A Mini Review of Word Embedding in Morpho

Huan LI (李卓桓) huan@bupt.edu.cn

Sep 30, 2018

Abstract

The text meaning is essential for AI; Word Embedding is the primary tool today for text meaning, so the Word Embedding is very important. However, most of the Word Embedding is based on words, like Word2Vec, GLOVE. From the view of Computational Linguistics, words are ruled by syntax, and they are formed by the basic unit of the language: morpheme. A morpheme is the smallest meaningful morphological unit of a language that cannot be further divided or analyzed. In other words, morpheme can be described as the minimal units of meaning. In this paper, we believe that the morpheme could help AI understand the text meaning better, and reviewed some Word Embedding technology with Morphology, which incorporating morphological information into Word Embedding.

1 Introduction

Morphology is a study of words. It mainly deals with word formation, examines the relationship between words, and analyzes their constituent elements.

A morpheme is the smallest unit of a word, which has a meaning, lexical or grammatical, and cannot be divided into smaller units. For instance, the word "unpredictable" consists of 3 morphemes – un + predict + able. Un is a prefix, which means "not" and is used in this example to negate the adjective "predictable." The suffix able is used to form adjectives and is usually placed at the end of a verb (useable, loveable, deniable, etc.).

In this paper, we will talk about the Word Embedding in Morphology.

2. Morphology

morphemes is Re-combin(e)-able. See Table 1 (Burstein 2016):

Table 1: (Morpheme Carries Meaning)

Affix	Morpheme	Meaning
Prefix	"re-"	"again"
Suffix	"-able"	"capable of"
Stem	"combine"	"to join"

And it is powerful. See Table 2 (Eckert and Sag 2011)

Table 2: (Study of Word Structure)

Word Structure
pre+pose
pre+pos+ition
pre+pos+ition+al
pre+pos+ition+al+ize
pre+pos+ition+al+iz+ation
pre+pos+ition+al+iz+ation+free
Pseudopseudohypoparathyroidism

That morphology knowledge will also help human to understand the words a lot. There has a famous book named《GRE 词汇精选》(红宝书) which is known by all the Chinese students who want to pass the GRE test, and it gives a lot of morphological tricks like Table 3:

Table 3: (GRE Book)

Word	Morphemes / Meaning
abandon	a+band(乐队)+on→ 一个乐队在演出 → 放纵
abash	ab+ash(灰)→ 中间有灰，灰头灰脸 → 尴尬
abate	a(加强)+bate(减弱，减少)→ 减轻
abbreviate	ab(加强)+brev(短)+iate→ 缩短
abdicate	ab(相反)+dic(说话，命令)+ate→ 不再命令 → 退位，辞职

That knowledge of the common prefixes would also help us in deciding the meaning of new words that we encounter.

3 Word Vector Presentations

When doing NLP(Natual Language Processing) with DNN(Deep Neural Network), we need to input the language to the computer. There are many technics can do this, such as One Hot Encoding for Character, or Vector Representing for Word.

The Vector Representing for Word is a very active field since Word2Vec (Mikolov et al. 2013) from Google, GLOVE (Pennington, Socher, and Manning 2014) from Stanford, and FastText (Niu et al. 2017) from Facebook, etc.

Most algorithms are derivative of Word2Vec: they map words in the training set into vectors. However, this method has many limitations, and the biggest one is that it has a closed vocabulary assumption, so that if a word had not been seen at training, then it could not be understood, this is an OOV(Out of Vocabulary) problem.

FastText has some breakthroughs. It considers each word as a Bag of Character n-grams. This is also called as a Subword-Units in the paper. Instead of dealing of individual words, FastText breaks words into several n-grams (Subword-Units). For instance, the tri-grams for the word `orange` is `ora`, `ran`, `ang`, `nge`. The word embedding vector for `orange` will be the sum of all these n-grams.

