No more beating about the bush: A Step towards Idiom Handling for Indian Language NLP

by

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, Dipti Misra Sharma

in

11th edition of the Language Resources and Evaluation Conference
(*LREC-2018*)

Miyazaki (Japan)

Report No: IIIT/TR/2018/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2018

No more beating about the bush: A Step towards Idiom Handling for Indian Language NLP

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, Dipti Sharma
FBK Trento, IIIT Hyderabad, Cardiff University, IIIT Hyderabad
ragrawal@fbk.eu, vighnesh.ck@students.iiit.ac.in, muralidaranV@cardiff.ac.uk, dipti@iiit.ac.in

Abstract

One of the major challenges in the field of Natural Language Processing (NLP) is the handling of idioms; seemingly ordinary phrases which could be further conjugated or even spread across the sentence to fit the context. Since idioms are a part of natural language, the ability to tackle them brings us closer to creating efficient NLP tools. This paper presents a multilingual parallel idiom dataset for seven Indian languages in addition to English and demonstrates its usefulness for two NLP applications - Machine Translation and Sentiment Analysis. We observe significant improvement for both the subtasks over baseline models trained without employing the idiom dataset.

Keywords: Idioms, Machine Translation, Sentiment Analysis, Indian Languages

1. Introduction

Idioms pose a problem to most NLP applications (Sag et al., 2002), including sentiment analysis, question answering, machine translation, parsing and so on. One of the most negatively affected subtasks among these is Machine Translation (MT) (Salton et al., 2014a). While parallel corpora can be used by MT systems to learn the language constructs, thereby generating decent translations from source to target language; the same cannot be said for the learning of idioms. Most machine translation systems existent today fail when it comes to the handling of idioms (Table 1). Past research (Salton et al., 2014a) has come up with results stating that a standard Statistical Machine Translation (SMT) system tends to achieve only about half the BLEU score of the same system when applied to sentences containing idioms, as compared to those that do not.

Since idioms encode a very specific kind of linguistic knowledge, it is not easy to learn their automatic handling computationally, without an idiom database. This makes idiom handling a challenging problem for various NLP subtasks including sentiment analysis and question answering in addition to MT. The situation is far worse for Indian languages, a majority of which are low resource languages (Post et al., 2012) with regard to the availability of NLP tools, and yet representing 1.3 billion native speakers. Moreover, these languages are under-studied, while also exhibiting linguistic properties that make idiom handling for various NLP subtasks even more challenging.

In this paper, we present a multilingual parallel dataset that maps 2208 commonly used idioms in English to their translations in seven Indian languages: Hindi, Urdu, Bengali, Tamil, Gujarati, Malayalam and Telugu[1]. The idioms are also annotated with the appropriate sentiments that they channel, and their meanings in the respective languages. We demonstrate the enhancement obtained using our resource for two major applications - machine translation and sentiment analysis. We observe a significant improvement in performance on conducting baseline experiments for the above mentioned tasks.

2. Related Work

One of the earliest known work in idiom handling is a comparative study (Volk, 1998) between two contemporary translation systems, namely machine translation and translation memory systems. The study concluded that neither of the systems could handle idioms, and proposed a method of integrating both the systems along with idiom databases to form a phrase archive, which could then be recognised more efficiently by the translation systems. A popular idiom corpus was the one built for Japanese (Hashimoto and Kawahara, 2008). This resource contains Japanese phrases labelled as either idiomatic or literal, which helps to better understand the semantics of the sentence. Another well known work on idiom handling is a system to identify idiomatic expressions from a large bilingual English-Korean corpus, using phrasal alignments to make sense of phrases as well as words, instead of the previously explored word alignment method that purely made sense of words alone (Lee et al., 2010). (Post et al., 2012) crowdsourced a parallel corpus between English and six Indian languages namely: Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. They compared the translational capabilities of their model with regard to Google Translate. However, there was no research specific to the domain of idioms in this work. Meanwhile, efforts to efficiently translate idioms resulted in a system that implemented a substitution method (Salton et al., 2014b). This system was tested on a parallel corpus of English and Brazilian-Portuguese, and would first substitute idioms with their literal meanings before translation, and later on substitute these literal meanings back to idioms after. There has been considerable work done on multiword expressions (MWE) in the last two years with regard

[1] This dataset is available at goo.gl/receLs

Table 1: Performance on Google Translate on idiomatic sentences.

