sciendo
DOI 10.2478/jazcas-2022-0023

CHINESE LANGUAGE WORD EMBEDDINGS BASED ON THE CORPUS HANKU

RADOVAN GARABÍK
Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

GARABÍK, Radovan: Chinese language word embeddings based on the corpus Hanku. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 996 – 1004.

Abstract: Vector models based on word embeddings are an indispensable part of advanced Natural Language Processing research and language analysis. We describe several Chinese language (Pǔtōnghuà) word embeddings, the differences from "western" language models caused by specific orthographic and linguistic features of the written Chinese language, and introduce a publicly available web interface for querying the vector models, aimed at linguistically or pedagogically oriented users.

Key words: word embeddings, Chinese, Pǔtōnghuà, corpus, NLP

1. INTRODUCTION

Recently, vector models based on word embeddings (Mikolov et al., 2013) became an indispensable part of advanced Natural Language Processing (NLP) research and language analysis. Originally conceived as a method working on raw, linguistically unannotated corpus (on the surface level of word forms), it has been often used in other configurations, e.g. on the space of lemmas, in order to better capture semantic values of the language, or on substring of words in the form of the fastText algorithm (Bojanowski et al., 2017), improving the analysis of inflected languages, without the need of "traditional" lemmatization and related NLP processes.

A vector space obtained by word embeddings is a very good model of semantic relations (compare Şenel et al., 2018); spatial relations between vectors correspond to semantic relations (similarities, differences, semantic categories, semantic clusters) between words. The models also extend into proper names; informally, we will speak about the "semantic closeness" and "synonyms" also for proper names, by which we mean the closeness of vectors in our models.

1.1 Chinese Language

Chinese as a macrolanguage is a group of language varieties of the Sinitic branch of the Sino-Tibetan languages. The modern prestigious and official variety (*Pǔtōnghuà* 普通话) is the common national speech of the Han nationality, using Beijing pronunciation as the standard pronunciation, Beijing speech as the basic dialect, and the model writing

of the modern vernacular prose as the norm for the grammar. It is based on northern dialects, in particular the standard written language is based on Beijing variant of Mandarin Chinese; and this is generally understood nowadays by the term "Chinese language". Modern Chinese language is in many respects, both inherently linguistic and sociolinguistic, quite different from other widespread languages:

- specific writing system, based on morphosyllabic script (*Hànzì* 汉字), where the basic units of the script – graphemes ("characters", *zì* 字) correspond to morphemes and syllables (with exceptions)[1] (Gajdoš, 2012)
- the language is almost completely isolating, words never change their form
- words are mostly bisyllabic
- the discrepancy between the spoken and written forms (Gajdoš, 2014)
- no space between words in writing
- in fact, the very notion of "word" is rather in flux; in Chinese corpus linguistics and NLP, word segmentation is a nontrivial challenge; the concept of "word" is even more weakened by the absence of a word stress and there is a significant disagreement among literate native speakers about the "correct" word segmentation (Sproat et al., 1996)
- significant amount of homophones

In the past, (Mandarin) Chinese has been marked by stark diglossia and stratification, with formal written texts being in Literary Chinese (*wényánwén* 文言文); in some aspects, this has been carried into contemporary language. A decisive factor for the discrepancy between the spoken and written forms, among other things, is the intellectualization of a language – the Literary Chinese is still one of the essential sources that affect the current (written) language in lexis and syntax. A consequence of these trends is the written language, which although based on the spoken language includes such "foreign" elements – the residue of the literary language *wényánwén* (Gajdoš, 2011). One important aspect of Literary Chinese is that words are mostly monosyllabic; later we discuss a vector model where this feature could be relevant.

2. CHINESE WORD EMBEDDINGS

There are some specifics when trying to make word embedding models of Chinese. Given the fact that most words are two characters long (corresponding to two syllables), the fastText algorithm would not be suitable for Chinese, either as written or even in some romanization. In many other languages (especially those using Latin/Greek/Cyrillic scripts) we can easily consider word embeddings to reflect a raw language, escaping the eventual trap of pre-existing linguistic bias, since the only

[1] Contemporary written Chinese often incorporates Latin script (Roman alphabet) elements, either as foreign (or even domestic) proper names (e.g. CNN, CCTV, QQ), abbreviations, or internet slang (e.g. CNM), often in combination with Arabic digits or Hànzì characters (2B, A片); this phenomenon is noticeably present already for some time (Hansell, 1994).

necessary prerequisite is tokenization, which can be performed quite efficiently and even universally (see e.g. Michelfeit et al., 2014). In Chinese, tokenization into words requires either statistical or rule based methods, introducing some amount of errors, and the exact way of segmenting text into words (or, looking from the opposite side, grouping individual characters into words) is subject to interpretation.

