MULTITASK LEARNING FOR DARPA LORELEI'S SITUATION FRAME EXTRACTION TASK

Karan Singla and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, USA

ABSTRACT

This paper describes a novel approach of multitask learning for an end-to-end optimization technique for document classification. The application motivation comes from the need to extract "Situation Frames (SF)" from a document within the context of DARPA's LORELEI program targeting humanitarian assistance and disaster relief. We show the benefit of our approach for extracting SF: which includes extracting document types and then linking them to entities. We jointly train a hierarchical document classifier and an auto-encoder using a shared word-level bottleneck layers. Our architecture can exploit additional monolingual corpora in addition to labelled data for classification, thus helping it to generalize over a bigger vocabulary. We evaluate these approaches over standard datasets for this task. Our methods show improvements for both document type prediction and entity linking.

Index Terms— Multitask learning, text classification, situation awareness

1. INTRODUCTION

Multi-task learning (MTL), a widely used approach in NLP, comes in many guises: joint learning (1; 2; 3), learning to learn (4), and learning with auxiliary tasks (5; 6; 7) are some of the names used for it. MTL has also helped achieve state-of-art results for wide range of NLP problems (1; 8; 6; 9). (8) show jointly modeling the target word sequence and its dependency tree structure helps to improve dependency parsing results. Recently (2; 10) show that multitask learning can also help transfer knowledge among tasks through shared lower level layers. For a thorough survey of multitask learning objectives in NLP, please refer to (11).

In this work, we jointly train two tasks, namely, *English auto-encoder* and a Hierarchical Attention based Document Classifier (HADC). HADC classifier, is similar to (12) where given an input document it exploits the hierarchical structure of a document. It uses Bi-LSTMs and self-attention mechanism for encoding words to sentence embeddings, and then sentences to document embedding. Our English auto-encoder is similar to (13) but adapted to a monolingual scenario. It aims to contain most information about a sentence in contextualized word embeddings (i.e., output of word level Bi-directional LSTM layer which takes word embeddings as input). The HADC and English auto-encoder share the word level word embedding and Bi-LSTM layer. In the multitask setup, we train both these tasks jointly using a weighted loss. We hypothesize that due to joint training both these tasks can inform each other through shared layers, which enables HADC classifier to be trained in an end-to-end fashion on a bigger vocabulary than labelled data.

Our motivating application comes from DARPA LORELEI program which annotates and aims to make systems that can provide that Situation awareness results in form of Situation Frames (SF) (14), notably for humanitarian assistance and disaster relief. Given a speech recording or a text document (including social media), a system should predict SFs. An example SF can be seen below.

```
{
  "DocumentID": "HIN_SN_000370",
  "Type": "medical-assistance",
  "TypeConfidence": 0.88,
  "PlaceMention": {
    "Start": 20,
    "End": 25,
    "EntityType": GPE,
  },
  "Justification": "Segment-0",
}
```

SF is defined by a *(Type, PlaceMention)* pair, where *Type* refers to a situation type and *PlaceMention* refers to a location. Linking of a situation *Type* to a *PlaceMention* is called *Localization* (15). There can be multiple frames in a document. For a system that can work across languages, people either use machine translation (MT) systems to go from a given language to English and then use an English-only system for SF prediction (16; 17; 18; 19; 15). Alternatively, one can use pre-trained fixed multilingual word embeddings as low level features directly for the SF system (20; 21; 22). For speech sources, a widely followed approach is to transcribe the audio sessions using an ASR and then translate them into English using a Machine Translation system (23; 18; 24). Since a majority of the annotated SF resources are in English, therefore in this paper, we focus on applying the concept of MTL learning for the task of predicting situation frames for En-

glish. We propose an end-to-end system which takes as input a pseudo-entity level document (using segments in which a *PlaceMention* appears) alongside the original source document, and predicts whether there should be a *Type* linked to this *PlaceMention* or not.

Our multitask models show improvements for F-score measure across languages (translated into English) for both *Type* and Localization prediction. We believe these improvements are due to the system's ability to avoid overfitting and generalize as our document classifier is trained in a multitask fashion with an English auto-encoder. Results also suggest that doing an end-to-end *localization* for extraction of situation frames gives better results.

2. DATA

LDC[1] provides annotated data packages in multiple languages along with their translations. A situation Type is selected from the fixed inventory of eleven labels, namely, *Evacuation*, *Food-supply*, *Search/rescue*, *Utilities*, *Infrastructure*, *Medical-assistance*, *Shelter*, *Water supply*, Terrorism, *Crime-voilence* and *Regime-change*

We use all data available for following languages: English (En), Spanish (es), Mandarin (Mn), Ugyhur (Ug), Tagrinya (Tg), Oromo (Or), Bengali (Bn), Hindi (Hi), Thai (Th) and Zulu (zu), Kinyarwanda(Kn) and Sinhalese (Sn) along with their English translations and SF annotations. We also collect additional annotations for English tweets (En-twt) and assume that all *PlaceMentions* as linked to the *Type* of the tweet. Table 1 shows frequencies of collected data.

3. MULTITASK LEARNING

We propose a multitask architecture for the extraction of situation frames. The main idea is to jointly train a hierarchical attention document classifier and an English auto-encoder. Both these tasks share the word level variables i.e., word embeddings and word level Bi-LSTM layer (context dependent word embeddings)

3.1. English Auto-encoder

Our auto-encoder uses contextualized word embeddings i.e., Bi-LSTM layer shared between two tasks. There are various implementations of LSTMs available; in this work we use an implementation based on (25) which comes as a part of Tensorflow (26). Sentence embeddings are obtained by applying a max-pooling operation over the output of a word level Bi-LSTM layers. Similar to (13) these sentence embeddings are also used to initialize the decoder LSTM through a linear transformation, and are also concatenated to its input embeddings at every time step. As shown in Figure 1, there is no other connection between the encoder and the decoder as we want contextualized word embeddings to capture all the information of a sentence. This is done because contextualized word embeddings (output of word-level Bi-LSTM layers) is shared with the Document classifier described in the following section.

3.2. Hierarchical Attention based Document Classifier (HADC)

Our HADC architecture is similar to (12), where given an input sentence X, we first feed word embeddings W to a Bi-LSTM layer to get contextualized word embeddings H. These contextualized word embeddings are then passed through a dense layer to get H'. H' is then passed to a self-attention layer to get a representation for each sentence S in the document. We use an attention layer (equations 1-3) with an internal context vector (12).

$$k_i = \tanh(WH'_i + b) \tag{1}$$

$$\alpha_i = \mathrm{softmax}(k_i^T a) \tag{2}$$

$$S = \sum_i \alpha_i h_i \tag{3}$$

The attention layer first applies a one-layer MLP to its inputs H'_i to derive the keys k_i. Then it computes the attention weights by applying a softmax non-linearity to the inner products between the keys k_i and the internal context vector a . Finally it computes the sentence representation R by taking a weighted average of its inputs. The context vector a is a model parameter that is initialized with uniform weights so that it behaves like an averaging operation at the beginning of the training.

For going from sentences to documents, we follow the same architecture again. i.e., keeping a Bi-LSTM layer to contextualize sentence embeddings and then an attention layer to get a document representation. This document representation is then fed to task specific linear and sigmoid layer. We use the sigmoid layer because our tasks are binary multilabel.

Multitask Loss: Our final multitask loss function is made of two terms, auto-encoder reconstruction loss A and HADC classification loss B. The $Loss$ is defined as $B + \alpha * A$, where α is empirically set to 0.3 for all multitask experiments.

3.3. Training Routine

We remove all punctuation and lower-case the data. Words that occur less than 5 times are replaced with the <unk> symbol. We initialize the word emdedding layer using 300 dimensional pre-trained Word2Vec embeddings[2]. We use batch size of 50 sentences for auto-encoder and 20 documents for HADC. An additional 300 random sentences from

[1] https://www.ldc.upenn.edu/

[2] https://code.google.com/archive/p/word2vec/

Language	En	En-twt	Mn	Ug	Tg	Or	Kn	Sn	Bn	Hi	Th	Zu
Entities / Entity Documents	3152	-	1368	1120	1712	2484	627	566	746	624	682	975
Documents	756	2934	132	256	656	1264	339	331	144	162	158	408

Table 1: Frequency for documents and gold standard entities for each language from LDC.

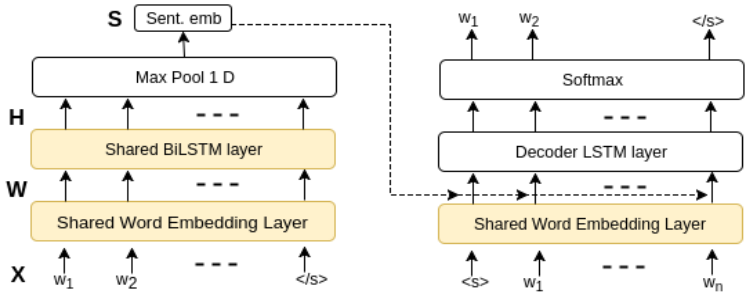

Fig. 1: English Auto-encoder architecture to learn contextualized word embeddings. Boxes in yellow shows variables which are shared among tasks.

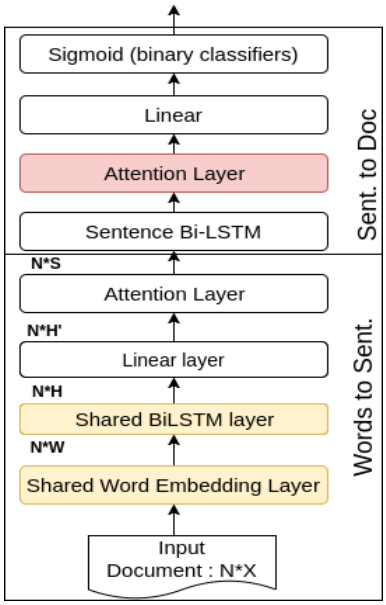

Fig. 2: Hierarchical Attention based Document Classifier (HADC) architecture which shares lower level layers with English Autoencoder.

HADC mini-batch are also added to the auto-encoder mini-batch. The LSTM hidden state dimension is 256 and 128 for word-level and sentence-level layers, respectively. We use dropout at the embedding layer and before the sentence-level layer in HADC with drop probability of 0.3. We use Adam optimizer (27) with a learning rate of 0.001 and an exponential decay of 0.98 after 10K steps (1 step is 1 mini-batch). The auto-encoder is pre-trained for 30K steps, before we begin joint multi-task training for document classification. We realized this pre-training helps simultaneous convergence of both the tasks.

Similar to prior works (16; 17; 18; 19; 15), we use the ReliefWeb corpus[3] of disaster-related documents to pre-train the model. The corpus contains disaster-related documents from various sources annotated for theme and disaster type, where theme labels are similar to topics discussed (food, water). We get inventory of 40 categories for the task of multi-label classification of documents and use approximately 120K documents for training and 52K documents for testing. We keep another 10K documents for the validation set. Our multitask model shows slightly better performance for ReliefWeb type prediction task. We achieve F-score of 72.5 and 73.1 for HADC and jointly trained multi-task HADC model respectively.

We have three different SF models:

- **HADC:** Pre-trained Relief web model is fine-tuned for situation type prediction by replacing last sigmoid layer.
- **HADC SepATT:** Here we use a different randomly uniform initialized internal context vector for sentence-level attention layer while predicting situation frames.
- **HADC Multi:** The English auto-encoder is pre-trained for 30K steps, before we start joint multitask training along with HADC SepATT model.
- **HADC Multi*extra:** Same as HADC-Multi, but here we use additional 500K sentences from Europarl(28) corpus for training English autoencoder

4. LOCALIZATION

An important aspect of extracting a SF is linking a situation type to a PlaceMention. For this we build upon the approach from (18). It follows the hypothesis that segments in which an entity mention appears is predictive of the situation type. So to localize, they use a simple solution of creating location-specific sub-documents and attempt to classify them using the same models. For each detected *PlaceMen-*

[3]ReliefWeb website. http://reliefweb.int/. Retrieved 31 Mar 2016

tion, all sentences/segments that contain said *PlaceMention* are collected to form a dummy "document" per PlaceMention. These dummy documents are then passed through the SF model again, creating a set of *Type* labels per *PlaceMention*. The PlaceMention-level Types are filtered by the document-level Types: Types not detected during the document-level pass were not allowed at the entity level.

We follow the same hypothesis as used by (18) but instead of creating situation frames by just filtering out PlaceMention-level types using types predicted for a document, we provide posteriors of document level types as an input to predict PlaceMention-level type. We combine the loss of PlaceMention-level and document-level type prediction using a scaling parameter β and do a joint optimization. This allows our model to train end-2-end for predicting *Type* for a given *PlaceMention*. We performed experiments using various β values $(0.3, 0.5, 0.8)$ and found 0.5 gives best results for Localization.

5. EVALUATION

We use English translations of Tigrinya (Tg), Oromo (Or), Kinyarwanda (Kn) and Sinhalese (Sn) for testing as they were official test languages for last two official LORELEI DARPA evaluations and remaining documents for training and validation (4950 documents and 11601 Entity-segments-documents using *PlaceMentions*). We randomly keep 10 % of data for validation.

System	Tg	Or	Kn	Sn
HADC	0.52	0.24	0.46	0.25
HADC SepATT	0.53	0.27	0.46	0.30
HADC Multi	0.50	**0.32**	0.51	0.41
HADC Multi*extra	**0.62**	0.31	**0.52**	**0.48**

Table 2: Situation Type F-scores using Human translations. *extra means 1M additional monolingual data from EUROPARL was used in training for English auto-encoder

System	Tg	Or	Kn	Sn
Nikos et al. (18)	0.21	0.08	0.20	0.13
HADC SepATT	0.24	0.12	0.24	0.19
HADC Multi	**0.27**	0.16	0.26	0.21
HADC Multi*extra	0.25	**0.20**	**0.30**	**0.25**

Table 3: Situation Type + Place (Gold Standard PlaceMentions) F-scores. *extra means 1M additional monolingual data from EUROPARL was used in training for English auto-encoder

Table 2 shows our results for F-score measure for the multi-label situation *Type* prediction at the document level. Our multi-task model *HADC-Multi* shows improvement over the baseline *HADC* model except for Tigrinya (Tg). This suggests multitasking helps achieve better results. The results for the experiments with additional monolingual data (*extra) shows best performance. This suggests that these improvements are mainly due to added vocabulary provided by additional monolingual data. With regards to doing *Localization*, we report f-score measures for the *(Type, PlaceMention)* tuple. Table 3 shows results for Localization. Similar to *Type* prediction results model with additional monolingual data (*extra) outperforms other compared models. All our end-to-end optimized models show gains when compared to the previous approach followed by (18).

6. CONCLUSIONS & FUTURE WORK

Our results suggest that joint multi-task learning of contextualized word embeddings using an English auto-encoder and end-to-end optimization for extracting situation frames is a promising direction. We believe our multitask architecture can be improved further by straightforward modifications to the output of shared layers like applying domain adversarial penalty to the contextualized word embeddings. We intend to apply, adapt and test our architecture to other NLP amd speech tasks in the future like sentiment classification and emotion recognition.

7. ACKNOWLEDGEMENTS

We would like to thank other ELISA team members who contributed to idea formulation, Dogan Can (USC) and Jonathan May (ISI). This work was supported by the U.S. DARPA LORELEI Program under Contract No. HR0011-15-C-0115 with the University of Southern California.

References

[1] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher, "A joint many-task model: Growing a neural network for multiple nlp tasks," *arXiv preprint arXiv:1611.01587*, 2016.

[2] Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Learning cross-lingual sentence representations via a multi-task dual-encoder model," *arXiv preprint arXiv:1810.12836*, 2018.

[3] Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 195–204.

[4] Jonathan Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, 1997.

[5] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-

task learning," in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[6] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," 2015.

[7] Otkrist Gupta, Dan Raviv, and Ramesh Raskar, "Multi-velocity neural networks for facial expression recognition in videos," *IEEE Transactions on Affective Computing*, 2017.

[8] Shuangzhi Wu, Dongdong Zhang, Nan Yang, Mu Li, and Ming Zhou, "Sequence-to-dependency neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, vol. 1, pp. 698–707.

[9] Anders Søgaard and Yoav Goldberg, "Deep multi-task learning with low level tasks supervised at lower layers," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, vol. 2, pp. 231–235.

[10] Karan Singla, Dogan Can, and Shrikanth Narayanan, "A multi-task approach to learning multilingual representations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 214–220.

[11] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[12] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[13] Mikel Artetxe and Holger Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *arXiv preprint arXiv:1812.10464*, 2018.

[14] Caitlin Christianson, Jason Duncan, and Boyan Onyshkevych, "Overview of the darpa lorelei program," *Machine Translation*, vol. 32, no. 1-2, pp. 3–9, 2018.

[15] Leon Cheung, