Initializing Convolutional Filters with Semantic Features for Text Classification

Shen Li[1,2] shen@mail.bnu.edu.cn

Zhe Zhao[4,5] helloworld@ruc.edu.cn

Tao Liu[4,5] tliu@ruc.edu.cn

Renfen Hu[3,†] irishere@mail.bnu.edu.cn

Xiaoyong Du[4,5] duyong@ruc.edu.cn

[1] Institute of Chinese Information Processing, Beijing Normal University
[2] UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University
[3] College of Chinese Language and Culture, Beijing Normal University
[4] School of Information, Renmin University of China
[5] Key Laboratory of Data Engineering and Knowledge Engineering, MOE

Appendix

The crux of our initialization technique is n-gram selection, which assists neural networks to extract important n-gram features at the beginning of the training process. In the following tables, we illustrate those selected n-grams of different classes and datasets to understand our technique intuitively. Since all of MR, SST-1, SST-2, CR, and MPQA are sentiment classification datasets, we only report the selected n-grams of SST-1 (Table 1). N-grams selected by our method in SUBJ and TREC are shown in Table 2 and Table 3.

Class	Very Positive	Positive	Neutral	Negative	Very Negative
Unigram	standout perfection releases	heartening virtuosic affectionately	kin reworked michelle	choppiness woozy meager	flopped indescribably atrociously
Bigram	mesmerizing music satisfying evenings most beautiful	with raw remarkable about this much	man vs kin 's the sides	left slightly been conjured ridiculous wig	definitely meaningless wasted nearly is meaningless
Trigram	best films of making it one enjoyable and satisfying	grounded in an remarkable about clung pleasant enough and	even one word speaking even one than to receive	conjured up only difficult to fathom dumbed down approach	devoid of substance is definitely meaningless with this silly

Table 1: Examples of the selected n-grams in SST-1 dataset. The results are self-explanatory. There are five classes in SST-1 dataset. The polarities of n-grams selected from very positive texts to very negative texts change smoothly. Adjectives with positive sentiment are easily selected in positive texts, e.g. "enjoyable", "beautiful" and "satisfying". Obviously, n-grams indicating negative emotions are more likely to be selected in negative texts such as "wasted" and "meaningless".

Class	Subjective	Objective
Unigram	amusing laughs i entertaining	discovers 233 decide boyfriend
Bigram	movie that it does but it the performances	his father him to he finds where he
Trigram	but it 's a movie that if you 're it 's not	is the story the help of falls in love in order to

Table 2: Examples of the selected n-grams in Subj dataset. We can observe that adjectives such as "amusing" and "entertaining" are more likely to be selected in subjective reviews, and neutral words such as "his" and "him" are more likely to be selected in objective reviews.

Class	ABBR.	ENTY.	DESC.	HUM.	LOC.	NUM.
Unigram	abbreviation stand acronym	fear disease animal	why different definition	wrote who portrayed	located country nationality	many average tall
Bigram	the abbreviation stand for abbreviation for	a fear fear of what color	how can how does why do	who was who is who invented	what country what city where is	how many when was how long
Trigram	is the abbreviation the abbreviation for stand for in	a fear of is a fear what color is	how do i how can i the difference between	who was the who is the who invented the	where can i u s state what city is	how many people when was the what year did

Table 3: Examples of the selected n-grams in TREC dataset. Strong indicators of question types are selected by NB weights. For example, “acronym”, “stand for”, and “the abbreviation of” are selected for the abbreviation question type. The n-grams that are related to entities’ attributes such as “disease” and “animal” are selected for the entity question type. Human’s actions (e.g. “portrayed”, “who invented”) are selected for the human question type. “what country” and “what city” possess large NB weights in questions about location. In questions of the number type, “how many” and “when was” are selected.