Corpus based Amharic sentiment lexicon generation

Girma Neshir[1], Andreas Rauber[2] and Solomon Atnafu[3]

[1]Addis Ababa University, IT Doctoral Program, AAU, Ethiopia,

[2]Technical University of Vienna, Institute of Information Systems Engineering, Austria,

[3]Addis Ababa University, Department of Computer Science, Ethiopia

[1]girma1978@gmail.com, [2]rauber@ifs.tuwien.ac.at and [3]solomon.atnafu@aau.edu.et

Abstract

Sentiment classification is an active research area with several applications including analysis of political opinions, classifying comments, movie reviews, news reviews and product reviews. To employ rule based sentiment classification, we require sentiment lexicons. However, manual construction of sentiment lexicon is time consuming and costly for resource-limited languages. To bypass manual development time and costs, we tried to build Amharic Sentiment Lexicons relying on corpus based approach. The intention of this approach is to handle sentiment terms specific to Amharic language from Amharic Corpus. Small sets of seed terms are manually prepared from three parts of speech such as noun, adjective and verb. We developed algorithms for constructing Amharic sentiment lexicons automatically from Amharic news corpus. Corpus based approach is proposed relying on the word co-occurrence distributional embedding including frequency based embedding (i.e. Positive Point-wise Mutual Information PPMI). Using PPMI with threshold value of 100 and 200, we got corpus based Amharic Sentiment lexicons of size 1811 and 3794 respectively by expanding 519 seeds. Finally, the lexicon generated in corpus based approach is evaluated.

keywords: Amharic Sentiment lexicon , Amharic Sentiment Classification , Amharic seed words.

Introduction

For carrying out Amharic sentiment classification, the availability of sentiment lexicons is crucial. To date, there are two generated Amharic sentiment lexicons. These are manually generated lexicon (1000) [Gebremeskel, 2010] and dictionary based Amharic SWN and SOCAL lexicons [Alemneh et al., 2019]. However, dictionary based generated lexicons has short-comings in that it has difficulty in capturing cultural connotation and language specific features of the language. This research builds corpus based algorithm to handle language and culture specific words in the lexicons [Alessia et al., 2015]. However, it could probably be impossible to handle all the words in the language as the corpus is a limited resource in almost all less resourced languages like Amharic. But still it is possible to build sentiment lexicons in particular domain where large amount of Amharic corpus is available. Due to this reason, the lexicon built using this approach is usually used for lexicon based sentiment analysis in the same domain from which it is built. The research questions to be addressed utilizing this approach are: (1) how can we build an approach to generate Amharic sentiment lexicon from corpus? (2) how do we evaluate the validity and quality of the generated lexicon?

Related works

In this section, we briefly present the key related works. Our work is closely associated to the work of [Passaro et al., 2015] which generated emotion based lexicon by bootstrapping corpus using word distributional semantics (i.e. using Positive Point-wise Mutual Information (PPMI)). Our approach is different from [Passaro et al., 2015] in that we generated sentiment lexicon rather than emotion lexicon. The other thing is that the approach of propagating sentiment to expand the seeds is also different. Besides, the threshold selection, the seed words' part of speech are different from language to language. For example, Amharic has few adverb classes unlike Italian [Yimam, 2000EC]. Thus, our seed words do not contain adverbs.

Proposed approach

There are variety of corpus based strategies that include count based (e.g. PPMI) and predictive based (e.g. word embedding) approaches. In this part, we present the proposed count based approach to generate Amharic sentiment lexicon from a corpus. The proposed framework of corpus based approach tries to generate Amharic sentiment lexicon. The framework has four components: (Amharic news) corpus collections, preprocessing module, PPMI matrix of word-context, algorithm to generate (Amharic) sentiment lexicon resulting in the generated (Amharic) sentiment lexicon. The proposed framework is shown in Fig.1.

