Towards Building a SentiWordNet for Tamil

Abishek Kannan
LTRC
IIIT Hyderabad
abishek.kannan
@research.iiit.ac.in

Gaurav Mohanty
LTRC
IIIT Hyderabad
gaurav.mohanty
@research.iiit.ac.in

Radhika Mamidi
LTRC
IIIT Hyderabad
radhika.mamidi
@iiit.ac.in

Abstract

Sentiment analysis is a discipline of Natural Language Processing which deals with analysing the subjectivity of the data. It is an important task with both commercial and academic functionality. Languages like English have several resources which assist in the task of sentiment analysis. SentiWordNet for English is one such important lexical resource that contains subjective polarity for each lexical item. With growing data in native vernacular, there is a need for language-specific SentiWordNet(s). In this paper, we discuss a generic approach followed for the development of a Tamil SentiWordNet using currently available resources in English. For Tamil SentiWordNet, a substantial agreement Fleiss Kappa score of 0.663 was obtained after verification from Tamil annotators. Such a resource would serve as a baseline for future improvements in the task of sentiment analysis specific to Tamil data.

1 Introduction

Tamil has over 70 million native speakers spread across the world and digitized data for Tamil is ever increasing on the web. From news data to movie review sites, usage of Tamil on the web is more than it ever was. In an era driven by social media, the focus shifts from an objective application (News) to a subjective environment (Surveys, Online Review Systems). Applications of sentiment analysis are endless and it is in demand because it has proved to be efficient. Thousands of text documents can be processed for sentiment in seconds compared to the hours it would take for a team to manually complete the task. Commercial Organisations are therefore incorporating sentiment analysis systems[1] for customer feedback and product review. For good governance, feedback from the public through social media and other surveys is monitored at a large scale. The public prefers to give feedback in its own vernacular. Analysing the sentiment in their feedback in various Indian languages hence demands language specific subjective lexicons. This served as the motivation for the creation of SentiWordNet for Tamil.

A translation based approach has been adopted to build this resource using various lexicons in English. Each of these lexicons comprises of English words with certain polarity. After several levels of preprocessing, a final set of English words was obtained. These words were then translated into Tamil using Google Translate[2]. The final set of words were annotated with either positive or negative polarity based on its prior polarity in English. The final lexicon was checked by Tamil annotators to remove any ambiguous entries and also for accuracy of translation.

The various tools used for the construction of SentiWordNet for Tamil include English SentiWordNet (Esuli and Sebastiani, 2006), AFINN-111 lexicon (Nielsen, 2011), Subjectivity Lexicon (Wilson et al., 2005), Opinion Lexicon (Liu et al., 2005) and Google Translate.

The rest of the paper is organized into various sections. Section 2 deals with related work and progress towards building SentiWordNets for Indian languages followed by Section 3 describing the resources and tools used. Section 4 contains a detailed explanation of the approach followed to build the Tamil SentiWordNet. Section 5 defines the evaluation scheme for verification of resource

[1] http://www.sas.com/en$_us/home.html$

[2] https://translate.google.co.in/

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 30–35,
Varanasi, India. December 2016. ©2016 NLP Association of India (NLPAI)

created. An insight on future work and extensibility of the SentiWordNet is provided in Section 6.

2 Related Work

Sentiment analysis has been an age-old task and has been improving steadily over the past few decades. ”It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining, and text mining” (Liu, 2012). Initially the analysis was only restricted to adjectives and adverbs but now many lexical resources contain nouns and verbs also. Various approaches have been proposed for building a SentiWordNet in the past.

Turney worked on sentiment analysis for customer reviews dataset, using an unsupervised learning algorithm (Turney, 2002). Wiebe proposed methods to generate a resource, with subjective information, for a given target language from resources present in English (Wiebe and Riloff, 2005). Translation methods included using a bilingual dictionary and a parallel corpora based approach.

For English, SentiWordNet was developed by Esuli (Esuli and Sebastiani, 2006) with improvements over the years (Baccianella et al., 2010). English SentiWordNet 3.0 is based off the Princeton English WordNet (Miller, 1995). Expansion strategies were suggested to increase the coverage of English SentiWordNet by assigning scores to antonym and synonym synsets.