Source	John is known for beating about the bush.
Target	जॉन बुश के बारे में मारने के लिए जाना जाता है।
Transliteration	john bush ke baare mein maarane ke lie jaana jaata hai.
Gloss	John bush GEN about LOC beat-INF PUR know-PASS-MASC-PRES
Meaning	John is known to be hitting in the matter of bush.
Source	The show kept me in stitches the entire time.
Target	शो ने मुझे पूरे समय टाँके में रखा।
Transliteration	sho ne mujhe poore samay taanke mein rakha.
Gloss	Show ERG I-DAT total time stitch LOC keep-PST
Meaning	The show kept me in stiches (injured) the entire time.

Legend: *GEN - Genitive case, LOC - Locative case, DAT - Dative case, PASS - Passive voice, MASC - Masculine gender, PRES - Present tense, PST - Past tense, INF - Infinitive form of a verb, PUR - Purpose of an action*

Category	Number of Idioms
Very negative (- -)	196
Somewhat negative (-)	657
Neutral (0)	726
Somewhat positive (+)	503
Very positive (++)	126
Total	2208

Table 2: Sentiment Annotation Statistics of our Dataset

to Indian languages. A prominent work was the detection of MWEs for Hindi language, mainly noun compounds and noun+verb compounds, using Word Embeddings and WordNet-based features (Patel and Bhattacharyya, 2015). Another important work was the annotation of MWEs for Hindi and Marathi, and classifying them into either compound nouns or light verb constructions (Singh et al., 2016). A very recent work on the topic of idiom handling (Liu and Hwa, 2016) is a system that implements a phrasal substitution by replacing idioms with their corresponding meanings and transforming the meanings to fit the context of the sentence with the right conjugation. So far, there has been no significant work done on creation of a multilingual idiom dataset. To the best of our knowledge, our resource is the first of its kind for Indian languages.

3. Creation of IMIL

3.1. Data Collection

A significant number of idioms in reference materials are ones that are seldom used, thereby hindering their effectiveness (Liu, 2003). We strive to create a multilingual parallel idiom dataset that covers the most commonly used idioms in everyday English, so that it can be used effectively for different NLP applications. We crawled the web through relevant websites to extract over 5000 idioms, their respective meanings, and their sample usages[2]. The list of the websites crawled through is provided here [3]. We then perform an intersection of the list of idioms obtained with those compiled from other well known American English corpora, including the American National Corpus (ANC) (Ide and Suderman, 2004); Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al., 2002), and Brown Corpus (Francis and Kucera, 1979).

The compilation of idioms from the above mentioned corpora was done using the method proposed by (Muzny and Zettlemoyer, 2013). A threshold count of 25 was set for eliminating non-frequent idioms after performing the intersection. The intersection and elimination was done for the removal of inaccurate programmatic detections but at the same time ensuring that it is a frequently occurring idiom. We were then able to filter out and consolidate a list of 2208 most commonly used idioms, thus rendering the application of the corpus as close to natural human language as possible.

3.2. Annotation Guidelines

We create a parallel idiom dataset for seven Indian languages in addition to English. The English idioms extracted in the first phase (Section 3.1.) are translated to the following languages :Hindi, Urdu, Malayalam, Bengali, Gujarati, Tamil and Telugu. The annotation for each language is performed by three native speakers and later verified by two professional linguists to deal with language specific idiosyncrasies. The resulting dataset is called “Idiom Mapping for Indian Languages” (IMIL).