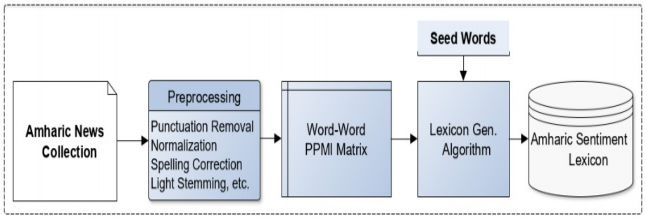


Fig. 1 Proposed corpus based lexicon building framework

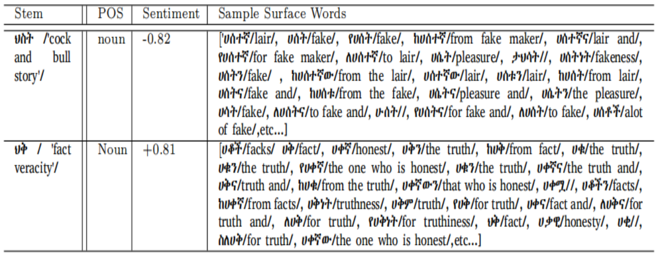
We developed algorithms for constructing Amharic sentiment lexicons automatically from Amharic news corpus. Corpus based approach is proposed relying on the word co-occurrence distributional embedding including frequency based embedding (i.e. PPMI). First we build word-context unigram frequency count matrix and transform it to point-wise mutual Information matrix. For an experimentally chosen threshold value, the top closest words to the mean vector of seed list are added to the lexicon. Then, the mean vector of the new sentiment seed list is updated and process is repeated until we get sufficient terms in the lexicon.

Results and discussions

For evaluating the proposed framework, we used the datasets which consist of 2705 sentence/phrase level sentiment annotated facebook news users' comments collected from the Government Office Affairs Communication (GOAC) between 2008 and 2010. We also used the Amharic sentiment lexicons including manual(1000), SWN(13679) and SOCAL(5683) [Neshir et al. , 2019].

Amharic seed words of size 519 are used to expand PPMI based lexicons. With experimentally obtained threshold value of 100 and 200, we got corpus based Amharic sentiment lexicons of size 1811 and 3794 respectively. Sample of generated lexicon is shown in Table 1.

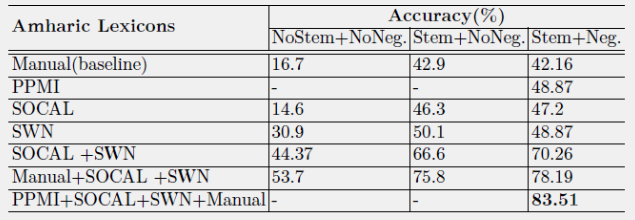
Table 1: sample of generated sentiment lexicon



Stem	POS	Sentiment	Sample Surface Words
ሀሰት /'cock and bull story'/	noun	-0.82	['ሀሰተኛ/lair/, ሀሰት/fake/, የሀሰት/fake/, ከሀሰተኛ/from fake maker/, ሀሰተኛና/lair and/, የሀሰተኛ/for fake maker/, ለሀሰተኛ/to lair/, ሀሴት/pleasure/, ታሀሳት//, ሀሰትነት/fakeness/, ሀሰትን/fake/ , ከሀሰተኛው/from the lair/, ሀሰተኛው/lair/, ሀሰቱን/lair/, ከሀሰት/from lair/, ሀሰትና/fake and/, ከሀሰቱ/from the fake/, ሀሴትና/pleasure and/, ሀሴትን/the pleasure/, ሀሳት/fake/, ለሀሰትና/to fake and/, ሁሰት//, የሀሰትና/for fake and/, ለሀሰት/to fake/, ሀሰቶች/alot of fake/,etc...]
ሀቅ / 'fact veracity'/	Noun	+0.81	[ሀቆች/facks/ ሀቅ/fact/, ሀቀኛ/honest/, ሀቅን/the truth/, ከሀቅ/from fact/, ሀቁ/the truth/, ሀቁን/the truth/, የሀቀኛ/the one who is honest/, ሀቁን/the truth/, ሀቀኛና/the truth and/, ሀቅና/truth and/, ከሀቁ/from the truth/, ሀቀኛውን/that who is honest/, ሀቀሟ//, ሀቆችን/facts/, ከሀቀኛ/from facts/, ሀቅነት/truthness/, ሀቅም/truth/, የሀቅ/for truth/, ሀቀና/fact and/, ለሀቅና/for truth and/, ለሀቅ/for truth/, የሀቅነት/for truthiness/, ሀቅ/fact/, ሀቃዊ/honesty/, ሀቂ//, ስለሀቅ/for truth/, ሀቀኛው/the one who is honest/,etc...]

