Classifying Frames at the Sentence Level in News Articles

Nona Naderi
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
nona@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
gh@cs.toronto.edu

Abstract

Previous approaches to generic frame classification analyze frames at the document level. Here, we propose a supervised based approach based on deep neural networks and distributional representations for classifying frames at the sentence level in news articles. We conduct our experiments on the publicly available Media Frames Corpus compiled from the U.S. Newspapers. Using (B)LSTMs and GRU networks to represent the meaning of frames, we demonstrate that our approach yields at least 14-point improvement over several baseline methods.

1 Introduction

Framing is generally conceptualized as a communication process to present an object or an issue. Various typologies are proposed for framing, for example some associate frames with specific issues (Entman (1993); Chong and Druckman (2007)), while others such as Card et al. (2015) believe that framing should be perceived as non-issue-specific, and be analyzed with a fixed set of framing dimensions. We take de Vreese's view that frames can be classified as either generic (issue-independent) or issue-specific (de Vreese (2005)). For example, *economic benefits* can be used as a generic frame for various issues; but the frame *marriage is about more than procreation* is specific to the issue of gay marriage. A single frame may relate to a complete text or only to shorter elements of a text, such as a paragraph or a single clause or sentence. Here, we focus on identifying generic frames at the sentence level in the Media Frames Corpus (Card et al., 2015). We employ both uni- and bi-directional LSTMs and gated recurrent networks, which have been used effectively to represent long sequences, to automatically learn frame representations.

In Section 2, we summerize the previous work on computational analysis of framing. The problem definition is introduced in Section 3, followed by the data preparation (Section 4). We describe the experimental setup in Section 5. We then report the results in Section 6, and finally conclude in Section 7.

2 Related Work

Researchers have taken different approaches to operationalize the concept of framing. Some work used various kinds of topic models to analyze frames. Tsur et al. (2015) interpreted various contexts of a specific topic as frames, and employed topic models and time series to infer them. In a similar study, Nguyen et al. (2015) modeled issues and frame topics using hierarchical topic models. They used bill texts, votes, and floor speeches of the U.S. Congress for their predictions. Baumer et al. (2015) investigated various lexical and syntactic features to characterize framing language in political news stories. They found that imagery, figurativeness, and other lexical features are important in identifying framing language.

The prior work on the analysis of issue-specific frames mostly focused on a limited list of issues and frames. Boltužić and Šnajder (2014) addressed the task of tagging user postings with a pre-existing set of frames for the two topics of *Pledge of Allegiance* and *gay marriage*. Their supervised classification model made use of entailment and semantic similarity features. To generalize their earlier work for various topics, they subsequently presented an unsupervised model to recognize frames on the topics of *abortion, gay rights, Obama,* and *marijuana* by means of textual similarity (Boltužić and Šnajder, 2015). On

Proceedings of Recent Advances in Natural Language Processing, pages 536–542,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_070

the same dataset, Hasan and Ng (2014) employed a probabilistic approach to classify forum posts based on users' stance and reasons. In a similar task, Misra et al. (2015) used a set of lexical and semantic similarity features to classify online forum discussions by "argument facets". Naderi and Hirst (2016) analyzed frames across genres and extracted various frames specific to the *gay-marriage* issue from Canadian parliamentary proceedings. Card et al. (2016) explored the use of persona features to classify entire news articles (on the issue of *immigration*) by their overall frames.

Various frames are used in news articles to persuade the audience by establishing a point of view or supporting one. Frame detection at the sentence level helps in analyzing these persuasive strategies in more detail. Here, we investigate the use of recurrent neural networks for identifying generic frames at the sentence level.

3 Problem Definition

Given a text about a controversial issue, our goal is to classify each sentence that expresses a frame relating to the issue (and not just the entire text with a single frame, as Card et al. (2016) did). We use articles from the Media Frames Corpus (see section 4 below), and our objective in this work is to identify the generic frames expressed in the sentences of these texts.

The following example, an excerpt from an article in the Media Frames Corpus (Card et al., 2015), is annotated with the primary frame *Quality of life* as the overall frame of the article. Individual sentences are annotated with frames (shown in boldface) such as *Fairness and equality* and *Cultural identity*. The annotations do not always cover the entire sentence, for example, only the first part of the third sentence is annotated, and the second part is not. Additionally, in some cases, portions of texts are annotated with multiple frames.