Subword-Units will be helpful when we meet an OOV word, like rare and complicated words, because we can analyze it from the characters.

However, the naive Subword-Units had included the morphological linguistics structure for words. The n-gram algorithm will produce too many combinations, which most of them would be meaningless. See Figure 1 (Kudo 2018)

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Figure 1: Subword Regularization

By introducing morphological knowledge, we can split the word into morphemes, which could help us to build representations for morphologically complex words from Subword-units of morphemes. Using Morph-Subword-Units build from morphemes will get better word representation. See Figure 2 (Perets 2009)

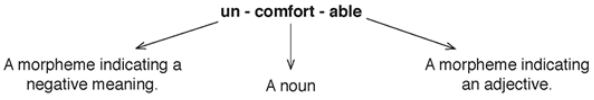

Figure 2: Morphology Tree

The languages other than English, such as Chinese, Japanese, and Korea, also have their word in different morphology. Those languages could use the same idea to present the smallest meaningful morphological unit of the language, for example, Chinese character components. See Figure 3 (Li et al. 2015).

Besides the character components, there's also some researchers go deeper with strokes n-grams. See Figure 4 (Cao et al. 2018).

Moreover, get the sememe for the component of the

transform	meaning	transform	meaning
艹 → 艸	grass	扌 → 手	hand
亻 → 人	human	氵 → 水	water
刂 → 刀	knife	車 → 车	vehicle
犭 → 犬	dog	攵 → 攴	hit
灬 → 火	fire	纟 → 糸	silk
钅 → 金	gold	耂 → 老	old
麥 → 麦	wheat	牜 → 牛	cattle
饣 → 食	eat	飠 → 食	eat
礻 → 示	memory	忄 → 心	heart
罒 → 网	nest	王 → 玉	jade
讠 → 言	speak	衤 → 衣	cloth
月 → 肉	body	辶 → 辵	walk

Figure 3: Chinese Character Component

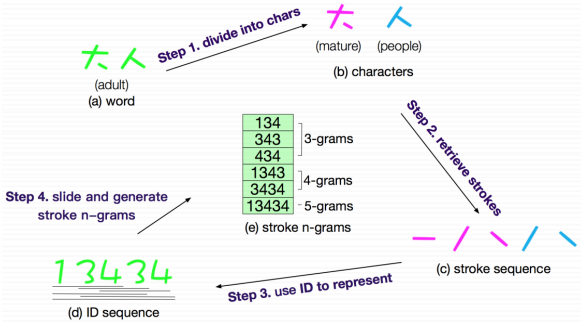

Figure 4: Chinese Character Strokes with N-Gram

Chinese characters from HowNet (Dong and Dong 2003). See Figure 5 (Niu et al. 2017)

法-B	法政 (law and politics), 法例 (rule), 法律 (law), 法理 (principle), 法号 (religious name), 法书 (calligraphy)
法-E	懂法 (understand the law), 法律 (law), 消法 (elimination), 正法 (execute death)
法-I	法律 (law), 法例 (rule), 法政 (law and politics), 正法 (execute death), 法官 (judge)
法-II	道法 (an oracular rule), 求法 (solution), 实验法 (experimental method), 取法 (follow the method)
道-B	道行 (attainments of a Taoist priest), 道经 (Taoist scriptures), 道法 (an oracular rule), 道人 (Taoist)
道-E	直道 (straight way), 近道 (shortcut), 便道 (sidewalk), 半道 (halfway), 大道 (revenue), 车道 (traffic lane)
道-I	直道 (straight way), 就道 (get on the way), 便道 (sidewalk), 巡道 (inspect the road), 大道 (revenue)
道-II	道行 (attainments of a Taoist priest), 邪道 (evil ways), 道法 (an oracular rule), 论道 (talk about methods)

Figure 5: Chinese Word with Sememe

Today, many Subword-Units algorithms want to improve the performance of the language model, like Morfessor, BPE, char-trigram, character, and analysis. See Figure 6 (Vania and Lopez 2017)