As discussed on dictionary based lexicons in [Alemneh et al 2019] for lexicon based sentiment classification, using stemming and negation handling are far improving the performance lexicon based classification. Besides, combination of lexicons outperforms better than the individual lexicon. We evaluated the generated Amharic sentiment lexicon in two ways: external to lexicon and internal to lexicon. External to lexicon is to test the usefulness and the correctness of each of the lexicon to find sentiment score of sentiment labeled Amharic comments corpus. Internal evaluation is to compute the degree to which each of the generated lexicons are overlapped (or agreed) with manual, SOCAL and SWN (Amharic) sentiment lexicons. Our lexicon detects subjectivity of Amharic facebook comments has shown an increment of 3.73 more than the subjectivity detection rate of the manual lexicon. For sentiment classification, the performance of our generated lexicon for classifying sentiment of Amharic facebook comments has an increment of 6.71 than the manual sentiment lexicon as shown in Table 2.

Table 2: Evaluation of Corpus based Generated Amharic lexicon for Amharic Facebook Sentiment Classification



Amharic Lexicons	**Accuracy(%)**		
	NoStem+NoNeg.	Stem+NoNeg.	Stem+Neg.
Manual(baseline)	16.7	42.9	42.16
PPMI	-	-	48.87
SOCAL	14.6	46.3	47.2
SWN	30.9	50.1	48.87
SOCAL +SWN	44.37	66.6	70.26
Manual+SOCAL +SWN	53.7	75.8	78.19
PPMI+SOCAL+SWN+Manual	-	-	**83.51**

In addition, the coverage result in a general corpus of 20 million tokens depicts that the coverage of PPMI based Amharic sentiment lexicon is better than the manual lexicon and SOCAL. However, it has less coverage than SWN. Unlike SWN, PPMI based lexicon is generated from corpus. Due to this reason its coverage to work on a general domain is limited. It also demonstrated that the positive and negative count in almost all lexicons seems to have balanced and uniform distribution of sentiment polarity terms in the corpus.

Conclusions and recommendations

This study revealed that it is possible to create sentiment lexicon for low resourced languages from corpus. This captures the language specific features and connotations related to the culture where the language is spoken. This cannot be handled using dictionary based approach that propagates labels from resource rich languages. To the best of our knowledge, the PPMI based approach to generate Amharic sentiment lexicon from corpus is performed for first time for Amharic language with minimal costs and time. Thus, the generated lexicons can be used in combination with other sentiment lexicons to enhance the performance of sentiment classifications in Amharic language. The approach is a generic approach which can be adapted to other resource limited languages to reduce cost of human annotation and the time it takes to annotated sentiment lexicons. Though the PPMI based Amharic sentiment lexicon outperforms the manual lexicon, prediction (word embedding) based approach is recommended to generate sentiment lexicon for Amharic language to handle context sensitive terms. Moreover, there are challenges to be addressed in the future researches including (i)selection of seed words matters, (ii)result is not quite good if either too big or too small threshold, (iii) semantic drift might occur if number of iterations is too big and (iv) the generated lexicon might not work in a general domain.

References

D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3), 2015.

S. Gebremeskel. Sentiment mining model for opinionated amharic texts. Unpublished Masters Thesis and Department of Computer Science and Addis Ababa University and Addis Ababa, 2010.

Girma Neshir Alemneh, Andreas Rauber, and Solomon Atnafu. Dictionary Based Amharic Sentiment Lexicon Generation, pages 311--326. 08 2019.

Lucia Passaro, Laura Pollacci, and Alessandro Lenci. Item: A vector space model to bootstrap an italian emotive lexicon. In Second Italian Conference on Computational Linguistics CLiC-it 2015, pages 215--220. Academia University Press, 2015.

Baye Yimam. (የአማርኛ-ሰዋሰው)yäamarIña säwasäw. Educational Materials Production and Distribution Enterprise(EMPDE), 2000E.C.