Example 1 *Immigration1.0-171*
[Overall frame of the article: ***Quality of life****]*
Immigrants say bias is 'swift kick' to citizenship ***[Fairness and equality]***
When Eduardo Flores moved to Texas in 1981, he was content straddling two cultures: working in the United States but retaining his Mexican citizenship ***[Cultural identity]****. Now, the anti-immigrant sentiment spawned by California's Proposition 187 is making him have second thoughts* ***[Cultural identity]****: Flores wants a claim on the rights available in his adopted land. Legal immigrants like Flores throughout the Southwest have been applying for citizenship at record levels, and many say they want the right to vote to stop the spread of laws like Proposition 187.* ***[Legality, constitutionality and jurisprudence]***

	Frame	N I+S	N I
1	Economic	7,070	2,597
2	Capacity and resources	1,516	846
3	Morality	1,185	259
4	Fairness and equality	1,368	559
5	Legality, constitutionality and jurisprudence	9,420	4,233
6	Policy prescription and evaluation	6,505	2,716
7	Crime and punishment	6,206	3,857
8	Security and defense	1,730	1,171
9	Health and safety	4,968	1,054
10	Quality of life	3,790	1,674
11	Cultural identity	4,644	2,264
12	Public opinion	2,496	937
13	Political	7,864	4,253
14	External regulation and reputation	888	438
15	Other	623	278
16	*Irrelevant*	1,256	–

Table 1: Frames and number of sentences for each (N), extracted from the Media Frames Corpus. I+S includes frames on immigration and smoking; I includes frames on only immigration

4 Data and Pre-processing

The Media Frames Corpus (Card et al., 2015) consists of news articles on three topics of *immigration, smoking,* and *same-sex* marriage.[1] In this corpus, each document is annotated with overall frame (this is what Card et al. (2016) used), and in each sentence, any text that cues a frame is also annotated with that frame, as seen in Example 1 above.

To create our dataset, we first gathered the an-

[1]We were able to download 4,315 articles from *smoking,* and 5,686 articles from *immigration* using the scripts provided at https://github.com/dallascard/media_frames_corpus. However, we were not able to obtain any of the *same-sex marriage* articles (according to the authors the inter-annotator agreement on the *same-sex marriage* set was much lower than the other two sets, Krippendorff alpha 0.08 compared to 0.16 for immigration and 0.23 for smoking).

notations that at least two annotators agreed upon; however, that process resulted in a small corpus because a majority of the articles on *smoking* were annotated only once. Therefore, we kept the cases that were annotated only once, and for the more controversial cases, where multiple frame dimensions were assigned, we kept only the annotations that were agreed upon by at least two annotators.

We then pre-processed the articles with a sentence splitter,[2] and gathered all the sentences annotated with the cue words for each frame. This resulted in 61,529 sentences in total. Table 1 shows the statistics of the resulting dataset.

The sentences were further lower-cased and all numeric tokens were converted to ⟨NUM⟩. Since frames 1, 5, 6, 7, and 13 account for more than 60% of the data, we focused on identifying these five frames; however, we also report the results based on all 15 frames, plus irrelevant category. For all classification tasks, we report 10-fold cross-validation results. For our experiments on immigration and smoking issues, in each fold, we use 30,023 sentences for training, 3,335 for validation, and 3,706 for testing.

As mentioned earlier, the majority of the articles on smoking were annotated only once and the reported inter-annotator agreement on this set is very low, therefore, we further removed the irrelevant category and replicated the experiments on only the immigration set, where at least two annotators agreed upon. Table 1 shows the statistics of the resulting immigration dataset. On this set, in each fold, we use 21,980 sentences for training, 2,442 for validation, and 2,713 for testing.

5 Methods

Here, we present our deep learning–based methods for frame classification. Treating a frame as a sequence of tokens, we explore the use of long short-term memories (Hochreiter and Schmidhuber, 1997) and bi-directional LSTMs (Graves et al., 2013) (BLSTMs), and gated recurrent units (GRU) (Cho et al., 2014) to model the frames. LSTMs and gated recurrent units are types of recurrent neural network that were designed to deal with long-term dependencies, and have been used effectively in the literature to represent long sequences.

To represent the frames, we use word embeddings of the sentences as an input of the model, followed by a single regular LSTM layer, and a sigmoid output layer for multi-class classification.[3] We decided to use a sigmoid function for the output layer because some sentences in our data are assigned multiple labels. We further replace the sigmoid function with a softmax function in the output layer for comparison. We have two settings for initializing our word representations: (1) publicly available GloVe pre-trained word embeddings[4] (Pennington et al., 2014) (300-dimensional vectors trained on Common Crawl data), and (2) embeddings that are constructed on the fly by the LSTM (without any pre-trained word embeddings; we use dropout of 0.2). [5]