We compiled three models of simplified Chinese, based on the same source, the Chinese web corpus Hanku (Gajdoš et al., 2016) and the Chinese literature subcorpus.[2] The size of the web corpus is 1215 480 206 unicode characters; tokenized into words (*cí* 词), the size is 744 709 741 tokens. As expected from a web corpus, it contains a significant number of repeated texts – after deduplicating (on the paragraph level), the size of the corpus is 819 793 592 unicode characters, 501 782 955 tokens. The size of the whole corpus (deduplicated web corpus and the Chinese literature subcorpus) the word embeddings are trained on is 949 902 689 unicode characters, 594 461 715 tokens. The vectors are trained using skip-gram models, with 200 dimensions and a context window of 7 tokens (slight variations in these hyperparameters, as well as switching the model to Continuous Bag of Words do not change the overall results much). The models are downloadable from our webpage[3] in text Gensim format.

2.1 Model trained on the level of individual words

This model, labelled cí 词 is the closest to the usual web embedding usage. Basic units of the text are words, composed of one or several graphemes (characters). Tokenization is performed by ZPar (Zhang – Clark, 2011), with several enhancements – non-Hànzì elements in the text are separated from Hànzì characters, punctuation characters are tokenized individually, sequences of digits forming numerals are grouped together and tokenized as single tokens, similarly sequences of Roman letters are uppercased, grouped and tokenized as single tokens corresponding to words written in Roman alphabet. Roman characters used in conjunction with Hànzì are treated as parts of the word – thus A片 and 二B would be one token each, not two.

2.2 Model trained on the level of individual characters

This model, called *zì* 字 is compiled at the level of characters – basic units of the text are individual Hànzì characters. Almost identically to the *cí* 词 model, Roman alphabet elements are still uppercased and tokenized as separate tokens[4]. Combinations of Hànzì characters and Latin letters or digits are split into individual Hànzì characters and non-Hànzì remains (e.g. A片 will be tokenized as two tokens, A and 片, but 2B will be one token, unlike its variant 二B that is tokenized as two tokens). In this way, we hope to uncover semantic relations of Hànzì characters, if there are any.

[2] The Hanku corpus contains three subcorpora – the web subcorpus, the subcorpus of literary Chinese and the subcorpus of legal Chinese.

[3] https://www.juls.savba.sk/data.html

[4] We forgo the discreteness characterizing Roman letters in Chinese texts.

2.3 Model trained on Hànyǔ pīnyīn representation of words

There is a rather straightforward, though not completely unambiguous, one way transformation of Hànzì characters into their Hànyǔ pīnyīn transliteration (the opposite way is much more ambiguous). We included an automated transcription into Hànyǔ pīnyīn in our source corpus; the transcription was performed by the *xpinyin* package[5], however, no disambiguation of characters with multiple readings has been performed.

Building a special mode of the web interface that translates characters on the fly into Hànyǔ pīnyīn would be rather simple, but there would be no additional linguistic value in such an endeavour – one can always use an on-line transliteration service (see e.g. DZ Translit[6]) to obtain the same results.

Then there is the possibility to compile a vector model directly on the transliterated words (where the syllables within one word are concatenated together). Tokens transcribed in this way correspond to the 词 model, the transcription is a surjective function (each character in our transliteration is assigned only one reading). The model therefore mirrors the semantic relations of the 词 model, with the exception of relations of homophones (multiple characters with identical pronunciation), where we expect the corresponding vectors to fall to a different region of the semantic space, roughly between the expected meanings of the homophone original words in Hànzì (something we are used to when dealing with homonyms in word embeddings in other languages). To facilitate using the model, we mark the tones using digits 1 to 5 (neutral tone has the number 5), not the usual diacritical marks.

3. WEB INTERFACE

3.1 Modes of Operation

Word embeddings are quite easy to use; there are several mature OpenSource software frameworks, libraries and packages in major programming languages, providing both training and querying the models; or the vectors themselves can be imported into a mathematical/statistical software of one's choice. Nevertheless, this approach is somewhat cumbersome for casual users (such as teachers or learners of the language), or in linguistic research. We built a web interface to the models, with the intention to be used by both experienced linguists (or lexicographers) and laymen. The interface and some of the possibilities it offers has already been described (Garabík, 2020) and we just summarize the main points here (focusing on the Chinese language models[7]):

[5] https://lxneng.com/posts/70

[6] http://quest.ms.mff.cuni.cz/cgi-bin/zeman/translit/translit.pl

[7] We build several models per language; all the other (non-Chinese) languages use common methodology and model types (based on lemma, word form and word form using the fastText algorithm), given specific features of the Chinese language and writing system outlined above, this methodology is neither completely applicable nor optimal for Chinese. This is the main reason we treat the Chinese models separately, taking advantage of the features of the writing system to arrive at better results.

– Any (syntactically correct and using words existing in the corpus, i.e. a query that results in a valid vector) query will display a table of nearest words from the embedding model and a visualization graph, displaying the surroundings of the result, in either 2D, 3D or 4D projection, using ISOMAP dimensionality reduction.
– At the most basic usage, the portal works as a souped-up thesaurus. Querying a single word displays a table (see Table 1 for an example) containing words semantically close to the searched term, with a numeric value quantifying the "closeness" (defined as $\sqrt{1-\cos^2\varphi}$, where φ is the angle between the vectors corresponding to the two words). Note that the closeness need not be directly comparable across different models. We also point out that word embeddings do not deal with homonymy/polysemy well – if the same word has two different meanings, its vector will be roughly a mixture of both vectors, i.e. not corresponding to any of them; or, more realistically, one of the meanings dominates and the vector points to this meaning's region of semantic space.
– Querying two or more words displays similar table, showing the vectors close to all of the words (i.e. a normalized sum of their vectors), which reflects words that are semantically similar to all of the input words; the interface additionally shows the semantic closeness ($\sqrt{1-\cos^2\varphi}$) of the first two words, as a simple number from the interval [0, 1] to give the user a hint about the level of their synonymy.
– Simple vector arithmetic, consisting of addition and subtraction, is supported. The result of the expression is used as a vector around which we look for semantically close words and display the table of them in a similar manner to the previous usage cases.
– It is possible to query (uppercased) non-standard words in Roman alphabet or combinations of Hànzì characters and Roman letters or digits; these are treated as bona fide words in the *cí* 词 model and give valuable insight into modern Internet slang, a subset of lexicon that is often not covered by existing dictionaries.

3.2 Usage and Examples

The models can be used as a substitution of a thesaurus; for a given query, we get not only the semantically closest words, but also their semantic closeness – "true" synonyms have the value close to zero.

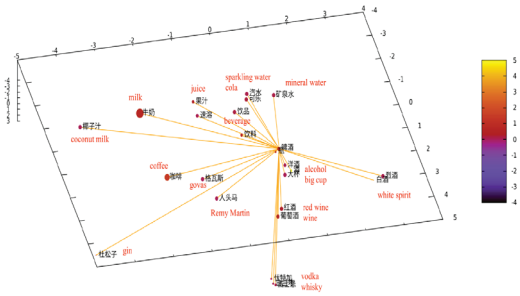
⊠	word	count	
0.000	龙	39537	G 百度 W
0.557	蛇	26177	G 百度 W
0.573	虎	13446	G 百度 W
0.574	争虎斗	47	G 百度 W
0.576	飞龙	1945	G 百度 W
0.598	凤	2769	G 百度 W

Table 1: Semantically closest words to the word lóng 龙 [dragon]. The second column is the word close to the query, the first column is the semantic closeness of the word, the third one number of

occurrences in the corpus, the fourth column contains links to external sources (Google search[8], Baidu search[9], English language Wiktionary[10]). Note that *zhēng hǔ dòu 争虎斗[11] is a phantom word, a relic of incorrect tokenization, as indicated by the low number of occurrences in the corpus (47). The translations of the words from top to bottom are: dragon; snake; tiger; *zhēng hǔ dòu fight against each other; flying dragon; phoenix (Chinese mythological bird).

	word	count	
0.000	TMD	1524	G 百度 W
0.313	他妈	4146	G 百度 W
0.341	煞笔	1043	G 百度 W
0.380	真尼玛	144	G 百度 W
0.387	它妈	209	G 百度 W
0.387	尼玛	4942	G 百度 W
0.397	妈逼	1638	G 百度 W
0.403	脑残	3800	G 百度 W
0.403	特么	2618	G 百度 W
0.404	傻B	805	G 百度 W
0.408	狗日	4421	G 百度 W
0.422	喷子	4146	G 百度 W
0.429	畜生	6882	G 百度 W

Table 2: Semantically closest words to the word (token) TMD – an example of using Roman letters as "native" parts of Chinese texts. The closest word tāmā 他妈 [damn it] with the semantic closeness of 0.313 is almost a synonym. We refrain from providing translations of the table, since we would have to include content warning for the benefit of our more sensitive readers.



Picture 1: 4D visualization of the word query píjiǔ 啤酒 [beer]. The fourth dimension is represented by different colours (probably not visible in the printed version of this article). We can see several semantic clusters around the term. We included translations of the words in the visualization.

