński, P.Jemioło, D.Sepioło
- 2xBSc: A.Bugaj, M.Szymkowski

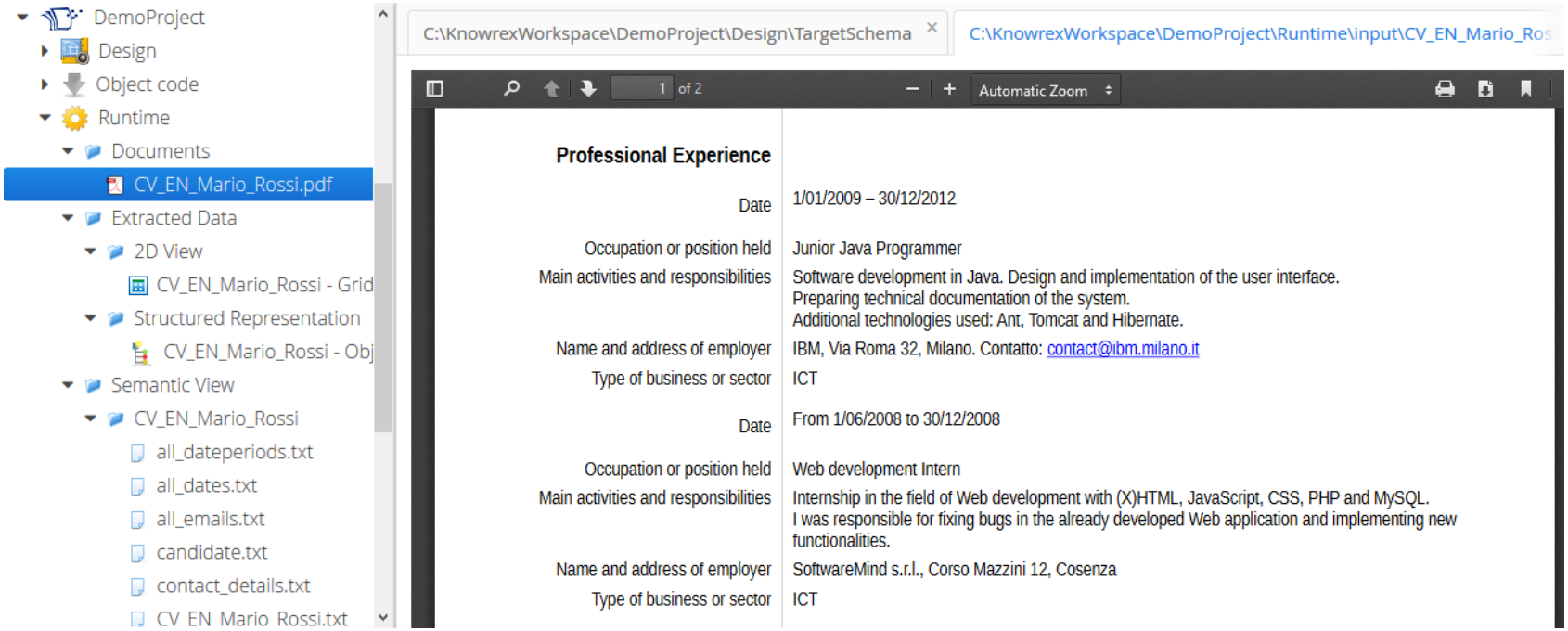
Research areas:

- KRR, logic, KBS, CSP
- semantic technologies, business processes, game theory
- explainability in AI



Research background

- PhD at Univeristy of Calabria – Knowrex project, Information Extraction



- Limitations: manual configuration bottleneck
- Solution proposal: automatic lexicon generation — entity set expansion problem (having a set of words/things, give more a superset of things *of the same kind*) — categorisation? similarity?

Back to AGH...

Areas of interests and applications

- Entity Set Expansion problem: Given a set of objects (words, things, ...) find a superset of things *of the same kind*
- recommendation engines, decision support systems

How to measure *similarity*?

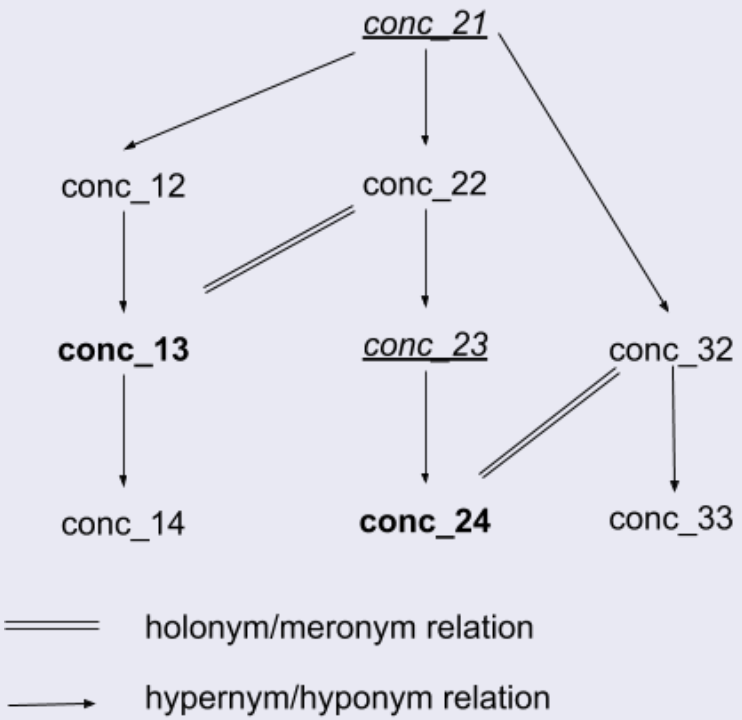
- Plethora of methods for assessing *similarity* of thingss
- Many levels: similarity of words, phrases, objects, documents, ...
- Which method to choose for a given problem and knowledge base?
- Which methods are intuitive and understandable yet perform well?