We further explore the use of bi-directional LSTMs to represent frame sentences. A bi-directional LSTM consists of two LSTMs running on the input sequence as well as the reverse of the input sequence, thereby allowing the hidden state to capture past and future information (Graves et al., 2013). The motivation behind using this model is to allow the recurrent neural networks to decide what sentence context is important for the classification. The input layer relies on the word embeddings that we mentioned above. We took two approaches to use the pre-trained embeddings: we allowed the embedding weights to be updated during the training (with dropout of 0.2), and we also prevented the embeddings from being updated. The output of the bi-directional LSTM layer (similar to the experiments with the LSTM model and GRU model) was passed to a dropout layer (Hinton et al., 2012) with a rate of 0.2-0.5 to avoid over-fitting, and then to a sigmoid layer to predict the class label of the input sentence. Similar to the experiment with the LSTM model, we replaced the sigmoid layer with a softmax layer for comparison. We further use gated recurrent units, which have shown to improve the performance of recurrent neural networks. All models (LSTM, BLSTM, and GRU) were trained with categorical cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 5 epochs. We experiment with 128 units for all models and restrict the vocabulary to 10,000 most frequent words (for the BLSTM model, we also used

[2] Using NLTK (Bird et al., 2009).

[3] Using https://keras.io/

[4] http://nlp.stanford.edu/projects/glove/

[5] We further used publicly available word2vec pre-trained word embeddings (Mikolov et al., 2013) (300-dimensional vectors trained on the Google News corpus), but achieved similar results.

Table 2: The performance of different models for classification of 5 frames on both immigration and smoking (10-fold cross-validation).

Model		Accuracy (%)
Majority Class (frame 5)		25.4
Uni-, bi-grams (tf-idf)		54.2
LDA 20-topics		53.3
LDA 50-topics		53.9
LDA 100-topics		53.2
Sum of vectors, pre-trained GloVe		60.2
fastText		62.0
LSTM (128 units) no pre-trained embeddings	10K	64.5
LSTM (128 units) GloVe	10K	66.7
LSTM (128 units) pre-trained GloVe	10K	67.5
BLSTM (128 units) no pre-trained embeddings	10K	64.6
BLSTM (128 units) GloVe	10K	66.8
BLSTM (128 units) pre-trained GloVe	10K	67.8
GRU (128 units) GloVe	10K	68.1
GRU (128 units) pre-trained GloVe	10K	**68.7**

Table 3: The performance of different models for 16-way classification (15 frames plus the irrelevant category) (10-fold cross-validation); B(LSTM) and GRU models use pre-trained GloVe embeddings

Model		Accuracy (%)
Majority Class (frame 5)		15.3
uni-, bi-grams (tf-idf)		38.7
50-topics		36.8
Sum of vectors, word2vec		43.2
Sum of vectors, GloVe		43.2
fastText		48.5
LSTM (128 units)	10K	52.1
BLSTM (128 units)	10K	52.5
GRU (128 units)	10K	**53.7**

Table 4: Confusion matrix for GRU with GloVe (5 classes) specified with frame number

		Predicted				
		1	5	6	7	13
Actual	1	**441**	32	28	31	35
	5	30	**705**	74	159	58
	6	57	149	**268**	55	109
	7	26	78	48	**558**	33
	13	40	40	58	27	**603**

the full vocabulary, but achieved a similar performance).

The baselines that we use are majority class and a random forest classifier[6] with 90 trees trained with bag-of-words representations of the sentences. We use both unigrams and bigrams, weighted using *tf-idf*. We further experiment with 20, 50, 100 topic features derived from the GibbsLDA++[7] implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). To represent the sentences with the topics, standard English stopwords were removed, and then tokens were lemmatized to their base form. To estimate the parameters, we used $\alpha = \frac{50}{K}$ (K= number of topics) and $\beta = 0.001$, and ran 1,000 Gibbs sampling iterations and estimated the model at every 100 iterations.

Further, we trained a random forest classifier with sentence vectors obtained by summing the pre-trained word embeddings.

Additionally, we used the fastText (Joulin et al., 2016) classifier based on the skip-gram model, where each word is represented as a bag of character n-grams and the classification is performed through a hierarchical softmax.

6 Results and Discussion

Multi-class Classification All classification results are reported in terms of accuracy. Tables 2 and 3 present the frame detection results for the sets of 5 and 15 frames, plus irrelevant category (sixteen-way classification) respectively on both immigration and smoking issues. The models specified with "*pre-trained*" do not update the embeddings during the training process, whereas the others do update them. All the models reported here used 500 maximum string length with

[6]Using scikit-learn (Pedregosa et al., 2011).
[7]http://gibbslda.sourceforge.net/

Table 5: The performance of different models for classification of 5 frames on only immigration (10-fold cross-validation).

