Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Panská 26, SK-81101 Bratislava
radovan.garabik@kassiopeia.juls.savba.sk

WORD EMBEDDING BASED ON LARGE-SCALE WEB CORPORA AS A POWERFUL LEXICOGRAPHIC TOOL

The Aranea Project offers a set of comparable corpora for two dozens of (mostly European) languages providing a convenient dataset for NLP applications that require training on large amounts of data. The article presents word embedding models trained on the Aranea corpora and an online interface to query the models and visualize the results. The implementation is aimed towards lexicographic use but can be also useful in other fields of linguistic study since the vector space is a plausible model of semantic space of word meanings. Three different models are available – one for a combination of part of speech and lemma, one for raw word forms, and one based on FastText algorithm uses subword vectors and is not limited to whole or known words in finding their semantic relations. The article is describing the interface and major modes of its functionality; it does not try to perform detailed linguistic analysis of presented examples.

1. Introduction

The Aranea Project (Benko 2014) offers a set of comparable corpora for two dozen of (mostly European) languages providing a convenient dataset for NLP applications that require training on large amounts of data. The corpora are built using the same methodology and compatible natural language processing tools and are available via NoSketch Engine interface (Rychlý 2007) at the web page of the project[1].

---

[1]   http://aranea.juls.savba.sk

Word embeddings is a collective name for various methods of representing words as vectors within a many-dimensional vector space. Although first mentioned as a theoretical concept in (Harris 1954), it gained momentum with the publication and open-sourcing of the *word2vec* software (Mikolov et al. 2013), and currently is an indispensable part of many NLP related tasks and processes. It is supported by several mature OpenSource frameworks, in particular, *gensim* (Řehůřek and Sojka 2010) is rather popular with researchers preferring the Python programming language – this framework is also used to generate our vector models. The vector space and the relation of words represented by vectors within is connected with more abstract semantic meanings of the words and their relations.

Our work presents an online accessible interface for vector models for main languages in the Aranea corpora to be used in lexicographic work. The implementation is somewhat Slovak-centric in the sense that some features are either available only for Slovak, or their implementation for other languages is not tuned for coverage or accuracy. This is understandable, since the implementation is geared towards use in Slovak lexicography, and indirectly because of the state of the art lemmatization (Garabík 2006) used in the Slovak corpora.

The models use automatic detection of bigrams, which aids to the lexicographic description of multiword expressions (though there are better tools available for collocation analysis, so this is useful in a supplemental role only). There is a possibility to filter out-of-dictionary lemmas, which is useful in uncovering non-obvious meanings of existing words. Otherwise, the lemmas obtained by statistical and heuristic guesser can be erroneous, but their inclusion often displays unexpected relations between words not covered by existing dictionaries (this does not exclusively cover only neologisms).

2. Vector models

At the time of writing, there are usable vector models for 22 languages and three more language models are in the test phase. Because of the aim to a provide useful resource for lexicographic work, we provide a special vector model (Slovak *ll* in table 1) using very low threshold for word frequency in the corpus (10 occurrences in the corpus), while other corpora use variable threshold,

depending on the size of the corpus (20 for the smallest corpora, 400 for the biggest ones). Although such a low threshold brings a lot of "noise" (uncommon typos, mislemmatized entries, errors in tokenization, foreign language citations etc.) into the results, it also helps to uncover rare, but relevant semantic relations or synonyms. The models use a context window 7 words wide and the skip-gram training algorithm.

Table 1: Overview of language models and their source corpora

language	corpus	corpus size
Arabic[*]	Araneum Arabicum	978 M
Bulgarian	Araneum Bulgaricum	1.2 G
Chinese (simplified)[#]	Hanku	1.2 G
Croatian	hrWaC v2.0	1.9 G
Czech	Araneum Bohemicum	5.2 G
Dutch	Araneum Nederlandicum	1.2 G
English	Araneum Anglicum	11.4 G
Estonian	Araneum Estonicum	430 M
Finnish	Araneum Finnicum	1.2 G
French	Araneum Francogallicum	8.7 G
German	Araneum Germanicum	9.1 G
Hungarian	Araneum Hungaricum	1.2 G
Italian	Araneum Italicum	1.2 G
Latin	Araneum Latinum	109 M
Latvian	Araneum Lettonicum	671 M
Polish	Araneum Polonicum	1.2 G
Portuguese	Araneum Portugallicum	1.2 G
Russian	Omnia Russica	29.7 G
Slovak	Araneum Slovacum + prim-8.0-juls-all	4.6 G
Slovak ll[&]	Araneum Slovacum + prim-8.0-juls-all	4.6 G
Slovene	slWaC v2.1	895 M
Spanish	Araneum Hispanicum	1.2 G
Swedish	Araneum Suedicum	1.2 G

* not lemmatized, wordform model serves as a fallback when the lemmata model is selected
\# lemmatization not applicable, the lemmata model differs from the wordform one only by the background presence of part of speech information (Gajdoš, Garabík and Benická 2016)
\& differs from the baseline Slovak model by using substantially lower threshold for word occurrence (frequency)

Test models

language	corpus	corpus size
Georgian	Araneum Georgianum	254 M
Norwegian$	Araneum Norvegicum	1.6 G
Romanian	Araneum Dacoromanicum	1.2 G

$ mixture of Nynorsk and Bokmål

Traditionally, to quantify semantic difference in word embeddings, cosine similarity is used – words close in meanings have their vectors almost parallel (angle θ between them close to zero and cos 0 = 1), unrelated words almost perpendicular (right angle, and cos $\pi/2$ = 0). Our users, however, prefer to align their spatial intuition with the semantic space model and think of semantically "close" words as "near" in the spatial sense and "unrelated" words as "far" in the spatial sense, therefore we define "semantic difference" as $\sqrt{(1-\cos^2 \theta)} = \sin \theta$, being close to zero for near-synonyms and close to one for unrelated words.

For most languages, there are three different models available: the model trained on the combination of part of speech and lemmas, the model trained on word forms, and a FastText model (Mikolov et al. 2018). Common linguistic preprocessing in all the models includes text normalization, deduplication, boilerplate removal, tokenization, and sentence level segmentation (Benko 2014).

The model trained on the combination of part of speech and lemmas (called lemmata model in this article) is trained on the sequence of the combination of part of speech (tagged by Araneum Universal Tagset Version 1.0) and lemmatized tokens. We keep the uppercase lemmas in the capitalization that is "natural" for the language in question. In particular, proper names in almost all languages and German nouns are capitalized. This helps in keeping casual users from the undue cognitive load, while allowing to distinguish homonymous common and proper names (and, in the case of German, nouns from other parts of speech). The model over lemmas loses information regarding inflected forms (and by implication, perhaps interesting syntactical features), but users generally expect to enter lemmas, and it is the lemmas that carry semantic information.

The main disadvantage of this model is that it exposes errors in lemmatization, particularly if the queried word was not known to the morphological database or is lemmatized incorrectly.

The model trained on word forms is the closest to commonplace word embedding usage. However, we normalize the capitalization of tokens according to their predominant capitalization (more than 90% of occurrences in the corpus at positions 1) not at the beginning of the sentence and 2) not immediately after direct speech punctuation, such as quotes or dashes). The model is otherwise independent of an existing linguistic annotation, therefore it is not tainted by eventual systematic errors or shortcomings of existing tools (especially systematic errors in lemmatization are known to skew the results significantly), and can be used even if other NLP processing components are not available for a given language.

The FastText model uses sub-word character n-gram vectors for certain values of n, added to the main word vector, calculated over the space of case normalized word forms. This model captures intra-word information and allows calculation of vectors for the input of words not present in the training corpora. This model is especially useful in searching for compound words or for languages with many such compounds (such as German) or querying for substandard or erroneous inputs. As a convenient side effect, the semantic closeness of the vectors extends to morphemes within words, which is especially visible if there are otherwise no close synonyms to the input.

3. User interface

Conceptually, the web interface (available at https://www.juls.savba.sk/semä. html) consists of several modules, and we strive to provide the most streamlined user experience possible. This is translated into the exact functionality being determined by user input, without additional (or visible) options.