[8] https://google.com

[9] https://www.baidu.com

[10] https://en.wiktionary.com

[11] The combination of characters is a part of the idiom *lóng zhēng hǔ dòu* 龙争虎斗 [fierce struggle between two evenly-matched opponents].

If there are at least two query terms (separated by either space or the plus sign), the interface calculates their semantic closeness and displays the value directly. For example, the closeness of the words *Sīluòfákè* 斯洛伐克 [Slovakia] and *Jiékè* 捷克 [Czech(ia)] is 0.312, much closer than *Rìběn* 日本 [Japan] and *Cháoxiǎn* 朝鲜 [North Korea] (0.638), which in our interpretation of the semantic model means that in a typical Chinese text, 日本 and 朝鲜 are perceived as rather different, but 斯洛伐克 and 捷克 are somewhat indistinguishable.

We noticed an interesting result – desemantized single characters in the *zì* 字 model are grouped together. This is somewhat surprising, because in Pǔtōnghuà these characters are almost exclusively used only for their phonetic value (e.g. in foreign language transcriptions) and not their original meaning, and the *zì* 字 model does not otherwise exhibit semantic properties of the individual characters, neither any obvious closeness of other classes of characters.

☒	word	count	
0.000	斯	723111	G 百度 W
0.322	尼	263755	G 百度 W
0.395	帕	43605	G 百度 W
0.397	尔	570967	G 百度 W
0.414	姆	67231	G 百度 W
0.456	迪	110605	G 百度 W
0.463	拉	482471	G 百度 W
0.483	弗	49477	G 百度 W
0.507	蒂	68818	G 百度 W

Table 3: Querying the 字 model for the character sī 斯. The original meaning of the character is desemantized and it is used only for its phonetic value sī. The whole region of our vector space (i.e. the semantic space) around this character is devoid of meaning – all the "semantically close" characters returned by our vector model (in the table) are used only for their phonetic values. From top to bottom: sī, ní, pà, ěr, mǔ, dí, lā, fú, dì.

3.3 Vector Arithmetic

One of the distinguishing, powerful and somewhat surprising features of word embedding models is working vector arithmetic – subtraction and addition of words has straightforward semantic interpretation, as a transfer to a different place in the multidimensional semantic space. Our web interface supports simple vector arithmetcs, consisting of addition and subtraction of (arbitrary number of) vectors.

The prototypical example used to demonstrate vector arithmetic in word embeddings is the "equation" *king* − *man* + *woman* = *queen* (or a local language equivalent), and we would like to use an appropriate Chinese language equivalent for demonstration purposes. The Chinese term for *king*: *guówáng* 国王 is an unassuming word not deeply connected with Chinese history, thus we use *huángdì*

皇帝 [emperor] instead. The equation 皇帝 − 男人 + 妇女 (i.e. *huángdì* 皇帝 [emperor] − *nánrén* 男人 [man] + *fùnǚ* 妇女 [woman]) gives *tàihòu* 太后 [empress dowager or the mother of an emperor] as the semantically closest frequent word (almost as expected; although not what we usually get for the query in "European" languages, it is quite understandable given Chinese history[12]). On the other hand, 皇帝 − 他 + 她 (*huángdì* 皇帝 [emperor] − *tā* 他 [he] + *tā* 她 [she]) gives *huánghòu* 皇后 [empress consort or wife of a ruling emperor[13]] and 皇帝 − 他 + 它 (*huángdì* 皇帝 [emperor] − *tā* 他 [he] + *tā* 它 [it]) gives *huángquán* 皇权 [imperial power]. Let's recall that Chinese does not use gendered personal pronouns and the distinction in writing between masculine, feminine and neutrum 3rd person pronouns has been introduced at the beginning of 20th century under the influence of "modern and progressive" western languages; nevertheless, the vector transfer clearly reflects semantic properties of these pronouns as written in modern Chinese.

Demonstrating a geographical example, we know the traditional Chinese drink is tea – what would, in the eyes of the word embeddings model, be the French equivalent? The query 茶叶 + 法国 − 中国 (i.e. *cháyè* 茶叶 [tea leaves] + *Fǎguó* 法国 [France] - *Zhōngguó* 中国 [China]) gives *hóngjiǔ* 红酒 [red wine] as the semantically closest word. We can interpret it as the typical product corresponding to tea leaves, if we make a transfer from the Chinese region of the semantic space to the "French" one (that is, France as written about in the Chinese language corpus). Similarly, 茶 + 法国 − 中国 (i.e. *chá* 茶 [tea] + *Fǎguó* 法国 [France] - *Zhōngguó* 中国 [China]) gives the result *kāfēi* 咖啡 [coffee] as the (whether right or wrong) typical French beverage corresponding to the *tea* in China in the mental image of an average(d) Chinese speaker.