Research objectives

- review approaches to comparing concepts (in structured sources)
- analyze possibilities of compare *instances*
- implement selected approaches, develop practical tools

Assumption/focus: Semantic knowledge bases

Semantic networks



- semantics given by structure
- nodes and edges – universal KR
- classes, objects; relations

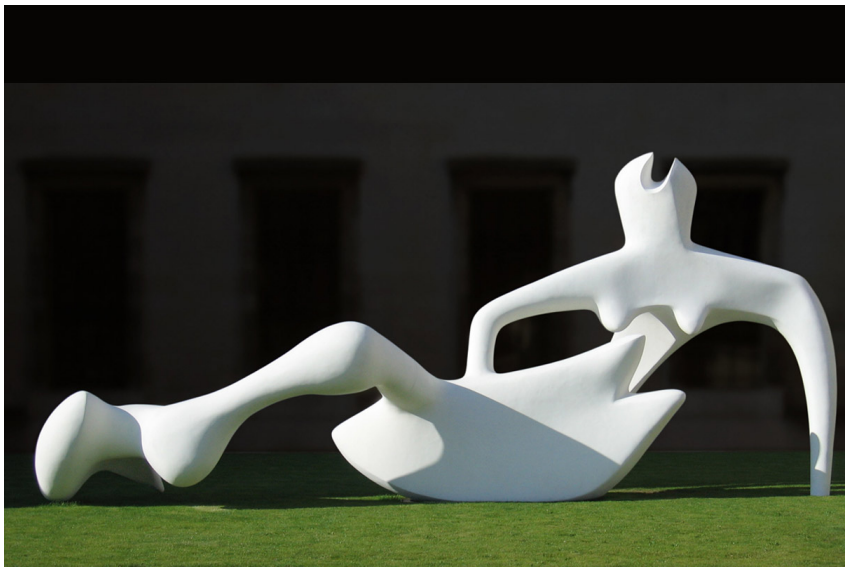
Modern knowledge graphs and semantic networks

- **DBPedia, Wikidata**: triple-based encyclopedias, knowledge about the world
- **BabelNet**: a multilingual semantic encyclopedia integrating information from several resources
- **WordNet**: a lexical database covering taxonomy of concepts, synonyms, antonyms, holo/meronyms, ...
- **Facebook**: persons, interests, activities, social interactions, communities

Presentation Outline

What is similarity?

1. **psychological perspective**: analyze the the common and disjoint *features* of the objects
2. **geometric perspective**: calculate the "distance" between the concepts:
   - structure-based metrics
   - embeddings-based methods