The input consists of two query fields, one for positive words (normalized sum of the corresponding vectors), one for negative words (sum of these vectors will be subtracted from the positive ones). The negative word field is hidden by default and is exposed when the user moves the mouse pointer over the "minus"

sign. The words should be separated by spaces or plus signs. It is also possible to enter negative words into the positive input field if prefixed by the minus sign – this way effectively constructing a simple arithmetic expression.

The following options are available:

- the model, one of lemmata, wordform, fasttext
- known lexicon filter (only for Slovak and only for the *lemmata* model)
- visualization method

The result of the query consists of three fields:

- informational message
- lexical similarities table
- visualization

The "informational message" field contains optional messages for the user, mostly about the status of the server (e.g. the backend is not working, the server is overloaded), or if an unknown word has been queried.

In case of a successful query (no error and the queried word is known to the model), the lexical similarities table and the visualization is displayed.

The lexical similarities table consists of three columns, the first one shows the similarity coefficient (rounded to three decimal places), the second column the word (semantically close to the query) and the third column links to external resources, with a letter-like symbol denoting the type of link.
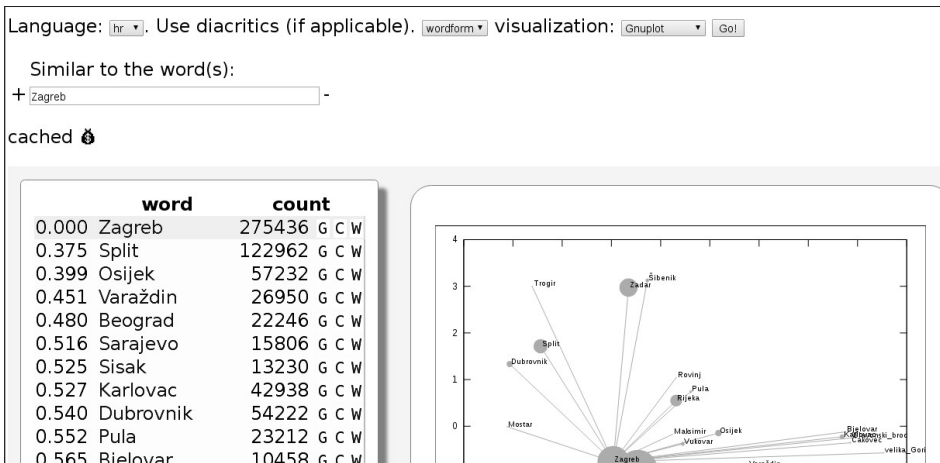
Figure 1: Example of the user interface

From top to bottom, left to right, there are these elements: language selector (hr), model selection (wordform), visualization selection (Gnuplot), input field (Zagreb), informational message ("cached"), lexical similarities table (with the columns containing semantic difference, word, raw count in the corpus, hyperlinks to external sources), and the visualization field.

The links to external resources are (together with the hyperlink symbols):

– link to the Aranea corpus[2], version Minus; a

– link to the Aranea corpus, version Maximum (if available; otherwise Maius; requires registration); A

– link to the Google search[3] for the word, restricted to the top level domain typical for the language[4]; G

– link to English language Wiktionary[5] entry for the word; W

– link to the Slovak National Corpus search interface[6] (only for the Slovak language; requires registration); P

---

[2]   http://aranea.juls.savba.sk

[3]   https://google.com

[4]   With the following exceptions: Arabic, English, Georgian and Latin are not restricted; German is restricted to the .de, .at and ch domains; Portuguese is restricted to the .pt, .br, .ao and .mz domains; Russian is restricted to the .su, .ru, .by, .бел and .рф domains.

