Affective Multi-Term Analysis using Distributional Semantic Similarity Measures

B.SUREKA, Mrs T.R AKILA DEVI

Abstract— Affective text analysis is a high quality text to handle large amount of noise data and very time consuming that combine the affective rating of multi-word terms. To generating the sentence rating have to improve the performance of unigram and bigram models by using the n-gram models. The affective rating for n-grams is estimated and the corpus based method using distributional semantic similarity metrics between the seed and the unseen words. Finally, it combines to form the sentence level rating through simple algebraic formula. The proposed framework produces the state of art results for word level tasks in English and Germany and the sentence level news headlines classification in semEval'07-Task 14 task. The inclusion of bigram terms provides significant performance improvement even if selection term is not applied.

Index Terms - Affective lexicon, emotion, lexical semantics, natural language understanding, opinion mining, and sentiment analysis

1 INTRODUCTION

Data mining is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving the methods of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure. Text mining is an interdisciplinary field that draws on information retrieval data mining, machine learning, statistics and computational linguistics. A substantial portion of information is stored as text such as news articles, technical papers, books, digital libraries, email messages, blogs and web pages. An important goal is to derive high-quality information from text and it is typically done through the discovery of patterns and trends by means such as statistical pattern learning, topic modeling, and statistical language modeling. Typical text mining tasks include text categorization, text clustering, concept or entity extraction, and production of granular taxonomies, sentiment analysis, document summarization, and entity-relation modeling.

Current text mining approaches can be divided into two main categories: bag-of-words and NLP-based techniques. The first approaches do not exploit morphological structures in the text they are rather ineffective and usually need to although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follows (i) use larger data sets and ignoring less frequently mentioned information, (ii) NLP-based approaches parse the sentences in the text and convert them into parse trees. Parse trees contain some of morphological structures in the text in a more machine readable format, and thus provide a better structure for text analysis.

Semantic similarity is the building block for numerous applications of natural language processing (NLP), such as grammar induction and affective text categorization. Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency or strength degree to which the word, phrase, sentence, or document in question is positive or negative towards a subject topic, person, or idea. The analysis of public opinion, such as the automated interpretation of on-line product reviews, semantic orientation can be extremely helpful in marketing, measures of popularity and success, and compiling reviews.

Distributional semantic models (DSM) are based on the distributional hypothesis of meaning assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. A wide range of contextual features are also used by DSM exploiting lexical, syntactic, semantic, and pragmatic information. DSM has been successfully applied to the problem of semantic similarity computation. Co-occurrence and context features are used to measure the strength of relations. Affective text analysis can happen at various levels, targeting different lexical units: words, phrases, sentences, utterances, as appropriate to the task that express the meaning of the utterance through some combination of the meanings (typically affective ratings) of the words they contain. Word ratings are provided by affective lexica, either manually annotated, such as Affective norms for English Words (ANEW) or, more typically, automatically expanded lexica such as SentiWord Net and WORDNET AFFECT. These word ratings are then combined through a variety of methods, making in use of part-of-speech tags, sentence structure.

2 RELATED WORK

Al Exandros Potamianos et al [2] this existing system describes a new unsupervised approach for the construction of DSM with application to lexical semantic similarity computation.

First, a corpus of snippets is harvested from the web. Then, a semantic network is constructed encoding the semantic relations between words in the corpus. Co-occurrence and context features are used to measure the strength of relations.

Michelle C et al [3] unsupervised adaptation algorithms of the semantic-affective models are proposed. Affective ratings at the utterance level are generated based on an emotional lexicon, which in turn is created using a semantic (similarity) model estimated over raw, unlabeled text. The proposed adaptation method creates task-dependent semantic similarity models and task dependent word/term affective ratings.

Shrikanth Narayanan et al [4] invested a method of affective text analysis and modeling that is capable of generating continuous valence ratings at the sentence level starting from word and multi-word term valence ratings. Motivated from the language modeling literature, a back-off algorithm is employed to efficiently fuse the valence of single-word and multi-word terms.

Peifeng Li et al [5] proposes a method to add more information to the dependency tree and provide an algorithm to prune dependency tree to reduce the noisy, and then to the sentence-level sentiment classification.

Zhang wei et al [6] investigated a system of automatic speech sentence segmentation from large multi-paragraph speech databases with corresponding text. The novelty of our approach lies in the facts that: a) an automatic speech sentence segmentation system b) an improved automatic speech sentence segmentation system.

Eneko Agirre et al [7] the paper system presents and compares Word Net based and distributional similarity approaches. The strengths and weaknesses of each approach regarding similarity and relatedness tasks are discussed, and a combination is presented.

Danushka bollegala et al [8] investigated a Web-based metrics that compute the semantic similarity between words or terms are presented and compared with the state of the art. Starting from the fundamental assumption that similarity of context implies similarity of meaning, relevant. The proposed algorithms work automatically; do not require any human-annotated knowledge resources.

3 FLOW DIAGRAM

In the Figure.1, keyword search module the user search the data in the search engine using the keywords. Keywords are words or phrases that describe the content. They can be used as metadata to describe images, text documents, database records, and Web pages. Here is used text as a keyword to search the data. In the word level tagging, gives the ratings for each word. In this multi-word tagging, more than a word will be selected and ratings for the search snippets is obtained. Multiword terms (MWTs) are relevant strings of words in text collection. Sentence level tagging is similar to Multi-word tagging; the sentence level tagging gives the ratings by fusing the single word and multi-word ratings.

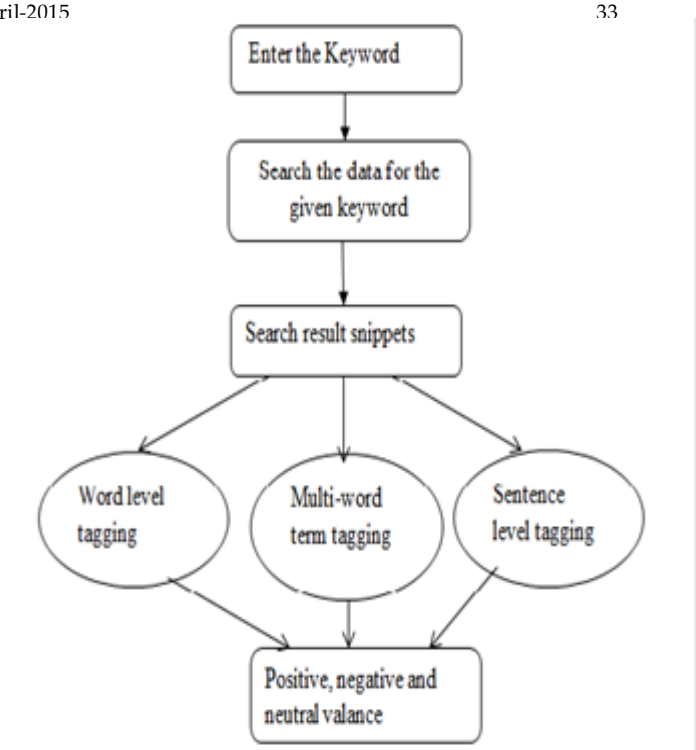

Fig.1. Search a query from the search engine

4 METHODOLOGY USED

A subset of words is automatically selected from the lexicon to serve as seed words for the affective model. The affective rating for a new word/term is estimated as a linear combination of the products between semantic similarities and affective ratings of the seed words. A seed word with high affective (or semantic) variance might be a less robust predictor of the affective scores of unseen words. Words with high affective variance typically have multiple part-of-speeches, tags and word senses, or their valence rating is highly context.

A.WORD LEVEL TAGGING

The similarity between two words was estimated as the cosine of their respective vectors. In our work, two types of metrics are investigated for weighting the strength of the link between a reference noun and its neighbours, namely, co-occurrence based and context-based similarity measures.

1) Co-Occurrence Based Similarity Metrics: To estimate the similarity between two words/terms using the frequency of co-existence within larger lexical units (sentences, documents). The underlying assumption is that terms that co-exist often are likely to be related semantically. One popular method to estimate co-occurrence is to pose conjunctive queries to a web search engine; the number of returned hits is an estimate of the frequency of co-occurrence.

2) Context Based Similarity Metrics: The basic assumption behind these metrics is that similarity of context implies similarity of meaning, i.e., terms that appear in similar lexical environment have a close semantic relation. Similarity between features extracted from text context using "bag of words" and metrics used as a cosine similarity.