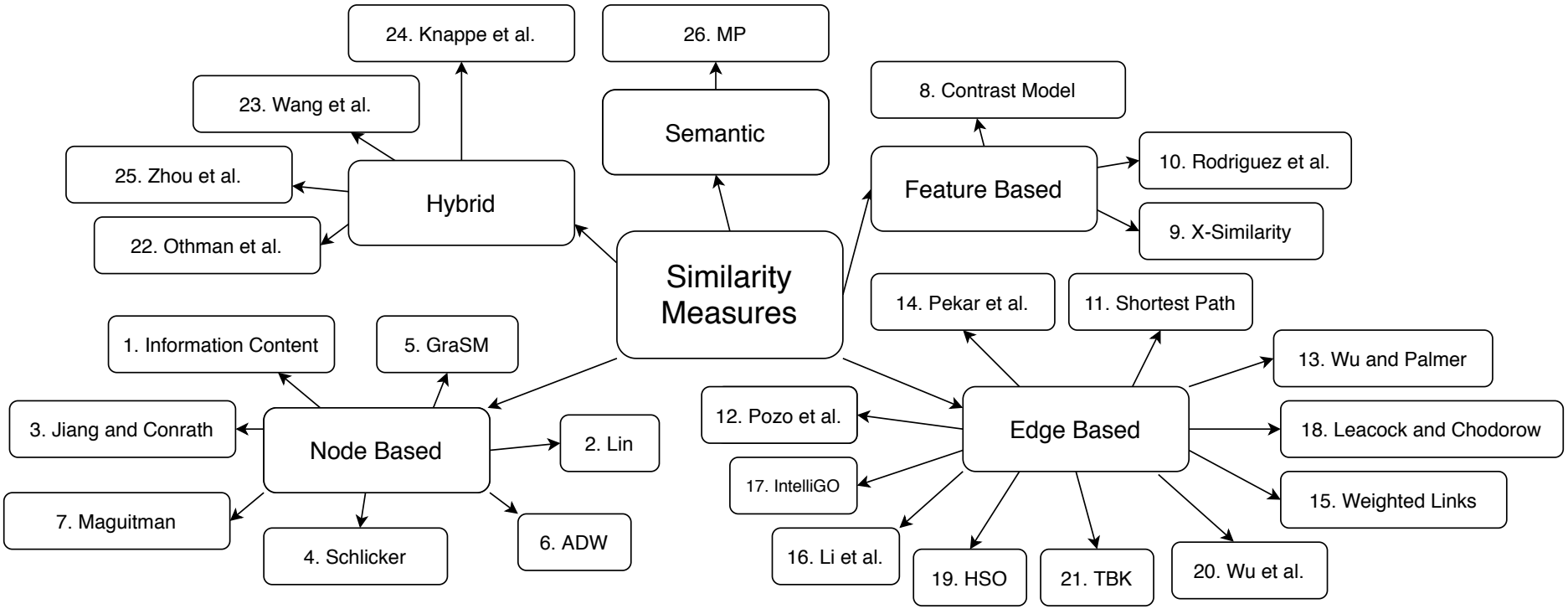


Survey and classification attempt

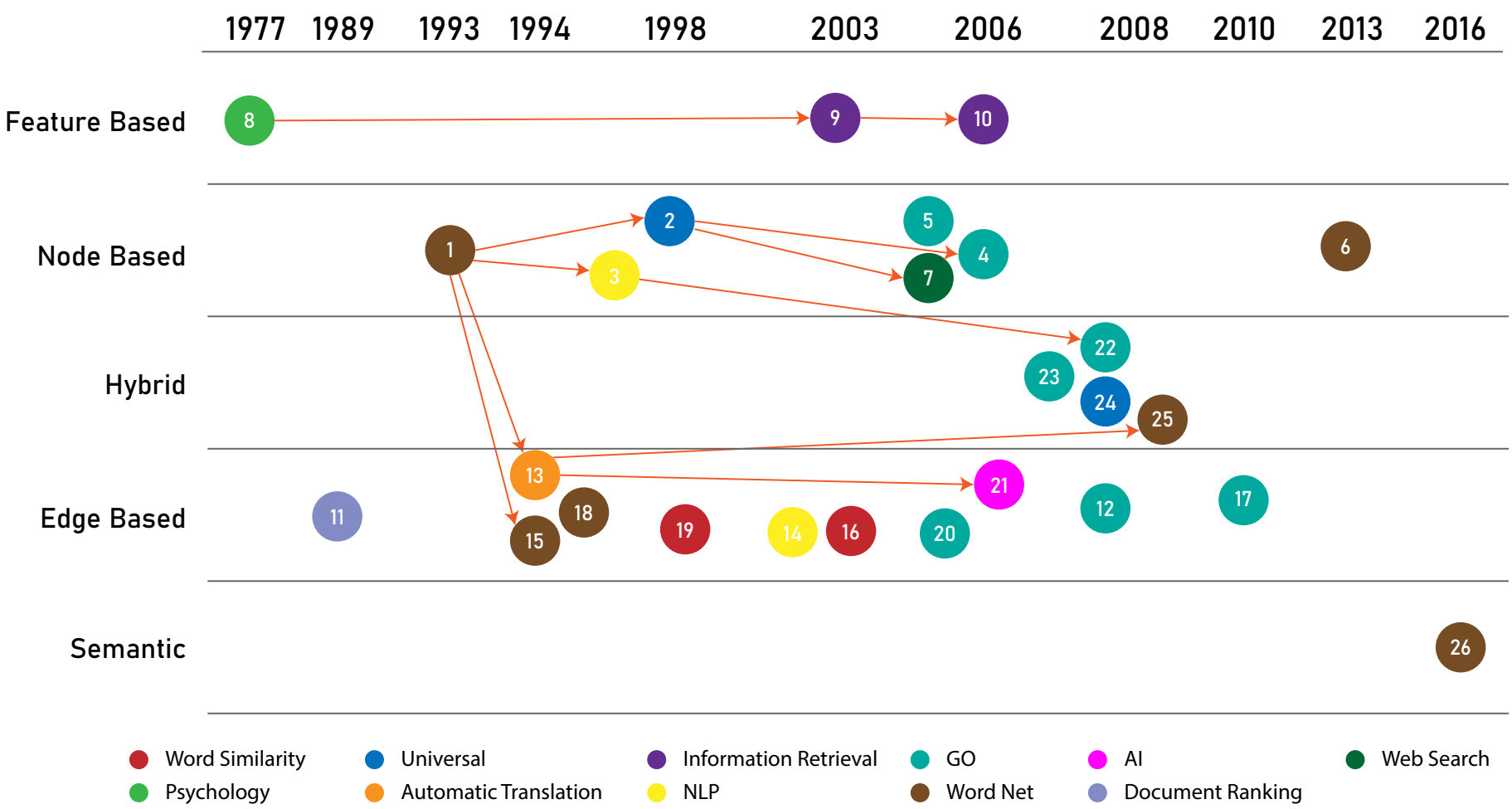
- Review of 60-70 papers on semantic similarity
- Selection: based on the attributes, such as: description, citations, references, etc. ... we identified "core", prominent, influential methods and/or methods visibly different from others
- An attempt to classify and model the methods landscape domain

Paper ID	Method name	Method type	Paper title	Authors	Citations	Year		Description	Semantic Similarity Definition Used	domain of application	Method Based On	Pros and Cons
tversky1977features	Contrast Model of Similarity	Feature-based	Features of Similarity	Amos Tversky	9716	1977	231.333	Similarity of the objects as linear combination or a contrast of the measures of their common and disctinctive features. He defines similarity as a matching process. He also introduced less known ration model where he describest similarity as a function of common features devided by the number of all common and disjunctive features of both stimuli.	Similarity as comparison of features (opposed to computation of metric distance between points).	Psychology	Novel	Pro: His approach is not influenced by many mathematical assumptions with geometrical and metric appraoches. Pro: It is very simple model.
rada1989development	Shortest Path	Edge-based	Development and application of a metric on semantic nets	R. Rada H. Mili E Bicknell M Blettner	903	1989	30.1	It is measured by subtracting the shortest path between the concepts in the hierarchy from doubled longest path in the hierarchy between the concepts.	The aggregate of interconnections between the concepts (average of the path lengths between pair of nodes).	Document Ranking	Not based on but connected with https://dl.acm.org/doi/10.1145/7902.7906	Pro: Distance approach sets baseline on Mesh (Medical Subject Headings) sets for performance.
richardson1994using	Weighted Links	Edge-based	Using WordNet as a knowledge base for measuring semantic similarity between words	R Richardson Alan F. Smeaton J Murphy	296	1994	11.84	Same method as above but the connections along the path have different weights. The score is obtained by summing up the weights.	Information content of the first class in the noun hierarchy that subsumes both classes.	WordNet	Resnik Information Content	Research is ongoing - that was the conclusion of the paper.
wu1994verbs	Wu and Palmer	Edge-based	Verbs semantics and lexical selection	Wu and Palmer	3892	1994	155.68	Ratio between the doubled distance from most specific comon concept to the root concept and sum of distances between the concepts and most specific comon concept and agan doubled distance from most specific comon concept to the root concept.	It is ratio between distance from root to closest common ancestor of 2 terms and the path between terms routed through root node.	Machine Translation	Resnik + https://www.researchgate.net/publication/221102495_Dynamic_Programming_Method_for_Analyzing_Conjunctive_Structures_in_Japanese	b
resnik1995using	Resnik	Node-based (Information Content)	Using Information Content to Evaluate Semantic Similarity in a Taxonomy	P. Resnik	4301	1995	179.208	A method of determining the similarity between concepts. Calculates the similarity of two concepts using the information content of their lowest common ancestor. The method uses shared information content that is information content of the concepts' parents (in the hierarchy) to determine the similarity between them. Intuition: if the common ancestor of two concepts has a high information content value, then the concepts share a lot of information and are similar. Values are in range [0, ∞), the higher the value, the greater the similarity. Requires "is-a" relations.	Semantic similarit in an is-a taxonomy. Similarity is the maximal information content over all concepts of which both words could be an instance.	WordNet	Novel	Pro: Not sensitive to a problem of varying link distances (as in edge based methods). Con: Not presented in the paper ("the method performs encouranginly well")
			Semantic Similarity					The distance between concepts in this method is the difference	Similarity definition derived from edge based methods: The			

Semantic Similarity Methods Diagram (Ontology)

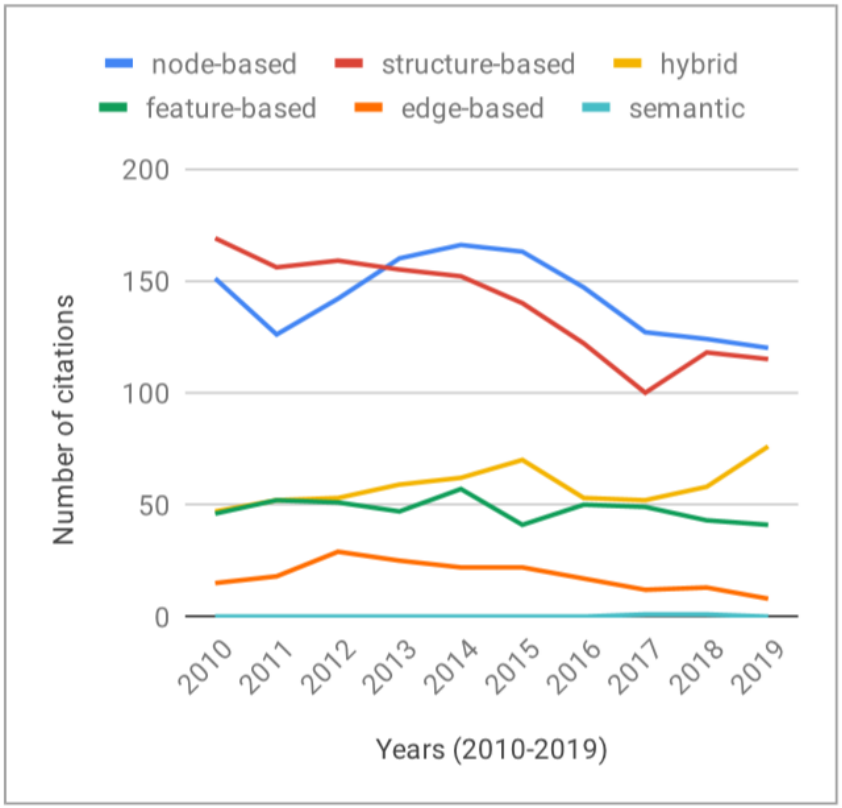


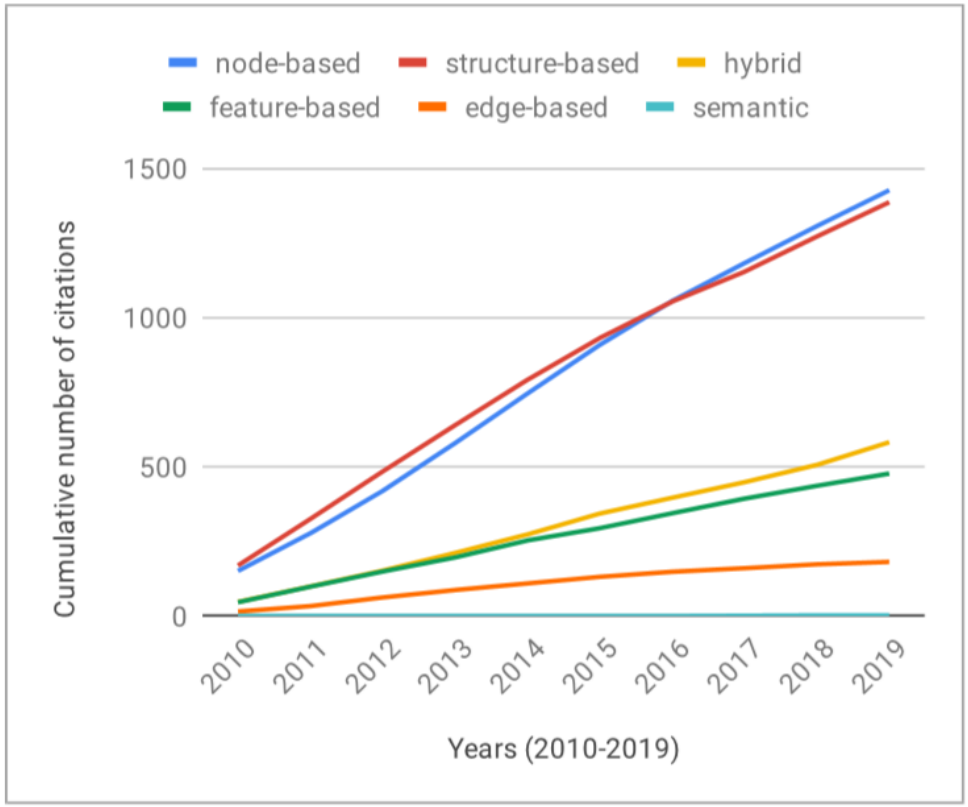
Semantic Similarity Methods Diagram (Ontology)



Bibliometric analysis

Trends in development of new methods over time

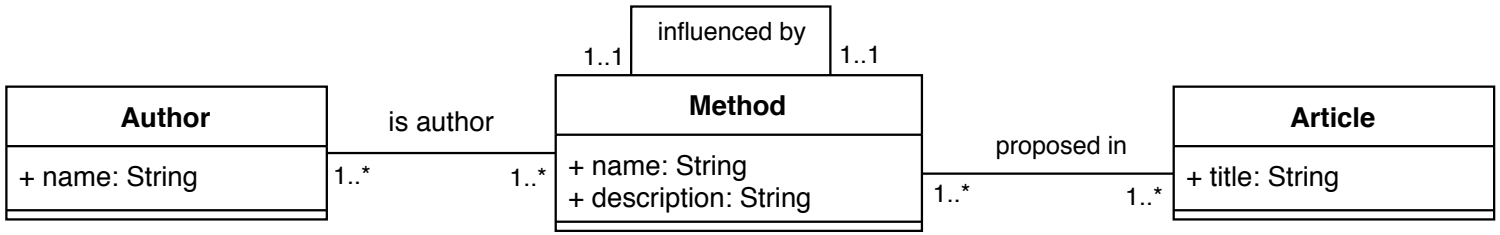




Tool supporting the analysis

"Historical atlas" of research methods:

- Data: ontology of methods, in json
- Two visualization methods: graph-base and chronological
- Universal: for analyzing any domain