[5]   https://en.wiktionary.org

[6]   https://bonito.korpus.sk

– link to the Slovak Dictionary Portal[7] (only for the Slovak language); S

– link to Yandex Search[8] (only for the Russian language); Я

– link to the search interface of the Dictionnaire de l'Académie Française[9] (only for the French language); F

– link to the hrWaC corpus search interface[10] (only for the Croatian language); C

– link to the slWaC corpus search interface[11] (only for the Slovene language); C

– link to the Slovene dictionary portal *Fran*[12] (only for the Slovene language); F

The individual words in the semantic similarities table are hypertext linked to a query of that word using the same language model and options. The number of rows (returned results) can be increased by clicking on the down arrow symbol at the bottom of the table.

4. Visualization

The visualization is especially important in quickly providing information about semantic clusters or the relation of semantically close words to the queried word. We are using the ISOMAP method of dimensionality reduction (Tenenbaum, De Silva and Langford 2000) to get a presentable scatter-like graph of the semantic neighborhood of the query. We are using reduction to two dimensions for the basic graph, reduction to three dimensions for a three-dimensional graph (displayed in 2D projection), and since there are people able to conceptualize and perceive four spatial dimensions (Francis and Brinkmann 2009), there is also a possibility to display a four dimensional graph, as a colour-coded 3D graph projected to a 2D screen.

---

[7] https://slovnik.juls.savba.sk
[8] https://yandex.ru
[9] https://academie.atilf.fr
[10] http://nl.ijs.si/noske/all.cgi/first_form?corpname=hrwac
[11] http://nl.ijs.si/noske/all.cgi/first_form?corpname=slwac
[12] https://fran.si

The graphs in SVG format are produced by the Gnuplot[13] software (Janert 2010), running on the server. Gunplot is a mature, well-established visualization multiplatform software with a rich set of capabilities, and its batch-like mode makes integrating into a web-centric interface rather straightforward and effortless. The software is also very efficient and generating the graphs takes a virtually negligible amount of resources. The raw dimensionality reduced coordinates in Gnuplot format are given as a link (⚙), providing the Gnuplot-enabled users with the ability to pan, zoom and rotate the graphs (though rotation is possible only in planes perpendicular to the ana-kata axis). Users can also use the raw data in the visualization or other statistical software of their choice.

There are also two additional visualization modules available, providing two different word clouds – a static image[14] and a dynamic rotating sphere[15]. These are provided purely for demonstration or illustrative purposes.

5. Usage Examples

5.1. Near Synonyms

Perhaps the most basic application of the interface is as an extensive collection of thesauri in various languages, each of them displaying not only synonyms of the queried word, but also quantifying the semantic difference of the synonym (in the lexical similarities table). In the visualization field, we can immediately spot prominent clusters of semantically similar words, which can bring new insights into semantic relations and behavior.

Table 2: Semantic similarities table for the query *djevojka*,
Croatian lemmata model

0.000	djevojka	130074
0.346	mladić	50694
0.420	djevojčica	57120

---

13 http://gnuplot.info
14 Based on wordcloud2.js software, https://wordcloud2-js.timdream.org.
15 Based on jsTagSphere software, http://jstagsphere.sf.net.

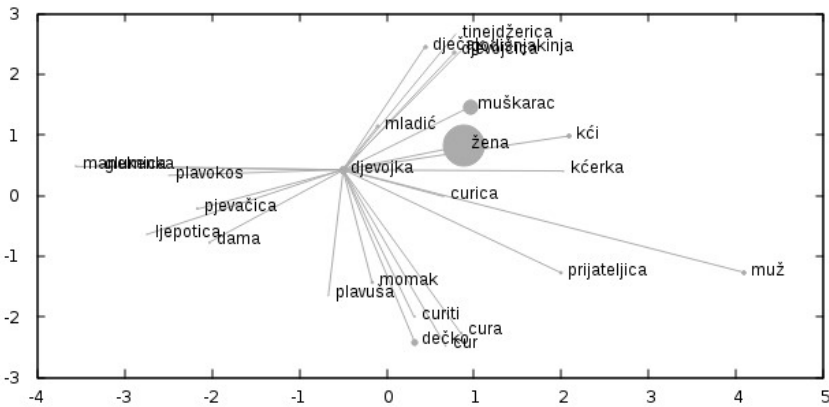| 0.502	dečko	99050
0.508	curiti	24738
0.509	dječak	57554
0.516	cura	35098
0.531	muškarac	233772
0.536	žena	670474
0.549	momak	42770
0.574	ljepotica	12390

Figure 2: 2D visualization of the query *djevojka*. Several
semantic clusters are visible

Table 3: Semantic similarities table for the query *kralj*, Croatian lemmata model

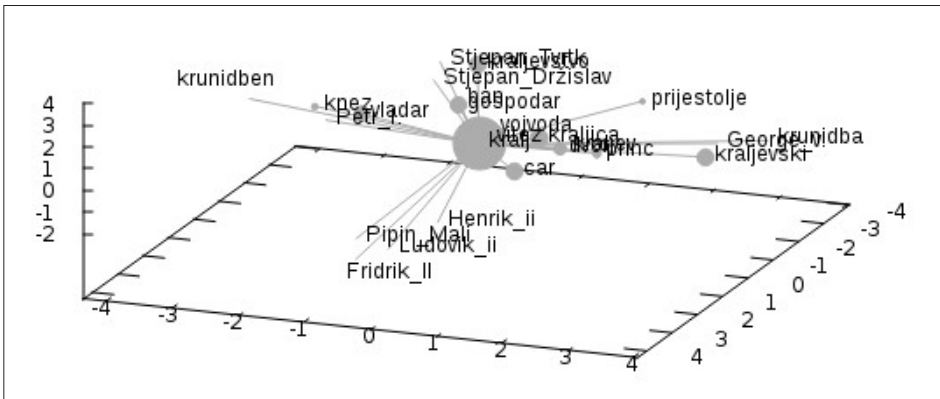| 0.000	kralj	83678
0.511	vladar	24612
0.541	car	26474
0.564	knez	10444
0.625	kraljica	28280
0.635	prijestolje	8526
0.641	Petr_i.	70
0.644	gospodar	25802
0.665	princ	12992
0.690	vitez	14742

Figure 3: 3D visualization of the query *kralj*. Several semantic clusters are visible

5.2. Vector Arithmetic

Since we have the words represented as vectors in a multidimensional Euclidean[16] space, we can perform simple vector arithmetic (addition and subtraction) corresponding to addition and subtraction of *meanings* of the words in the traditional sense.

The usual example used to demonstrate word embedding arithmetic is to subtract the masculinity from the word *king*, add femininity and get the word *queen*, using the equation:

$$king + woman - man = queen \quad (1)$$

which we demonstrate on the Croatian model, i. e. the equation will be of the form:

$$kralj + žena - muškarac = x \quad (2)$$

First, the semantic space around the single query *kralj* is show in Table 3 and Figure 4; we can recognize several semantic groups, e.g. that of important European kings (*Henrik II*, *Ludovik II*, *Pipin Mali*, *Fridrik II* – bottom group),

---

important Croatian/Bosnian rulers (*Stjepan Tvrtko*, *Stjepan Držislav* – upper group), other titles and rulers (*gospodar*, *ban*, *vojvoda*, *car*, *princ*, *kraljica* – in the middle of the graph) and several semantically less connected, solitary words (*prijestolje*, *kraljevski*).

Table 4: The most similar vectors to the query x = *kralj + žena – muškarac*

0.193	kralj	83678
0.608	kraljica	28280
0.621	knez	10444
0.642	vladar	24612
0.649	Petr_i.	70
0.661	car	26474
0.684	prijestolje	8526
0.688	dvor	19908
0.698	Pipin_Mali	112
0.699	krunidba	658
0.711	vitez	14742
	...	
0.987	žena	670474

Table 5: The most similar vectors to the query x = *premijer + žena – muškarac*

0.201	premijer	105350
0.472	premijerka	18732
0.493	vlada	401842
0.523	Jadranka_Kosor	29904
0.537	premijerka_Jadranka	7882
0.538	Ivo_Sanader	14840
0.540	Zoran_Milanović	19544
0.553	Kosor	32200
0.580	premijerka_Kosor	5390
0.585	potpredsjednik_vlada	12096
0.598	Iva_Sanader	12796
	...	
0.978	žena	670474

The results of equation (2) nearest to the calculated vector *x* are in Table 4. Unsurprisingly, the nearest word not equal to the input is *kraljica*, with other vectors being close because of their closeness to *kralj* (in other words, there are no other clearly feminine near-synonyms). Note that we calculate the semantic difference relative to the vector *x* and always include all the input words with positive arithmetic signs in the results, to give the user an idea of how really close is the result to the words presented in the table. This alleviates some of the concerns risen in (Nissim et al. 2019)[17]. The connecting lines in the visualization graph originate at the position of the vector *x* as well; however, their length is necessarily distorted and generally does not correspond to the semantic closeness. Also, note that the word nearest to the *x* of equation (2) happens to be *kralj* itself; *kraljica* is the second nearest.

In modern times, of course, a monarchy is not typical for Croatia, and we do not expect so many texts about monarchies in the web corpus. We thus repeat the query with a more modern example:

$$premijer + \check{z}ena - mu\check{s}karac = x \qquad (3)$$

The results are in Table 5; as expected, the nearest word (apart from premijer itself) is premijerka, but we also see several close proper names, most prominently (feminine) Jadranka Kosor[18]. The table also incidentally illustrates automatic bigram detection – all the bigrams (Jadranka Kosor, premijerka Jadranka, Ivo Sanader, etc.) were automatically inferred from the corpus data; and errors in the lemmatization – Iva Sanader is such an error[19]. It also demonstrates the closeness (in the abstract semantic space, which our vector space is hopefully a reasonably adequate model of) of proper nouns to common nouns.

---

[17]   The idea to originate the connecting lines in the visualization at the resulting vector (and not the nearest word) and to include semantic closeness between this origin and original words in the tables has been inspired by (Nissim et al. 2019), brought to our attention by the anonymous reviewer of this article. We are also grateful for the comments provided by the reviewer.

[18]   Prime minister of Croatia from 2009 to 2011.
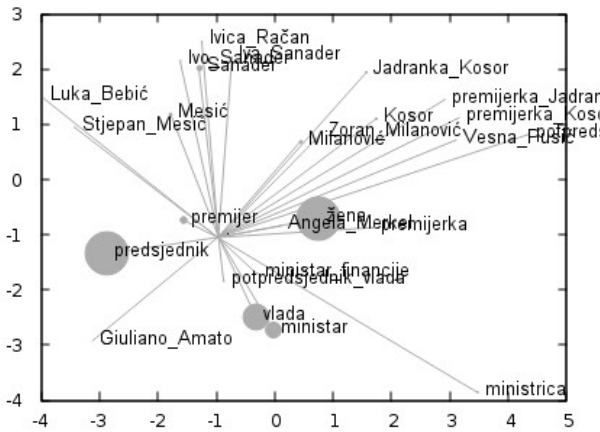
[19]   The correct lemma is Ivo Sanader.

Figure 4: 2D visualization of the query *premijer + žena – muškarac*; connecting lines originate at the vector representing the arithmetic result

5.3. Quantification of Semantic Difference

Another mode of operation is activated automatically if exactly two terms are entered into the positive query field, and it calculates and displays semantic difference, defined as sin θ, where θ is the angle between the two vectors. To illustrate the results, Table 6 contains several pairs of synonyms, not quite synonyms and also word pairs with a rather low semantic difference that are however not considered synonyms in the traditional sense at all.

Table 6: Quantification of semantic differences for several word pairs

first word	second word	sin θ
kralj	premijer	0.885
kralj	knez	0.564
kralj	car	0.541
lingvistika	jezikoslovlje	0.661
premijer	predsjednik	0.628
apoteka	ljekarna	0.427
tri	dva	0.162
osam	sedam	0.120
euro	kuna	0.327
Beč	Budimpešta	0.339
Slovačka	Češka	0.240

6. Conclusion

Word embedding is an indispensable method in modern Natural Language Processing. By presenting simple, yet powerful web-accessible interface to various word vector models build upon the Aranea corpora family, we hope to bridge the gap between contemporary NLP and traditional linguistic and lexicographic research and allow lexicographers to consult the rich information that word embeddings trained on huge corpora can provide.

References

Benko, Vladimír. 2014. Aranea: Yet Another Family of (Comparable) Web Corpora. *Text, Speech and Dialogue. 17th International Conference, TSD 2014*. Eds. Sojka, Petra et al. Springer International Publishing Switzerland. Brno. 257–264.

Erjavec, Tomaž; Ljubešić, Nikola; Logar, Nataša. 2015. The slWaC corpus of the Slovene Web. *Informatica: an international journal of computing and informatics* 39/1. 35–42.

Francis, George K.; Brinkmann, Peter. 2009. Human four-dimensional spatial intuition in virtual reality. *Psychonomic Bulletin & Review* 16/5. 818–823.

Gajdoš, Ľuboš; Garabík, Radovan; Benická, Jana. 2016. The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. *Studia Orientalia Slovaca* 15/1. 21–33.

Garabík, Radovan. 2008. Storing Morphology Information in a Wiki. *Lexicographic Tools and Techniques*. IITP RAS. Moscow. 55–59.

Harris, Zellig S. 1954. Distributional structure. *Word* 10/2–3. 146–162.

Janert, Philipp K. 2010. *Gnuplot in action: understanding data with graphs*. Manning Publications Co. New York.

Ljubešić, Nikola; Klubička, Filip. 2014. {bs, hr, sr}wac-web corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. Eds. Bildhauer, Felix; Schäfer, Roland. Association for Computational Linguistics. Gothenburg. 29–35.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*. Université de Montreal. Scottsdale.

Mikolov, Tomas; Grave, Edouard; Bojanowski, Piotr; Puhrsch, Christian; Joulin, Armand. 2018. Advances in Pre-Training Distributed Word Representations. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association. Miyazaki.

NISSIM, MALVINA; VAN NOORD, RIK; VAN DER GOOT, ROB. 2019. Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. *Computational Linguistics* 46/2. 487–497. doi.org/10.1162/coli_a_00379.

ŘEHŮŘEK, RADIM; SOJKA, PETR. 2010. Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Framework*. Eds. Witte, René et al. ELRA. Valletta. 46–50.

RYCHLÝ, PAVEL. 2007. Manatee/Bonito – A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Masaryk University. Brno. 65–70.

Slovenský národný korpus – prim-8.0-juls-all. 2018. Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied. Bratislava. http://korpus.juls.savba.sk (accessed 28 December 2019).

TENENBAUM, JOSHUA B.; DE SILVA, VIN, LANGFORD, JOHN C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290/5500. 2319–2323. doi.org/10.1126/science.290.5500.2319.

Vektorski prikaz riječi utemeljen na velikim mrežnim korpusima kao moćan leksikografski alat

Sažetak

Projekt *Aranea* sadržava niz usporednih korpusa za 24 (uglavnom europskih) jezika. On pruža prikladan skup podataka za aplikacije za obradu prirodnoga jezika (NLP) koje zahtijevaju učenje na velikoj količini podataka. U radu se prikazuju modeli vektorskoga prikaza riječi koji su uspostavljeni učenjem na korpusima *Aranea* te mrežno sučelje kako bi se propitali modeli i vizualizirali rezultati. To može biti korisno za leksikografsku praksu, ali i u drugim područjima leksikografskoga proučavanja jer je vektorski prostor vjerodostojan model semantičkoga prostora značenja riječi. Postoje tri moguća modela: prvi za kombinaciju vrste riječi i leme, drugi za sirove forme riječi i treći koji se temelji na algoritmu FastText koji upotrebljava vektore na razini nižoj od riječi i nije ograničen na cijele riječi ili poznate riječi pri pronalaženju semantičkih odnosa. U radu se opisuju sučelje i osnovni modeli njegova funkcioniranja, ali se ne pokušava provesti iscrpna jezična analiza prikazanih primjera.

Keywords: corpus, word embedding, vector similarity, semantic similarity, web corpora, visualization

Ključne riječi: korpus, vektorski prikaz riječi, vektorska sličnost, semantička sličnost, mrežni korpusi, vizualizacija