B. MULTI-WORD TERM TAGGING

Multiword terms (MWTs) are relevant strings of words in text collections. The multiword terms points to relevant information, corresponding to topics and subtopics in the text collection, and be quite useful specially for highly refining generic queries. LocalMaxs algorithm is introduced for automatically extracting multiword terms. This algorithm requires neither empirically suggested thresholds nor complex linguistic filters nor language specific morph-syntactic rules. These features make this algorithm a suitable approach to extract MWTs from text collections written in any language. The LocalMaxs is an algorithm that works with any text collection as input and automatically produces multiword terms (MWTs) from that text collection.

- Localmaxs algorithm

  The Localmaxs algorithm accepts a text as input and generates multiword units from that text. For each phrase, the algorithm looks at all sub phrases which are one word shorter contained within that phrase as well as all super phrases which are one word longer that contain the phrase. When the phrase's association measure achieves its maximum value compared to those for all the sub phrases or super phrases, the phrase is considered a probable multiword unit.

  More precisely, the local maxs algorithm is as follows.

  - Let assoc be an association measure.
  - Let W be a phrase.
  - Let length() be a function returning the number of words in a phrase.
  - Let sub W be the set of all sub phrases contained in W of size (length(W) - 1).
  - Let super W be the set of all super phrases in the text of size (length(W) + 1) which contain W as a sub phrase.

  Then for all x phrases in sub W, and for all y phrases in super W, W is a multiword unit if:

  - (length(W) = 2) and (assoc(W) > assoc(y)), or (length(W) > 2) and (assoc(W) >= assoc(x)) and (assoc(W)> assoc(y))

C. SENTENCE LEVEL TAGGING

Linear fusion assumes that the words should be weighted equally of their strong or weak content. Sentence have increase in polarized term might end up having low absolute valence. In the propose work weighted average scheme increase the absolute valence values. Each word's affective score is the problem in the sentence level score. In many compound expressions where their semantic and affective content cannot be accurately estimate the sum of words. Negation can alter the meaning and affective scores. Sentence level rating can be found by combining the single and the multi word terms (i.e.) unigram and bigram models.

D. FUSION OF N–GRAM MODELS

A subset is used as seed words and the affective ratings of new words/terms are all expressed as a weighted linear combination of their semantic similarities to these seed words multiplied with the affective ratings of the seeds. The language modelling literature, a back-off algorithm is employed to efficiently fuse the valence of single-word and multi-word terms. Specifically, a term detection criterion is used to select the appropriate n-gram terms, starting with bigrams and potentially backing off to unigrams.

The methods for combining word ratings into sentence ratings vary significantly. Usually a word selection step is involved, which removes non-relevant words from the sentence and keeps only those considered to carry affective significance. The actual combination of ratings is usually done by some simple numeric method, like taking the arithmetic mean. For bigrams that appear rarely in our corpus it may be advantageous to back-off to a unigram where adequate statistics to accurately estimate affective scores exist. The method has no requirement that limiting it to estimating word ratings or even limiting it to any specific language. To apply to bigrams, only the semantic similarity metric has to be extended to handle both unigrams and bigrams.

The main fusion strategies to improve the performance of unigram and bigram models are interpolation, Back off and the weighted interpolation.

1) Interpolation: For sentence that consists of the word Sequence w1,w2…wN create a unigram and bigram affective model, respectively.

2) Back-Off: Instead of using the interpolating the affective scores of different n-grams models, a criterion for alternating between the unigram and bigram model is proposed.

3) Weighted interpolation: weighted interpolation extends the interpolation and back-off models. The final collection of terms will include all unigram in the sentence and some of bigrams to produce the final sentence rating.

5 IMPLEMENTATION AND RESULTS

A. Similarity Metric Selection

The co-occurrence based similarity metrics can estimate the co-occurrence counts either using web hits. They can co-occur within a document at any distance. To calculate the frequency statics [Fig. 2] the co-occurrences (i.e.) the individual word's number of occurrences as well as number of times that the two words co-exists within a set distance by using the web search engine.