Visual “guide” about similarity metrics

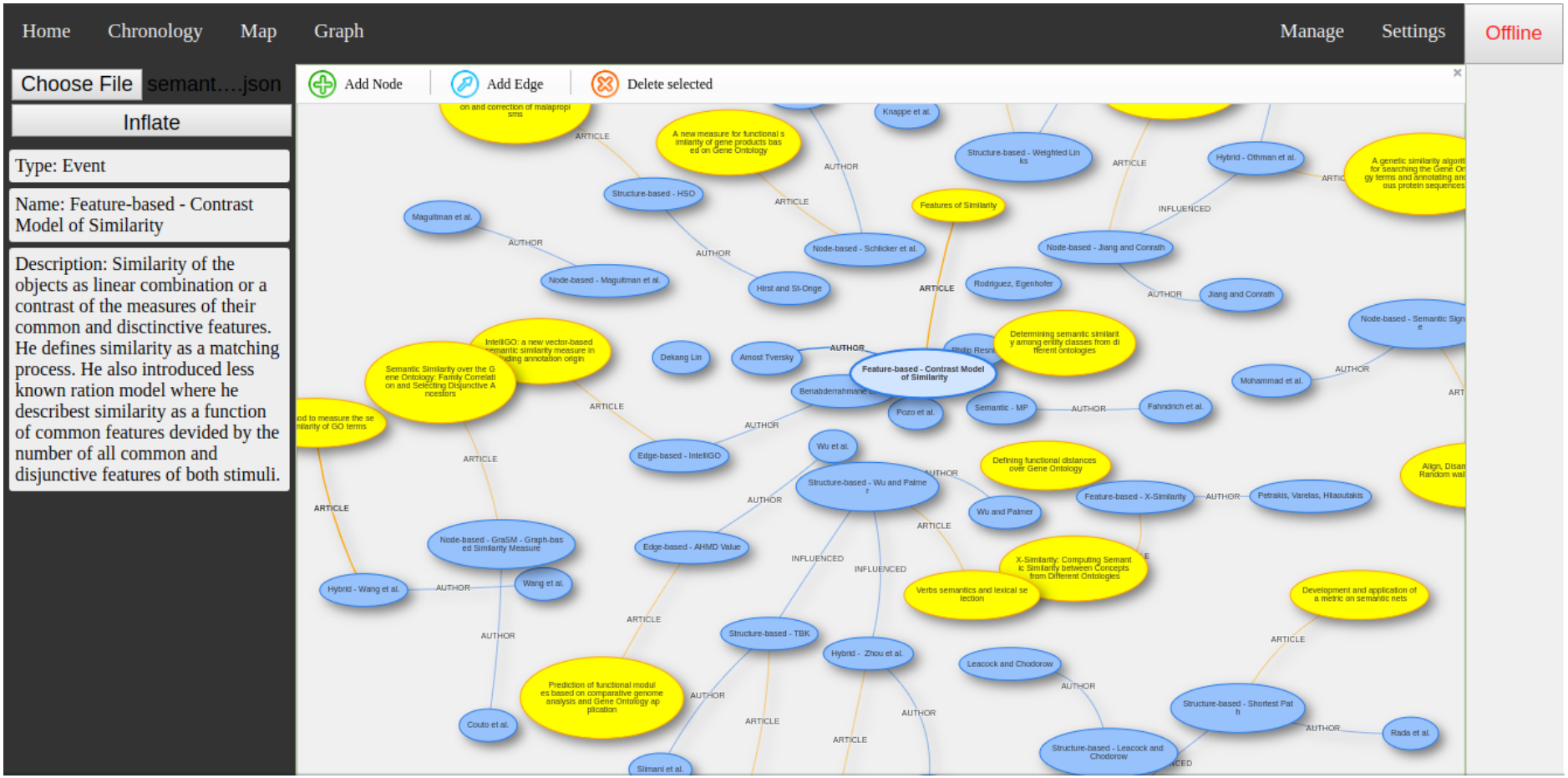


Figure: See `https://gitlab.com/SzymonMajk/chartas-front`.

“Historical atlas” of research proposals

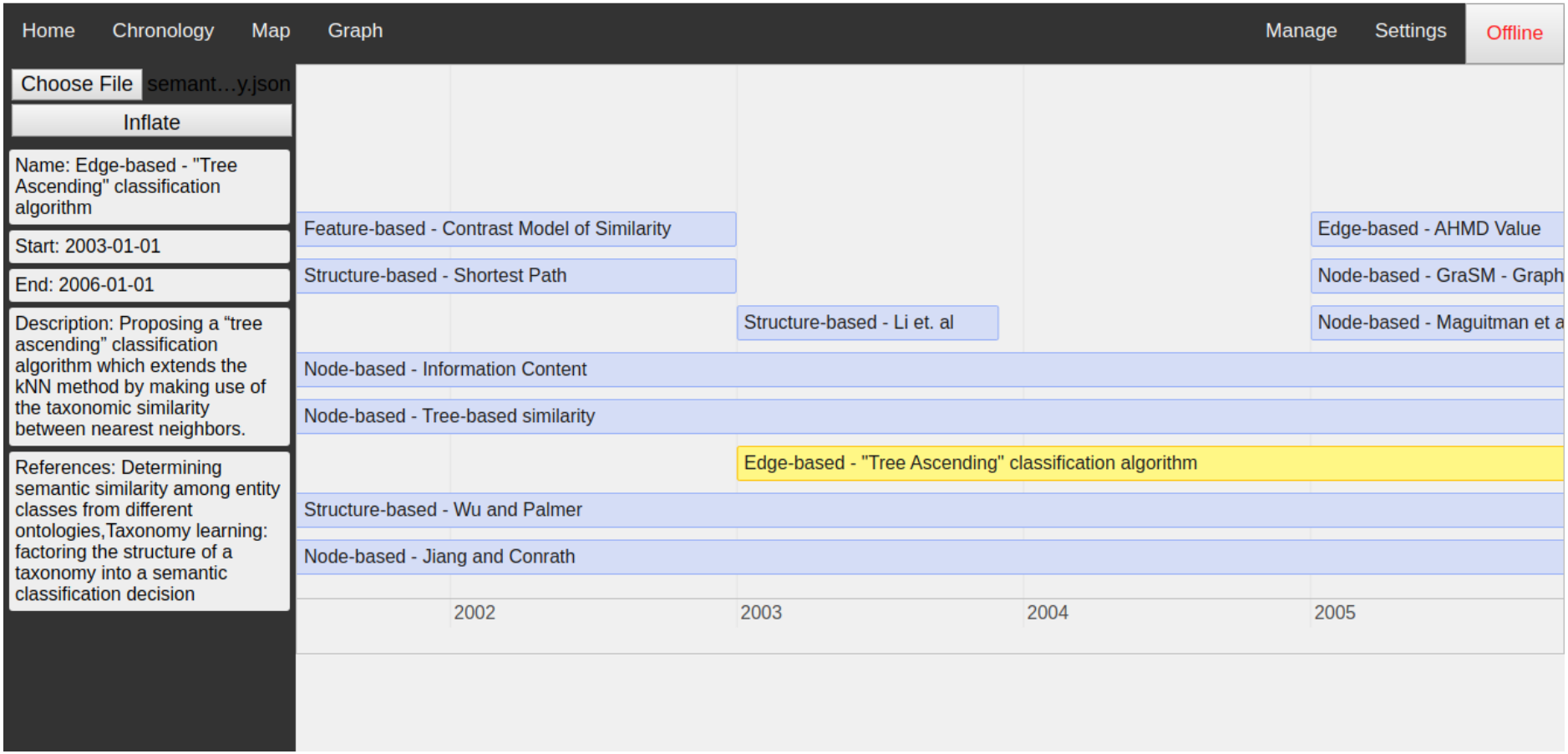


Figure: See `https://gitlab.com/SzymonMajk/chartas-front`.

Summary (of the first thread)

Research questions

How can we measure similarity? What is the state-of-the-art? What methods are best for which situations? How to support literature research?