Unit	**Output of $\sigma(\boldsymbol{wants})$**
Morfessor	^want, s$
BPE	^w, ants$
char-trigram	^wa, wan, ant, nts ts$
character	^, w, a, n, t, s, $
analysis	want+VB, +3rd, +SG, +Pres

Figure 6: From Characters to Words to in Between

There has already a lot researchs like "implicitly incorporating morphological information into Word Embedding" (Xu and Liu 2017), See Figure 7, and "Better Word Representations with Recursive Neural Networks for Morphology" [luong2013better], See Figure 8 (Luong, Socher, and Manning 2013).

At last, that prior knowledge of morphological is valuable. However, how can we get the morphological

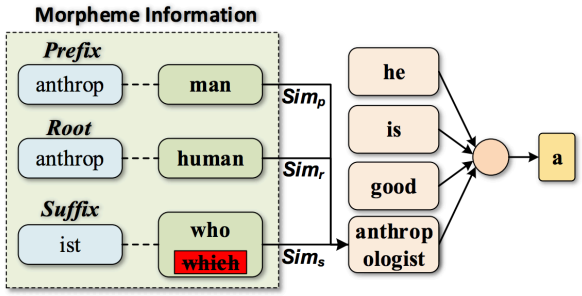

Figure 7: Morphological Information into Word Embedding

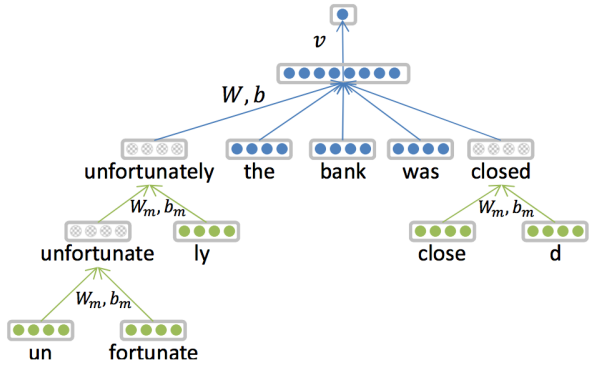

Figure 8: Morphological Sub Words

knowledge? Of course, we can get them from the priory researchers. However, it could be better if we could induct the morphology knowledge from unsupervised machine learning algorithms. See Figure 9 (Soricut and Och 2015)

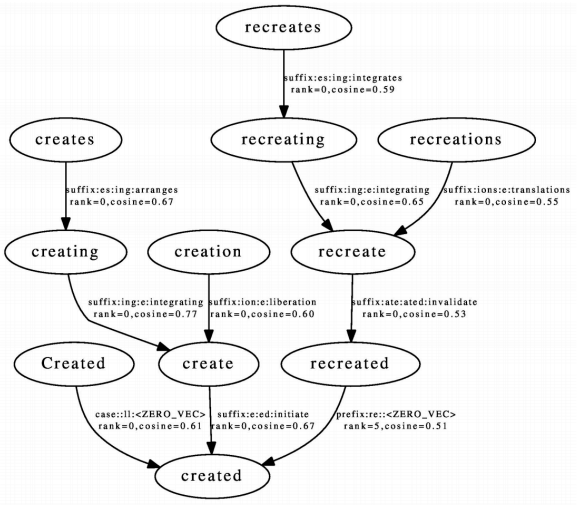

Figure 9: Unsupervised Morphology Induction

Conclusion and Future Work

Morphology knowledge is, and it could be able to help the word embedding to be more informative. There's already lots of word embedding research based on the morpheme, but it seems they are all very early stage, and there still have not a solution for the industry like Word2Vec or GLOVE.

We believe this direction is a right direction to improve the performance of traditional word embedding for NLP tasks, and it is worth to do more research based on the previous studies.