Model		Accuracy (%)	F_1(%)
Majority Class (frame 13)		24.1	–
Uni-, bi-grams (tf-idf)		64.8	62.4
LSTM (128 units) GloVe	10K	70.5	69.9
LSTM (128 units) pre-trained GloVe	10K	70.5	70.2
BLSTM (128 units) GloVe	10K	70.0	69.7
BLSTM (128 units) pre-trained GloVe	10K	70.2	69.8
GRU (128 units) GloVe	10K	70.2	69.7
GRU (128 units) pre-trained GloVe	10K	**71.2**	**70.7**

Table 6: The performance of different models for 15-way classification on immigration set (10-fold cross-validation)

Model	Accuracy (%)	F_1 (%)
Majority Class (frame 13)	15.7	–
uni-, bi-grams (tf-idf)	49.7	44.5
LSTM (128 units)	57.7	56.0
BLSTM (128 units)	57.4	56.0
GRU (128 units)	**58.7**	**57.1**

Table 7: The performance of one-against the others classification achieved by GRU model on immigration set (10-fold cross-validation)

Frame	Accuracy	F_1	Majority class
1	**92.5**	**92.2**	85.3
5	**84.3**	**83.8**	76.0
6	84.9	82.6	84.6
7	**89.9**	**89.6**	78.2
13	**89.3**	**89.3**	75.9

mini-batches of 50 (we also experimented with smaller string length and mini-batches; however, the models achieved lower accuracies). On the combined set, the best accuracy (68.7%) was obtained by the GRU model using 300-dimension GloVe word vectors without being updated, 500 maximum string length with mini-batches of 50. This was slightly better than the results of uni-directional LSTM and bi-directional LSTM models, which achieve similar performance. We did not observe any performance improvement for the models when the word embeddings were updated. The models achieved very similar results with sigmoid and softmax functions. None of the models that learned the embeddings on the fly outperformed their counterparts initialized with GloVe embeddings, this shows that the semantics that are captured in word embeddings are useful for representing frames. The LSTM, BLSTM, and GRU models all outperformed the baseline random forest classifier with sentence vectors obtained by summing the pre-trained word-embeddings, this shows that this baseline classifier cannot learn the sentence representation of frames.

Using the full vocabulary (about 30,000) did not impact the performance of the BLSTM model with GloVe embeddings. All LSTM, BLSTM, and GRU models yielded at least a 10-point improvement over the random forest classifier trained with topics. A confusion matrix for the best-performing GRU model is shown in Table 4. The *policy prescription and evaluation* frame is often misclassified as the *legality, constitutionality, and jurisprudence* frame, which can be expected, as these frames are more likely to have overlapping expressions.

Tables 5 and 6 present the frame detection results for the sets of 5 and 15 frames on immigration corpus respectively. LSTM and BLSTM models perform similarly on the immigration set as well. The best performance (71.2%) is achieved again by GRU model, which is about 6-point above the bag-of-words baseline.

One-against-others Classification We wanted to see how different frames were effected by the model, so we performed a one-against-others classification, where each frame is tested against the rest of frames in the corpus. Table 7 presents the results. We only considered the five most frequent frames. The *political* and *crime and punishment* frames are recognized better than the other frames. While the training set for frame *economic* is smaller than the training set for *legality* frame, *economic* frame was detected more accurately. This is probably due to the unambiguous

cues and phrases regarding monetary and financial expressions, such as *dollars* and *middle class* that are associated with this frame. The most ambiguous frame is *policy prescription and evaluation*.

7 Conclusion

In this study, we motivated the importance of recognizing generic frames at the sentence level in news articles. In order to represent frames effectively, we employed recurrent neural net models. We showed that our approach achieved better performance compared to classifiers trained with topics and other strong baseline models. There are several potential directions for future work. First, we could study the interaction between the primary frame and the frames at the sentence level found in the article. Another interesting direction is to apply our model to other genre of discourse.

Acknowledgments

We are grateful to Suzanne Stevenson and Frank Rudzicz for helpful comments. This research is financially supported by Natural Sciences and Engineering Research Council of Canada.

References

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1472–1482.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*. pages 49–58.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, pages 110–115.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. volume 2, pages 438–444.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1410–1420.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259,* .

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science* 10.

Claes H. De Vreese. 2005. News framing: Theory and typology. *Information Design Journal + Document Design* 13(1):51–62.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4):51–58.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, pages 273–278.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 751–762.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* .

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. 2015. Using summarization to discover argument facets in online idealogical [sic] dialog. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. pages 430–440.

Nona Naderi and Graeme Hirst. 2016. Argumentation mining in parliamentary discourse. In Matteo Baldoni et al., editor, *Principles and Practice of Multi-Agent Systems*, Springer International Publishing, pages 16–25.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea Party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1438–1448.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1629–1638.