Following are the guidelines that the annotators were asked to follow:

1. An idiom, its meaning, and a sample usage is provided in English followed by slots for the seven languages. If there is an equivalent idiom in the target language, then the corresponding idiomatic translation is to be provided. In case this is not possible, a phrasal translation that aptly conveys the meaning of the source idiom is to be added instead. This information is to be mentioned along with the translation, using the tags ‘P’ (phrasal) or ‘I’ (idiomatic).

[2] This was done using various python libraries, primarily BeautifulSoup4.

[3] https://goo.gl/s4R4uH

2. In case it is neither possible to find an appropriate idiom nor an equivalent phrasal translation, “Skip” has to be entered in the target slot.

3. The sentiment of each idiom should be marked at each node in its parsed tree structure (detailed in Section 4.2.). The annotation scheme along with the statistics for each sentiment is given in Table 2.

4. Experiments and Results

We demonstrate the application of our dataset by conducting experiments for two different NLP tasks:

4.1. Machine Translation

Statistical Machine Translation (SMT) (Koehn, 2009) and Neural Machine Translation (NMT) ((Sutskever et al., 2014), (Cho et al., 2014), (Bahdanau et al., 2014)) are the two major MT paradigms today which require large parallel corpora for training. Such corpora containing sufficient idiomatic sentences are not available for Indian languages. We employ $IMIL$ and conduct experiments to analyse MT quality when the system is fed with an idiom mapping in addition to the parallel training corpora.

We employ the multilingual Indian Language Corpora Initiative (ILCI) corpus (Jha, 2010) for training [4]. It contains 50,000 sentences from the health and tourism domains aligned across eleven Indian languages. We choose three Indo-Aryan languages (Hindi, Bengali, Urdu) and one Dravidian language (Telugu) as candidate languages for our experiments to maintain brevity. We employ preprocessing to eliminate misalignments - the resultant dataset has a size of 47,382 sentences (Training - 44000, Validation - 1382, Test - 2000). We create 250 manually annotated sentences with idiomatic usages, of which 50 are appended to the validation set, and 200 to the test set. The resultant size of the training set, validation set and test set ($Test_{Concat}$) is 46,200, 1432 and 2200 sentences respectively. We conduct experiments using both NMT as well as SMT approaches. For the former, we concatenate $IMIL$ to the $ILCI$ training set, yielding a training set containing 46200 sentences. We train an NMT model on this set using the architecture proposed by (Bahdanau et al., 2014) and call it NMT_{IMIL}. We compare the performance of this model with a baseline NMT model trained on the $ILCI$ train set (44000 sentences). This model is called NMT_{Base}. The results obtained are given in Table 3. Although NMT_{IMIL} produces better output than NMT_{Base} in terms of BLEU score, the translation quality obtained is found to be substandard on manual inspection due to inadequate inflectional learning. It is, however, much better as compared to the literal translation produced by NMT_{Base} for idioms.

Additionally, we train a Phrase Based Statistical Machine Translation system (PBSMT) (Zens et al., 2002) using our dataset as an additional resource for the phrase table generation. We use Moses (Koehn et al., 2007) for phrase extraction as well as lexicalized reordering as proposed by (Kunchukuttan et al., 2014). We append the 2208 idioms to the phrase-tables rather than concatenating them to the training set[5]. The training set is thus 44000 sentences with the other splits remaining the same as mentioned above. We also add an additional feature in the phrase table to indicate whether the idiom can have a non-idiomatic usage as well, i.e. 0 if it cannot and 1 if it can[6]. We compare the performance with a standard PBSMT model trained on the $ILCI$ parallel corpus, called $PBSMT_{Base}$. We observe significantly higher improvement in scores (an average inrease of 2.69 % BLEU) using this method than that obtained using NMT (0.73 % BLEU). This can be attributed to a more sophisticated handling of idioms using phrase tables rather than direct concatenation to the training corpus.

4.1.1. Discussion

Although the performance of the translation system improves with the inclusion of our idiom dataset (IMIL), there are a few issues that we noted in the idiomatic translations produced by the system. An idiom is not a fixed multi-word expression but allows considerable variation in how the idiomatic expression is going to be realized in a sentence based on syntactic and morphological properties of (a) the tokens inside the idiom (b) the composite expression itself. If the equivalent idiom in the target language belongs to a similar syntactic category the translation is likely to be correct. Where the syntactic categories of the expressions differ, the translation quality is affected.

Let us take two idioms for illustration: ‘without batting an eyelid’ and ‘cannot stomach someone or something’. The syntactic category of the two idioms are PP and NP respectively which determines how they are used in a sentence. The first idiom can be used as a part of a Verb Phrase like ’VP(VP (uttering a lie) PP(without batting an eyelid)). If the equivalent idiom in target language can be used as a part of a verb phrase just like English and hence the translation sounds good. For example, when the system output for Telugu translation is “(VP (VP(saṅkōcapaḍakuṇḍā) VP(abad’dhaṁ annāḍu))”, the translation is perfectly okay. Even though the target phrase “(VP saṅkōcapaḍakuṇḍā)” is not a PP like in English, it can constitute a larger verb phrase just like in English. In the second idiom, the target language phrase learnt from our parallel dataset is not a verb phrase but a noun phrase ‘bardaasht ke baahar’. Hence a source sentence ‘He could not stomach the truth’ when translated as ‘wah sach bardaasht ke baahar hai’ is not a good translation because of this syntactic incompatibility of the target

[4] This corpus is available on request from TDIL: https://goo.gl/VHYST

[5] We use the xml markup feature provided by Moses for suggesting phrasal translations to the decoder.

[6] This facilitates learning of the decoder for idioms having possible literal usage as well.

Table 3: Impact of IMIL on Sentiment Analysis

S1	John is always beating about the bush
Base	(2 (2 John) (3 (3 (2 (2 is) (2 always)) (2 (2 beating) (2 (2 about) (2 (2 the) (2 bush))))) (2 .)))
IMIL	(2 (2 John) (1 (1 (2 (2 is) (2 always)) (1 (2 beating) (2 (2 about) (2 (2 the) (2 bush))))) (2 .)))
S2	He hit the ceiling when he came to know the truth.
Base	(3 (2 He) (2 (2 (2 (3 hit) (2 (2 the) (2 ceiling))) (2 (2 when) (2 (2 he) (2 (2 came) (2 (2 to) (2 (2 know) (2 (2 the) (3 truth))
IMIL	(1 (2 He) (1 (1 (0 (3 hit) (2 (2 the) (2 ceiling))) (2 (2 when) (2 (2 he) (2 (2 came) (2 (2 to) (2 (2 know) (2 (2 the) (3 truth))
S3	Mary is in the pink of health.
Base	(2 (2 (2 Mary) (2 (2 is) (2 (2 in) (2 (2 (2 the) (2 pink)) (2 (2 of) (2 health))))) (2 .)))
IMIL	(3 (2 (2 Mary) (3 (2 is) (3 (2 in) (4 (2 (2 the) (2 pink)) (2 (2 of) (2 health))))) (2 .)))

Legend: *S1, S2, S3 : Sample Sentences, 0: Extremely negative, 1: Negative, 2: Neutral, 3: Positive, 4: Extremely positive, Base:* $Stanford_{Base}$*, IMIL:* $Stanford_{IMIL}$

Model	Direction	Bengali	Urdu	Telugu
$PBSMT_{Base}$	hin=>	28.42	41.38	14.82
	hin<=	28.17	42.64	19.47
NMT_{Base}	hin=>	25.73	39.57	11.43
	hin<=	26.42	43.51	15.14
$PBSMT_{IMIL}$	hin=>	**31.84**	**44.74**	**16.46**
	hin<=	**32.06**	43.97	**21.94**
NMT_{IMIL}	hin=>	25.91	40.18	12.18
	hin<=	27.81	**44.15**	15.97

Table 4: Results obtained on $Test_{concat}$ by our models in terms of BLEU score. *hin: Hindi*

	Compositional	Actual
break a leg	-	+
kick the bucket	0	-
apple of my eye	0	++
under the weather	0	-

Table 5: Examples of non-compositionality of sentiments in idioms

idiom. Idioms have to be matched for their syntactic compatibility while translating them. Secondly, there are components in idioms which are determined by the other tokens outside the idiom. e.g. 'worth one's salt' is realized as 'worth his salt', 'worth her salt' and so on agreeing with the subject. These changes should be accommodated in the target language as well.

As part of future work, the automatic generation of the bidirectional lexical and phrasal translation probabilities as proposed by (Klementiev et al., 2012) can be explored along with the feature addition in the phrase table for further improvement in performance for languages where large monolingual corpora are available. This could facilitate the coverage of words and phrases surrounding the idiom by the the decoder in addition to the idiom itself.

4.2. Sentiment Analysis

This is one of the most interesting applications of the database due the non-compositional behavior of idioms in terms of semantic as well as sentiment information. Table 4 gives some of such examples, motivating the need for a sentiment-annotated idiom database.

Model	CALA	CARLA
$Stanford_{Base}$	67.01	70.23
$Stanford_{IMIL}$	68.73	73.56

Table 6: Sentiment Analysis results on $Test_{Concat}$. *CALA: Combined Approximate Label Accuracy. CARLA: Combined Approximate Root Label Accuracy.*

This is the primary motivating factor for the need of an idiom sentiment database like $IMIL$, which can help towards better Sentiment Analysis, especially the phrase-level approaches ((Wilson et al., 2005), (Socher et al., 2013)). $IMIL$ can be employed for Sentiment Analysis for any candidate language among the languages in consideration. Due to space constraints, we demonstrate the application of $IMIL$ to Sentiment Analysis for English, using Recursive Neural Tensor Networks (RNTNs) proposed by (Socher et al., 2013). The RNTNs can learn the phrase sentiments from a sentiment treebank containing trees with a sentiment annotated at each node in the parsed tree structure of a sentence. Our dataset is seamlessly integratable with the Stanford Sentiment Treebank, since the sentiment annotation scheme (mentioned in Section 3.2.) is in alignment with the method employed by (Socher et al., 2013).

We generate the parse trees for the 2208 idioms from $IMIL$ and annotate them with sentiments at each node level. We append this treebank to the training set employed by (Socher et al., 2013). The model trained on this set is called $Stanford_{IMIL}$. For development and testing, we append 50 and 200 annotated sentence trees with idiomatic usage to the original validation and test sets respectively. The resultant statistics are as follows: Training set - 10744 trees, validation set - 1151 trees, test set - 2410 trees. The test set is called $Test_{Concat}$. We compare the performance with a baseline model trained on the original training set employed by (Socher et al., 2013). We call this model $Stanford_{Base}$. The results of both the models on $Test_{Concat}$ are given in Table 5. $Stanford_{IMIL}$ shows significant improvement over $Stanford_{Base}$. We observe that although the training is done on only idiomatic phrase trees than sentence trees, $Stanford_{IMIL}$ produces a 1.52 % increase for Combined Approximate Label Accuracy and a 3.33 % increase for Combined Approximate Root Label Accuracy. This is attributed to the ability of the RNTN to learn the sentiment at higher nodes of the tree from the subtrees using the tensor-based composition function.

4.2.1. Discussion

We inspect the outputs generated by the model after it is trained using our parallel dataset IMIL[7]. We plot the sentiment trees using the outputs generated by StanfordBase and StanfordIMIL. Table 3 shows the performance of the model for idiomatic sentences, before and after training with IMIL. [8]. It can be observed that the model is able to handle the non-compositional behavior of idioms with respect to sentiments on being trained with IMIL as additional data. This would be very challenging to accomplish in the absence of labelled sentiment trees for idioms. StanfordIMIL is also able to learn the correct sentiment trees for the entire idiomatic sentences, although the training is done only on the idiom phrases.

5. Conclusion

This paper is an effort in the direction of idiom handling for various Natural Language Processing tasks, with an emphasis on Indian languages. We present $IMIL$, a multilingual parallel idiom dataset consisting of 2208 idioms, spanning across seven languages in addition to English. We demonstrate its usefulness for two applications, namely Machine Translation and Sentiment Analysis. We conclude that Phrase-based SMT is better able to handle idiomatic sentences than Neural Machine Translation, producing an average increase of 2.69 % BLEU score over a baseline model trained over the same corpus. A promising improvement is also observed for Sentiment Analysis, primarily due to the inability of the baseline model to learn the non-compositional sentiments of idioms, which is addressed with the presence of an idiom sentiment dataset. We conclude that $IMIL$ is a valuable resource with potential applications in varied NLP subtasks, especially with regard to Indian languages.

6. References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bhattacharyya, P. (2017). Indowordnet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Fellbaum, C. (1998). Towards a representation of idioms in wordnet. In *Proceedings of the workshop on the use of WordNet in Natural Language Processing Systems (Coling-ACL 1998)*.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Brown University*, 2.

Hashimoto, C. and Kawahara, D. (2008). Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features. In *Proceedings of the conference on empirical methods in natural language processing*, pages 992–1001. Association for Computational Linguistics.

Ide, N. and Suderman, K. (2004). The american national corpus first release. In *LREC*.

Jha, G. N. (2010). The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R., and Bhattacharyya, P. (2014). Sata-anuvadak: Tackling multiway translation of indian languages. *pan*, 841(54,570):4–135.

Lee, H.-G., Kim, M.-J., Hong, G., Kim, S.-B., Hwang, Y.-S., and Rim, H.-C. (2010). Identifying idiomatic expressions using phrase alignments in bilingual parallel corpus. In *PRICAI*, pages 123–133. Springer.

Liu, C. and Hwa, R. (2016). Phrasal substitution of idiomatic expressions. In *HLT-NAACL*, pages 363–373.

Liu, D. (2003). The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, 37(4):671–700.

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

Muzny, G. and Zettlemoyer, L. (2013). Automatic idiom identification in wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421.

Osherson, A. and Fellbaum, C. (2010). The representation of idioms in wordnet. In *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010), Mumbai, India. Narosa Publishing House.*

[7] We choose English due to smooth integrability with Stanford SA

[8] A graphical view is provided at https://goo.gl/5Wcqba

Patel, D. S. S. B. K. and Bhattacharyya, P. (2015). Detection of multiword expressions for hindi language using word embeddings and wordnet-based features. In *12th International Conference on Natural Language Processing*, page 291.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. *Computational Linguistics and Intelligent Text Processing*, pages 189–206.

Salton, G., Ross, R., and Kelleher, J. (2014a). An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese.

Salton, G., Ross, R., and Kelleher, J. (2014b). Evaluation of a substitution method for idiom transformation in statistical machine translation.

Simpson, R. C., Briggs, S. L., Ovens, J., and Swales, J. M. (2002). The michigan corpus of academic spoken english. *Ann Arbor, MI: The Regents of the University of Michigan.*

Singh, D., Bhingardive, S., and Bhattacharya, P. (2016). Multiword expressions dataset for indian languages. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Volk, M. (1998). The automatic translation of idioms. machine translation vs. translation memory systems. *Machine Translation: Theory, Applications, and Evaluation, An Assessment of the State-of-the-art, St. Augustin, Gardez Verlag.*

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Zens, R., Och, F. J., Ney, H., and Vi, L. (2002). Phrase-based statistical machine translation. *KI*, 2479:18–32.