The future work would consider to:

1. Train a Tokenizer to parse Word to Morphemes
2. Train Morpheme Embedding on Wikipedia Dataset

3. Use GRU to convert Morpheme Embeddings for each word to Word Embedding(Morpheme Word Embedding)
4. Benchmark all the morpheme-based word embedding algorithms
5. Compare the results of standard NLP tasks on Morpheme Word Embedding against to other traditional word embeddings.

Acknowledgements

I would like to thank professor Xiaojie LI for I started thinking about embedding in morphology when I am on the class Computational Linguistics that he taught.

I'd also like to thank my friend Tongjun LI, who is the organizer of Wechat group "NLP Fans" with hundreds of members, where I could discuss my idea over there.

See Also

1. Morphology: Word Structure (video)
2. Morphology: The Study of Word Structure (slide)
3. 汉字拼义理论：心理学揭开争鸣百年的汉字之谜 1
4. 汉字拼义理论：心理学揭开争鸣百年的汉字之谜 2 3 汉字拼义理论：心理学揭开争鸣百年的汉字之谜
5. morphology (morpheme & allomorph)
6. A review of word embedding and document similarity algorithms applied to academic text
7. Morphological Recursive Neural Network (morphoRNN)
8. "后 Word Embedding" 的热点会在哪里?
9. 论文阅读笔记 Improved Word Representation Learning with Sememes
10. BETTER WORD REPRESENTATIONS WITH RECURSIVE NEURAL NETWORKS FOR MORPHOLOGY (slide)
11. Word embeddings in 2017: Trends and future directions
12. HowNet - 义原（Sememe）顾名思义就是原子语义，即最基本的、不宜再分割的最小语义单位
13. SentencePiece is a re-implementation of Subword-units, an effective way to alleviate the open vocabulary problems in neural machine translation. SentencePiece supports two segmentation algorithms, byte-pair-encoding (BPE) [Sennrich et al.] and unigram language model [Kudo.].
14. Deep contextualized word representations

References

Burstein, Jill. 2016. "Morphology: Word Structure." 2016. https://www.youtube.com/watch?v=zlkGtu035xc.

Cao, Shaosheng, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. "Cw2vec: Learning Chinese Word Embeddings with Stroke N-Gram Information."

Dong, Zhendong, and Qiang Dong. 2003. "HowNet-a Hybrid Language and Knowledge Resource." In *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*, 820–24. IEEE.

Eckert, Penny, and Ivan A. Sag. 2011. "Linguistics 1: Introduction to Linguistics." 2011. https://web.stanford.edu/class/linguist1/Slides/morph1-slides.pdf.

Kudo, Taku. 2018. "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates." *arXiv Preprint arXiv:1804.10959*.

Li, Yanran, Wenjie Li, Fei Sun, and Sujian Li. 2015. "Component-Enhanced Chinese Character Embeddings." *arXiv Preprint arXiv:1508.06669*.

Luong, Thang, Richard Socher, and Christopher Manning. 2013. "Better Word Representations with Recursive Neural Networks for Morphology." In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 104–13.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Com-

positionality." In *Advances in Neural Information Processing Systems*, 3111–9.

Niu, Yilin, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. "Improved Word Representation Learning with Sememes." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:2049–58.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (Emnlp)*, 1532–43.

Perets, Claire. 2009. "Scared of Hebrew Grammar? Hebrew Morphology Made Simple." 2009. https://www.engheb.com/htm/blog-hebrew_morphology_made_simple.htm.

Soricut, Radu, and Franz Och. 2015. "Unsupervised Morphology Induction Using Word Embeddings." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1627–37.

Vania, Clara, and Adam Lopez. 2017. "From Characters to Words to in Between: Do We Capture Morphology?" *arXiv Preprint arXiv:1704.08352*.

Xu, Yang, and Jiawei Liu. 2017. "Implicitly Incorporating Morphological Information into Word Embedding." *arXiv Preprint arXiv:1701.02481*.