For comparison, 茶 + 日本 − 中国 (i.e. *chá* 茶 [tea] + *Rìběn* 日本 [Japan] − *Zhōngguó* 中国 [China]) gives *qīngjiǔ* 清酒 [sake] as the Japanese semantic equivalent of Chinese *tea* (again, from Chinese perspective).

CONCLUSION

Word embeddings in modern written Chinese benefit from a specific approach, compared to naïve straightforward application of existing algorithms and software tools and packages. Models based on words (*cí* 词) give expected results, conditioned on word segmentation of adequate quality. By tokenizing sequences of Roman letters

[12] In ancient China, empresses were unheard of – there was only one ruling empress, *Wǔ Zétiān* 武则天 of the Zhōu (late Táng) dynasty (and the wife of a ruling emperor was usually not politically significant). Since many emperors ascended the throne as children, the emperor's mother would often possess notable political power. Perhaps the best known example is Empress Dowager *Cíxī* 慈禧 of the Qīng dynasty.

[13] As opposed to the female ruling monarch; both of these roles are covered by the English term *queen*. This is sometimes disambiguated in a European context by two two-word terms *queen regnant* and *queen consort*.

and combinations of non-Hànzì and Hànzì characters we obtain information of semantic relations of these unconventional words, often used in online Chinese slang, a register seldom covered in existing dictionaries.

We provide a web interface for casual or less technically oriented users that provides basic query methods within the word embedding models, returning a list of semantically related results, allowing quantifying semantic relatedness, and providing several visualization methods.

Bibliography

BOJANOWSKI, Piotr – GRAVE, Edouard – JOULIN, Armand – MIKOLOV, Tomáš: Enriching word vectors with subword information. In: Transactions of the Association for Computational Linguistics, 2017, No. 5, pp. 135–146.

GAJDOŠ, Ľuboš – GARABÍK, Radovan – BENICKÁ, Jana: The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. In: Studia Orientalia Slovaca, 2016, Vol. 15, No. 1, pp. 21–33.

GAJDOŠ, Ľuboš: The discrepancy between spoken and written Chinese methodological notes on linguistics. In: Studia Orientalia Slovaca, 2011, Vol. 10, No. 1, pp. 155–159.

GAJDOŠ, Ľuboš: Čínsky jazyk a čínske písmo. In: Historická revue, 2012, Vol. 23, No. 7, pp. 47–50.

GAJDOŠ, Ľuboš: Synsémantické slová v rámci stratifikácie čínskeho jazyka. In: Miscellanea Asiae Orientalis Slovaca. Bratislava: Univerzita Komenského 2014, pp. 121–131.

GARABÍK, Radovan: Word Embedding Based on Large-Scale Web Corpora as a Powerful Lexicographic Tool. In: Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 2020, Vol. 46, No. 2, pp. 603–618.

中华人民共和国中央人民政府: 国务院关于推广普通话的指示, 1956. Available online: http://www.gov.cn/test/2005-08/02/content_19132.htm

HANSELL, Mark: The Sino-Alphabet: The Assimilation of Roman Letters into the Chinese Writing System. In: Sino-Platonic Papers, 1994, Vol. 45, pp. 1–28.

MICHELFEIT, Jan – POMIKÁLEK, Jan – SUCHOMEL, Vít: Text Tokenisation Using unitok. In: 8th Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU 2014, pp. 71–75.

MIKOLOV, Tomáš – CHEN, Kai – CORRADO, Greg – JEFFREY, Dean: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR 2013.

ŘEHŮŘEK, Radim – SOJKA, Petr: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.

ŞENEL, Lutfi Kerem – UTLU, İhsan. – YÜCESOY, Veysel – KOÇ, Aykut. – ÇUKUR, Tolga: Semantic structure and interpretability of word embeddings. In: EEE/ACM Transactions on Audio, Speech and Language Processing, 2018, Vol. 26, No. 10, pp. 1769–1779.

SPROAT, Richard W. – SHIH, Chilin – GALE, William – CHANG, Nancy:. A stochastic finite-state word-segmentation algorithm for Chinese. In: Computational Linguistics, 1996, Vol. 22, No. 3, pp. 377–404.

ZHANG, Yue – CLARK, Stephen: Syntactic Processing Using the Generalized Perceptron and Beam Search. In: Computational Linguistics, 2011, Vol. 37, No. 1, pp. 105–151.