Obtained results

- Review and classification + modeling of the domain
- Result: a guide for newcomers to the domain
- MSc students studying the subject at AGH UST

Paper

"Tracing the Evolution of Approaches to Semantic Similarity Analysis", by W.T.Adrian, S.Skoczeń, S.Majkut, K.Kluza, A.Ligęza, presented at IC3K / KEOD conference (November 2020)

Presentation Outline

1. Introduction

2. Review and analysis of semantic similarity metrics

3. **Implementation and extension of selected metrics**
   - Yang & Powers similarity metric
   - Alvarez & Lim similarity metric
   - Experiments and results

4. Conclusion

Reviewing metrics of semantic similarity

Looking for a measure that is:

- understandable and intuitive
- based on a structured knowledge base
- "explainable" (*n*-dimensional vectors were not what we focused on)

Paper ID	Method name	Method type	Paper title	Authors	Citations	Year		Description	Semantic Similarity Definition Used	domain of application	Method Based On	Pros and Cons
tversky1977features	Contrast Model of Similarity	Feature-based	Features of Similarity	Amos Tversky	9716	1977	231.333	Similarity of the objects as linear combination or a contrast of the measures of their common and disctinctive features. He defines similarity as a matching process. He also introduced less known ration model where he describest similarity as a function of common features devided by the number of all common and disjunctive features of both stimuli.	Similarity as comparison of features (opposed to computation of metric distance between points).	Psychology	Novel	Pro: His approach is not influenced by many mathematical assumptions with geometrical and metric appraoches. Pro: It is very simple model.
rada1989development	Shortest Path	Edge-based	Development and application of a metric on semantic nets	R. Rada H. Mili E Bicknell M Blettner	903	1989	30.1	It is measured by subtracting the shortest path between the concepts in the hierarchy from doubled longest path in the hierarchy between the concepts.	The aggregate of interconnections between the concepts (average of the path lengths between pair of nodes).	Document Ranking	Not based on but connected with https://dl.acm.org/doi/10.1145/7902.7906	Pro: Distance approach sets baseline on Mesh (Medical Subject Headings) sets for performance.
richardson1994using	Weighted Links	Edge-based	Using WordNet as a knowledge base for measuring semantic similarity between words	R Richardson Alan F. Smeaton J Murphy	296	1994	11.84	Same method as above but the connections along the path have different weights. The score is obtained by summing up the weights.	Information content of the first class in the noun hierarchy that subsumes both classes.	WordNet	Resnik Information Content	Research is ongoing - that was the conclusion of the paper.
wu1994verbs	Wu and Palmer	Edge-based	Verbs semantics and lexical selection	Wu and Palmer	3892	1994	155.68	Ratio between the doubled distance from most specific comon concept to the root concept and sum of distances between the concepts and most specific comon concept and agan doubled distance from most specific comon concept to the root concept.	It is ratio between distance from root to closest common ancestor of 2 terms and the path between terms routed through root node.	Machine Translation	Resnik + https://www.researchgate.net/publication/221102495_Dynamic_Programming_Method_for_Analyzing_Conjunctive_Structures_in_Japanese	b
resnik1995using	Resnik	Node-based (Information Content)	Using Information Content to Evaluate Semantic Similarity in a Taxonomy	P. Resnik	4301	1995	179.208	A method of determining the similarity between concepts. Calculates the similarity of two concepts using the information content of their lowest common ancestor. The method uses shared information content that is information content of the concepts' parents (in the hierarchy) to determine the similarity between them. Intuition: if the common ancestor of two concepts has a high information content value, then the concepts share a lot of information and are similar. Values are in range [0, ∞), the higher the value, the greater the similarity. Requires "is-a" relations.	Semantic similarit in an is-a taxonomy. Similarity is the maximal information content over all concepts of which both words could be an instance.	WordNet	Novel	Pro: Not sensitive to a problem of varying link distances (as in edge based methods). Con: Not presented in the paper ("the method performs encouranginly well")
			Semantic Similarity					The distance between concepts in this method is the difference	Similarity definition derived from edge based methods: The			

Yang & Powers similarity measure

- An edge-based method basing on graph traversal (way of traversal as well as path calculation depends on a variant)
- To compute the similarity we use the following formula:

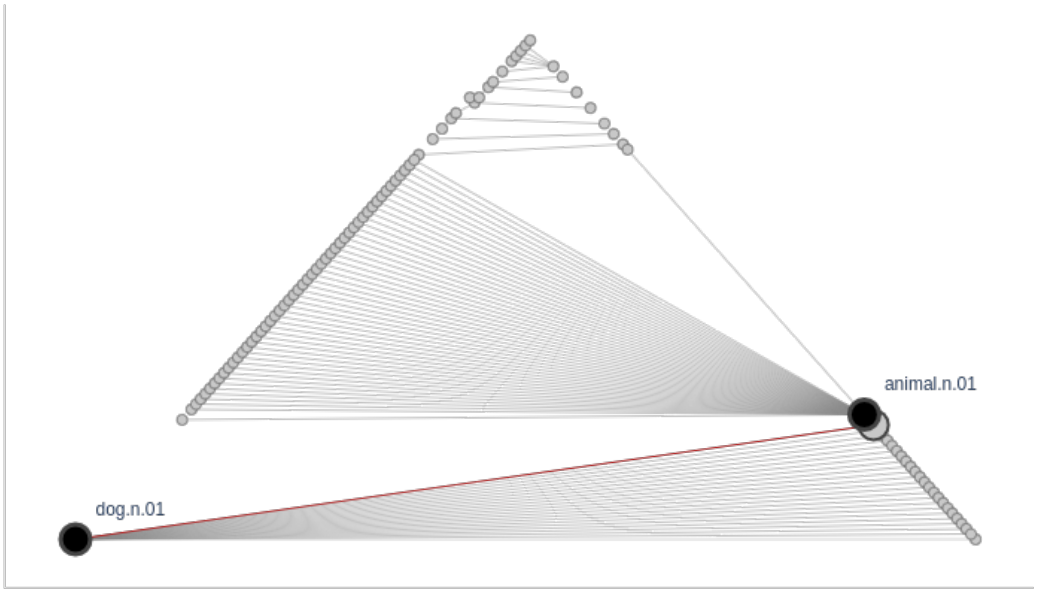
$$Sim(c_1, c_2) = \begin{cases} \alpha_t \prod_{i=1}^{dist(c_1,c_2)} \beta_{ti} & \text{if } dist(c_1, c_2) < \gamma \\ 0 & \text{if } dist(c_1, c_2) \geq \gamma \end{cases}$$

where:

- c_1, c_2 denotes the concepts being compared
- α is the link-type factor
- β is the depth factor
- γ is path length threshold
- $t \in \{hh, hm\}$ denotes relation type (hypernym-hyponym, holonym-meronym)
- $dist(c_1, c_2)$ is a number of edges in the path between both concepts

Creation of graph for Yang & Powers metric

- There are 6 variants overall
  - Traversal: Uni-Directional and Bi-Directional Search
  - Result: max, sum, mean
- We implemented Bi-Directional traversal and maximal value (Sim_{max_B})
- To create a graph we start with all meanings of the given words
- Algorithm recursively traverse graph finding all hypernyms/hyponyms/holonyms/meronyms of words until it will find common node (both traversal processes find the same node).