Less resourced languages depend on resources present in English to build such lexical tools. Whalley and Medagoda propose a method to build a sentiment lexicon for Sinhala using Sentiwordnet 3.0 (Whalley and Medagoda, 2015). The Sentiwordnet is mapped to an online Sinhala dictionary. Scores for each lexicon and its synonyms is assigned based on English Sentiwordnet scores. Similar work is prevalent in literature for many Indian languages as well. Joshi built a SentiWordNet for Hindi using English SentiWordNet and linking English and Hindi WordNets (Joshi et al., 2010). Polarity scores were copied from the words in English SentiWordNet to the corresponding translated words in Hindi SentiWordNet.

Another approach was proposed by Amitava Das (Das and Bandyopadhyay, 2010) (Das and Bandyopadhyay, 2011) (Das and Gambäck, 2012) in order to build SentiWordNet for three Indian languages (Bengali, Hindi and Telugu). This approach used two resources available in English which provided subjectivity information: SentiWordNet 3.0 and Subjectivity Lexicon. A bilingual dictionary based translation was carried out in order to obtain the target lexicon. A Wordnet based approach, to assign scores to synsets, and an automatic corpus based approach were also suggested.

3 Resources Used

For the creation of Tamil SentiWordnet, English SentiWordNet 3.0 and Subjectivity Lexicon were the two most reliable resources. On reviewing English SentiWordNet and comparing it with the Subjectivity Lexicon, it was found that many words had contradicting sentiments in both the lists. Therefore, for a better estimate of sentiment for each word and reduction of ambiguities, two more resources, AFINN-111 and Opinion Lexicon, were also used. The resources used are described below:

List Name	**Number of Tokens**
SentiWordNet	2000K
Subjectivity Lexicon	8222
AFINN-111	2477
Opinion Lexicon	6789

Table 1: Resource Table.

- **English SentiWordNet** is a lexical resource for opinion mining which has a rich dataset of about 2 million lexical entries. SentiWordNet assigns, to each synset of WordNet, sentiment scores: positive and negative. Each synset is uniquely identified by a synset ID corresponding to the synset ID in Princeton WordNet. Other information includes Part-Of-Speech Tag (Adjective, Adverb, Noun, Verb). Positive and negative scores are a decimal ranging from zero to one. Objectivity score defines how factual a given word is and is obtained by 1 - (Positive Score + Negative Score).

- **Subjectivity Lexicon** is also a highly reliable lexicon for sentiment information and is robust in terms of performance. It is used as a part of OpinionFinder[3] (Wilson et al., 2005).

[3] http://mpqa.cs.pitt.edu/opinionfinder/

The list contains for a given word, Part-of-Speech tag, its polarity and subjectivity parameter. Subjectivity parameter classifies the word as either strongly or weakly subjective.

- **AFINN-111** is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). Words have been manually labeled by Finn rup Nielsen (Nielsen, 2011).

- **Opinion Lexicon** comprises of a relatively large dataset of positive and negative words without any specific scores. This dataset is based off annotated twitter corpora (Hu and Liu, 2004).

- **Translation-Dictionary** . In order to translate the final collection of words from English to Tamil, we used Google Translate[4] by running every single word on the Google Translate web application. The final list of translated words was cross checked by Tamil annotators in order to remove multi-word entries, incorrect translations and other ambiguous words.

4 Approach

Figure 1 shows a step by step procedure followed to build the Tamil SentiWordNet. The methodology is generic and can be used to build a SentiWordNet in any language. The entire procedure is divided into three parts :

- **Collecting Source Lexicon** - In order to build a SentiWordNet for any Indian language, one can use available resource(s), with sentiment information, from English.

- **Translation to Target Lexicon** - Once source lexicon is acquired, it needs to be translated to target lexicon using a translation method such as usage of a bilingual dictionary, an online translation resource, or a parallel corpus.

- **Evaluation of Target Lexicon** - The created target lexicon needs to be evaluated for errors. This paper adopts manual evaluation by language specific annotators and reports annotator agreement score.

4.1 Source Lexicon

In order to obtain the source lexicon, multiple filtering techniques were applied to the existing resources in English. Source Lexicon acquisition starts with Subjectivity Lexicon and SentiWordNet, which are the primary resources for sentiment analysis in English. SentiWordNet polarity scores are obtained from learning through large English corpora. A threshold of 0.4 was considered (Das and Bandyopadhyay, 2010), as those words which have a score lower than the threshold may lose subjectivity upon translation to the target language. Words which have scores above 0.4 are assumed to be strongly subjective. Upon filtering words from English SentiWordNet based on the above criteria, a total of 16,791 tokens were obtained.

The Subjectivity Lexicon contains 8,222 words in total. From this set, all words which were annotated as weakly subjective were removed (Riloff et al., 2006). A total of 2,652 weakly subjective words were discarded resulting in a new set of only strongly subjective words. As mentioned before, this list also contains Part-of-Speech tags. Those words which were tagged 'anypos' were also removed to prevent context related ambiguities. Since the main aim was only to capture positive or negative sentiment, words tagged as neutral were also removed. The final list of words from Subjectivity Lexicon comprised of 4,526 tokens.

On merging the two filtered lists it was found that 2,199 tokens were common between the both. Among these duplicates only words which had the same Part-of-Speech tag in both the lists were sent forward and the others were discarded. Some of the duplicates included words which had conflicting tags in the SentiWordNet and the Subjectivity Lexicon. For example, the word 'pride' was tagged as positive in the Subjectivity Lexicon and the same word was given a higher negative score in SentiWordNet. Subjectivity of such words depend upon context and hence were also removed.

The final list now contained words which were strongly subjective and would more likely hold their subjectivity after translation. To ensure this, the list was manually checked. The final list now contained 15,823 tokens.

Since many entries in the SentiWordNet had opposing scores to that in the Subjectivity Lexicon, it was decided to add two more lists to increase the reliability of the source lexicon. AFINN-

[4] https://translate.google.co.in/

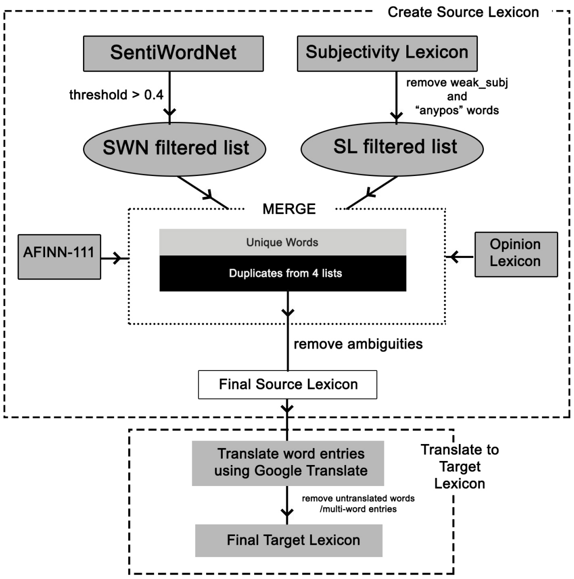

Figure 1: Flow of Design for Tamil SentiWordNet

111 (Nielsen, 2011) and Opinion Lexicon (Liu et al., 2005) were the two lists that were added and they were used to filter out more words which had ambiguous sense. For example, the word 'rid' had a positive score in SentiWordNet and it was tagged as negative in the Subjectivity Lexicon. Such words had to be avoided because they could be either positive or negative, depending on the context. A total of 4,954 words were present in more than one list. If a given word was present in a majority of the lists it appeared in, the majority opinion (positive or negative) was considered. The final list only contained strongly subjective words which served as the source lexicon before translation.

4.2 Target Lexicon

Each of the words from the final list was then translated using Google translate. Bilingual dictionaries may not account for all the words because of language variations (Bakliwal et al., 2012). This method of translation is also labour intensive. Context dependent word mapping between two languages is a tough task in general. Though Google Translate has its own challenges, this method was used for faster translation of words and better translation performance.

Some words were not translated into Tamil because the target language lacks such words. Multi-word entries in the source lexicon were challenging to translate as sometimes the first word would get translated but not the rest. In a few cases a multi-word entry would get translated to an accurate single word and in some cases a single word entry would get translated to a multi-word entry in the target language. Such cases had to be individually checked during pre-processing before evaluation.

The final list of words which were properly translated contained 10,225 single word entries tagged as positive or negative. One must note that this method does not copy English SentiWordNet scores for positivity or negativity. Copying scores was suggested previously in literature (Das and Bandyopadhyay, 2010). This was not followed because English SentiWordNet scores are based on English corpus. Same scores may not necessarily work for Tamil. Subjectivity of a word may transcend across languages but not in the same magnitude. Hence, a given word is only marked as carrying either positive or negative sentiment.