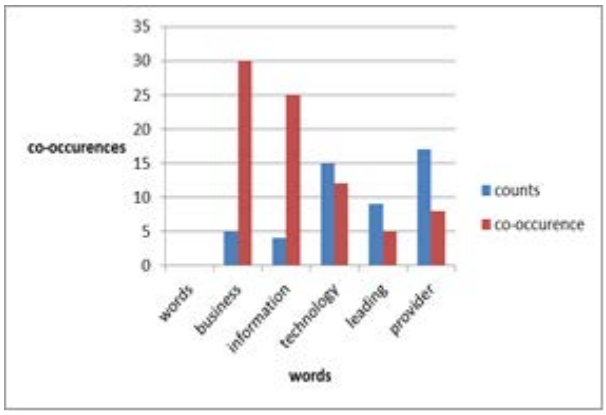

Fig.2: Similarity metrics

6 Conclusion

A method of creating sentence affective ratings based on the combination of partial affective ratings of word n-grams. An affective lexicon expansion algorithm capable of creating continuous n–gram affective ratings based on a set of manually labelled seed words and semantic similarity ratings calculated over web data. Multi term word tagging more than a word will be selected and ratings for the search snippets is obtained. Sentence level ratings were obtained from n-gram ratings using linear and non-linear fusion methods. Interpolation and back-off models were proposed for combining unigram and bigram affective ratings.

Reference

[1] Nikolaos Malandrakis, Student Member, IEEE, Alexandros Potamianos, Senior Member, IEEE, Elias Iosif, Student Member, IEEE, and Shrikanth Narayanan, Fellow, IEEE "Distributional Semantic Models for Affective Text Analysis",IEEE international transactions on audio, speech, and language processing, vol. 21, no. 11, november 2013

[2] E L Ias Ios I F And Al Exandros Potamianos "Similarity Computation Using Semantic Networks Created From Web-Harvested Data", IEEE international conferences, vol.1, pp.1-30, 2012

[3] Nikolaos malandrakis1, alexandros potamianos1, kean j. hsu2, kalina n. babeva2, michelle c. feng2, gerald c. davison2, shrikanth narayanan1 " Affective language model adaptation via corpus selection" IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) 2014.

[4] Nikolaos Malandrakis1, Alexandros Potamianos2, Shrikanth Narayanan1 "continuous models of affect from text using n-grams" 2013.

[5] Peifeng Li, Qiaoming Zhu, Wei Zhang "A Dependency Tree based Approach for Sentence-level Sentiment Classification" 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2011.

[6] Zhang Wei, Pang Minhui, Du Ranran, Liu Yayu "Automatic Speech Sentence Segmentation from Multi-paragraph Databases" International Conference on Measuring Technology and Mechatronics Automation 2010.

[7] Eneko Agirre† Enrique Alfonseca‡ Keith Hall‡ Jana Kravalova‡§ Marius Pas¸ca‡ Aitor Soroa† "A Study on Similarity and Relatedness Using Distributional andWordNet-based Approaches"2011.

[8] Elias Iosif, Student Member, IEEE, and Alexandros Potamianos, Senior Member, IEEE "Unsupervised Semantic Similarity Computation between Terms Using Web Documents" IEEE transactions on knowledge and data engineering, vol. 22, no. 11, november 2010.

[9] Danushka bollegala, yutaka matsuo, and mitsuru ishizuka, member, ieee "a web search engine-based approach to measure semantic similarity between words" ieee transactions on knowledge and data engineering, vol. 23, no. 7, july 2011.

[10] Dat Huynh, Dat Tran, Wanli Ma "Combination Features for Semantic Similarity Measure" Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I, IMECS 2014, March 12 - 14, 2014, Hong Kong.

[11] Hsun-hui huang and yau-hwang kuo, member, ieee "cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach" ieee transactions on fuzzy systems, vol. 18, no. 6, december 2010.

[12] M. Baroni and A. Lenci, "Distributional Memory: A General Framework For Corpus-Based Semantics," Comput. Linguist., vol. 36, no. 4, pp. 673–721, 2010.

[13] K. Moilanen, S. Pulman, and Y. Zhang, "Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression," in Proc. WASSA Workshop ECAI, 2010, pp. 36–43.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Kernel models for affective lexicon creation," in Proc. Interspeech, 2011, pp. 2977–2980.

[15] A. Hassan and D. Radev, "Identifying text polarity using random walks," in Proc. ACL, 2010, pp. 395–403.