Example

For the pair (dog,animal):

$$t = hh, \alpha_t = 0.7, \beta_t = 0.85$$

$$sim = 0.7 \cdot \prod_{1}^{2} 0.85 = 0.595$$

Alvarez & Lim similarity measure

- An edge-based method (considers the shortest path between words in the taxonomy) - but not only
- Three main components are taken into account to compute the distance:

$$\text{dist}(w_1, w_2) = \arg\min_{(i,j)} \begin{bmatrix} pl(c_{1i}, c_{2j}) \\ \cdot d_{nca}(c_{1i}, c_{2j}) \\ \cdot (1 + \text{gloss}(c_{1i}, c_{2j})) \end{bmatrix}$$

where:

- c_{ij} denotes the j-th meaning of the i-th word
- $pl(c_{1_i}, c_{2_j})$ is the path length between c_{1_i} and c_{2_j}
- $d_{nca}(c_{1_i}, c_{2_j}) = 1 - \frac{depth(c)}{maxdepth}$
- $depth(c)$ is the depth of the concept c in the created graph
- $\text{gloss}(c_{1_i}, c_{2_j}) = 1 - \frac{|g_{1_i} \cap g_{2_j}|}{\max(|g_{1_i}|, |g_{2_j}|)}$ g_{1_i}, g_{2_j} - descriptive definitions of concepts

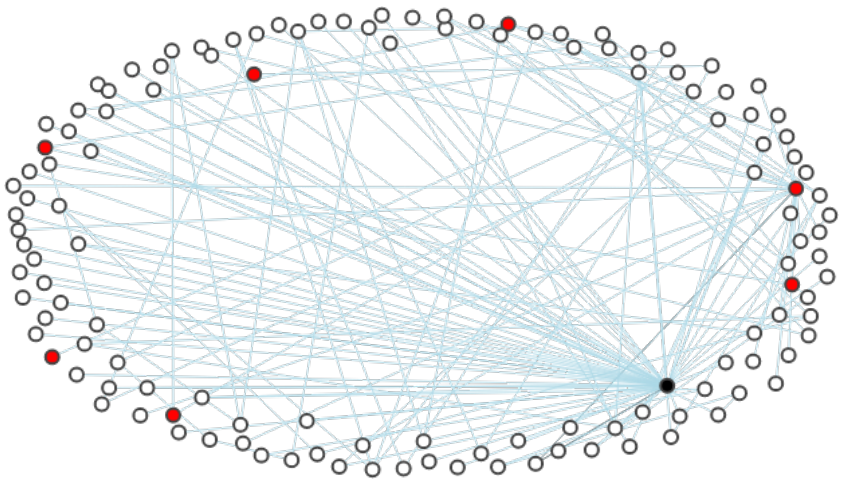
- distance to similarity: $sim(w_1, w_2) = \exp(\frac{-dist(w_1, w_2)}{b})$, b set experimentally

Gloss

- dog = "a common animal with four legs, especially kept by people as a pet"
- pet = "an animal that is kept by people as a companion and treated kindly"

Creating a graph for Alvarez & Lim metrics

- The algorithm inserts into the graph hypernyms of words found in the path between the given concept and a root in WordNet
- For each concept $r \in \{hyp(c) \cup mer(c) \cup hol(c)\}$, $c \in \{s(w_1) \cup s(w_2)\}$, where s is a set of synonyms and hyp, mer, hol are the sets of hyponyms, meronyms and holonyms respectively, we recursively add the hypernyms existing in the path from r to root.
- Edge weight: $weight(c_{1_i}, c_{2_j}) = 1 - \frac{depth(c_{1_i}) + depth(c_{2_j})}{2 * maxdepth}$



Example

For the pair (dog,animal):

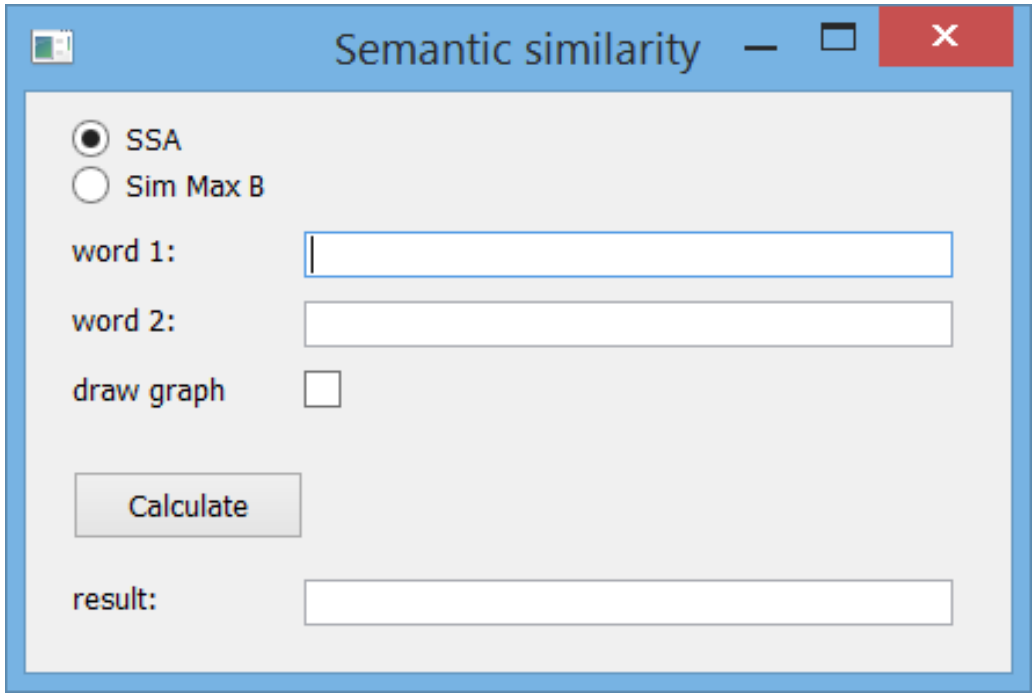
$$pl = 1.4; depth = 5; gloss = 1$$

$$dist = 1.4 * (1 - \frac{5}{20}) * (1 + 1) = 2.1$$

$$sim = \exp(\frac{-2.1}{4}) \approx 0.592$$

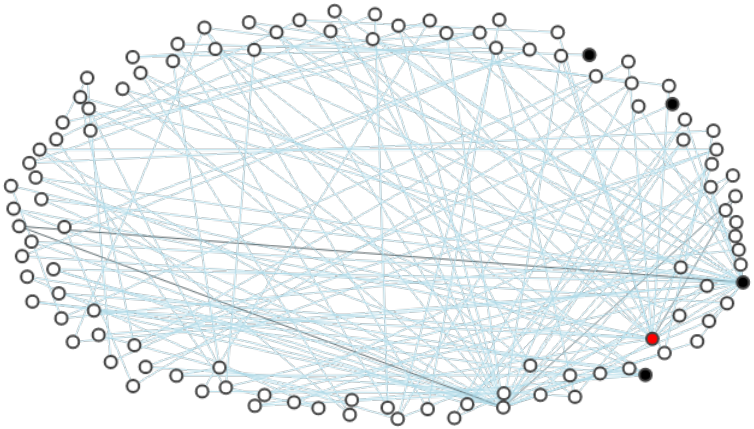
Implementation

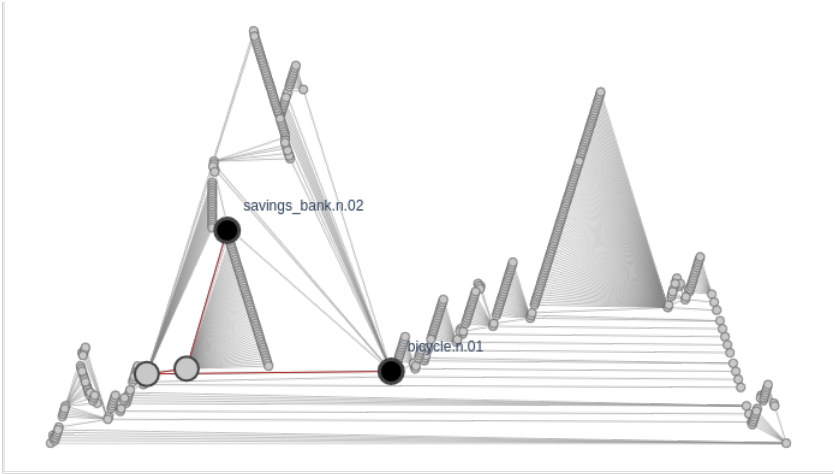
- Python 3, libraries: numpy, NLTK, networkx, plotly
- Simple GUI made in PyQt

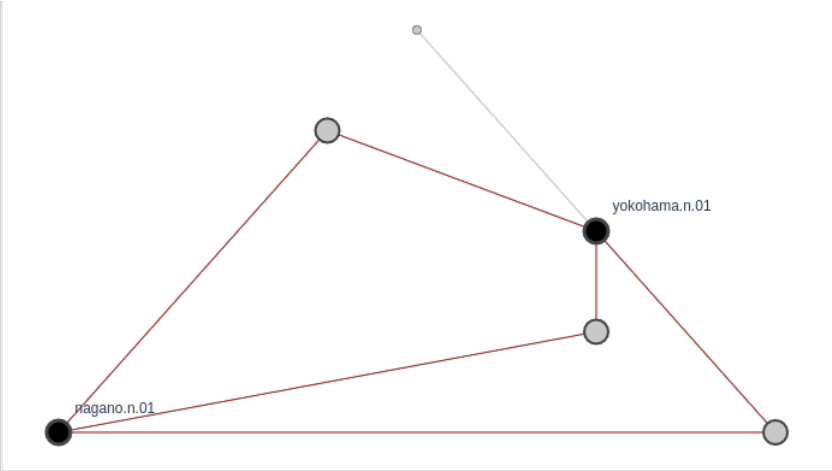


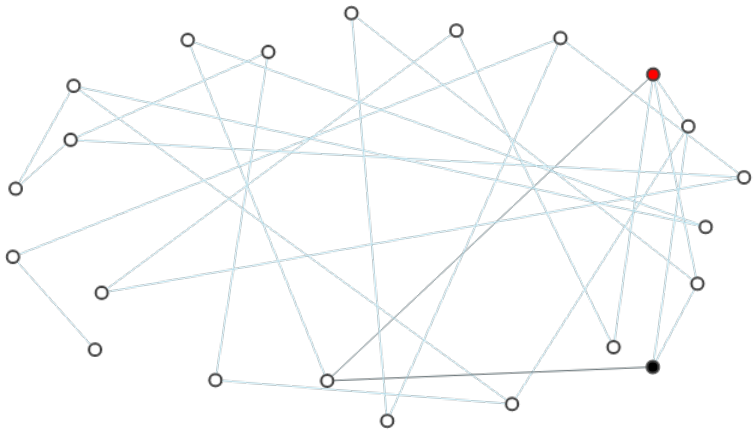
The metrics are extended to take into consideration also the instance similarity by the use of *instance hypernym* relations while creating a graph.

Visualization possibilities









Experiments over WordSim353 dataset

- results better than most of the knowledge-based metrics
- worse than hybrid and embedding-based metrics

Word 1	Word 2	Sim_{maxB}	SSA
tiger	cat	8.5	8.95
tiger	tiger	10	10
tiger	animal	4.17	3.79
plane	car	5.95	6.72
train	car	8.5	5.24
money	cash	5.95	6.23
king	queen	9.0	10
football	soccer	8.5	9.22
vodka	brandy	5.95	6.66
food	fruit	4.17	3.56
money	dollar	2.92	4.1

Tests on our "instance dataset":

Instance 1	Instance 2	Sim_{maxB}	SSA
Warsaw	Cracow	5.95	8.13
Roma	Vienna	5.95	7.59
Roma	Budapest	5.95	7.59
Roma	Hamburg	4.16	6.07
Newton	Galileo	4.17	4.69
Newton	Mozart	0.7	0.8
Vistula	Thames	5.95	7.12
Vistula	Balaton	2.92	3.14

Summary (of the second thread)

Research questions

How can we measure similarity of instances in a graph-oriented knowledge base so it is human-readable, intuitive, and accurate?

Obtained results

- Implementation of selected methods and extension to instance similarity
- Practical (educational) tool with visualization options
- First results and intuitions towards metrics combining structure of the knowledge base and vector representation learning

Paper

"Adapting selected knowledge-based similarity metrics for instance similarity", by W.T.Adrian, A. Bugaj, P. Swędrak, presented at LENLS17 workshop (Nov. 2020)

Presentation Outline

Conclusion

Main results

- Analysis and classification of semantic similarity metrics
- Tool development: historical atlas of methods, a tool calculating similarity between words, other implemented metrics and experiments
- 2 conference/workshop papers, MSc students involved in the topic

Challenges

- Keep up-to-date about the state-of-the-art and new proposals
- Embeddings methods!

Plans for future

- Experiments on richer knowledge bases and selected problems
- Towards new metrics for semantic similarity combining structure-based and embeddings-based methods

The End

Thank you for your attention!

Do you have any questions?

Contact me at: wta@agh.edu.pl

Contact us at: kraken@agh.edu.pl