5 Evaluation Methodology

After translation to target lexicon, the list comprised of a set of English words along with their corresponding Tamil translations. Each of these words is either marked with positive or negative polarity based on its polarity in source lexicon(s).

This list was sent to 5 Tamil annotators to verify the correctness of the translation. Words which did not retain subjectivity after translation were removed. In case of conflict over any word, the majority of opinion was taken into account. When we say majority, we assume that at-least 4 out of 5 annotators agree on a given sentiment. If not, the word is removed from the list for future inspection. Words which did not transfer contextual meaning were also removed. For example, the word *inclination* was translated wrongly in the target language. The translation only captured the words meaning as a slope or an incline and not a person's tendency to act or feel in a certain way. Words which were wrongly translated were also removed.

The evaluation resulted in 190 words being marked as ambiguous and 540 words being marked as wrongly translated. These words were eliminated and the final lexicon contained a total of 9495 words with strong subjectivity. 3336 words were tagged as positive and 6159 words were tagged as negative. In order to capture inter-annotator agreement Fleiss Kappa[5] score for the final set was also calculated. Fleiss Kappa is calculated using the following formula:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

$\bar{P}$ is the sum of observed agreement. $\bar{P}_e$ is the sum of agreement by chance. Fleiss kappa score is calculated using five raters.There are two categories (positive/negative). A substantial agreement score of κ = **0.663** is reported for Tamil SentiWordNet.

Initial Token Count	10225
Wrongly Translated	540
Ambiguous Entries	190
Final Token Count	9495
Inter-Annotator Agreement (**Fleiss Kappa**)	
0.663	

Table 2: Evaluation Details

6 Conclusion and Future Work

The Tamil SentiWordNet lexicon serves as a baseline for future improvements. The approach followed can be used to build SentiWordNet for any Indian Language. Other methods of translation to target language include usage of a bilingual dictionary or a parallel corpora for English and Target language pair. Various techniques can be applied to improve the accuracy and expand the lexicon content. Tamil WordNet (Rajendran et al., 2002) is available publicly[6] and contains 1916 synset entries. Lexicon can be expanded by using, for a given word in the Tamil SentiWordNet, its corresponding synsets in the Tamil WordNet. Synonyms and antonyms can be classified with similar and opposite subjectivity respectively.

The SentiWordNet can also be expanded using a corpus based approach to capture language-specific words. SentiWordNet lexicon can be used as a seed list and the corpus can be tagged based on this seed list. Machine learning techniques can then be applied on this corpus to find new words to be added to the lexicon.

Currently, the lexicon has only been divided into two classes (positive and negative) This classification can be replaced by a five point scale in the future (Nakov et al., 2016). Furthermore, for getting subjectivity scores of individual words in the SentiWordNet one can use sentiment annotated Tamil corpora. The accuracy of any lexical resource is best calculated when it is actually usable in practical applications. With Tamil social media data being more readily available, manually annotating this data for positive and negative sentiment and using Tamil SentiWordNet to annotate the same data to check for accuracy is one possible method. One of the challenges which needs to be addressed in the future is capturing the sentiment of multi-word entries.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*.

Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for indian languages. *Asian Federation for Natural Language Processing, China*, pages 56–63.

Amitava Das and Sivaji Bandyopadhyay. 2011. Dr sentiment knows everything! In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: systems demonstrations*, pages 50–55. Association for Computational Linguistics.

[5] https://en.wikipedia.org/wiki/Fleiss’˙kappa

[6] https://www.amrita.edu/center/computational-engineering-and-networking/research/computational

Amitava Das and Björn Gambäck. 2012. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 38–46. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

S Rajendran, S Arulmozi, B Kumara Shanmugam, S Baskaran, and S Thiagarajan. 2002. Tamil wordnet. In *Proceedings of the First International Global WordNet Conference. Mysore*, volume 152, pages 271–274.

Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 440–448. Association for Computational Linguistics.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

J Whalley and N Medagoda. 2015. Sentiment lexicon construction using sentiwordnet 3.0. *ICNC'15 - FSKD'15, School of Information Science and Engineering, Hunan University, China*.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.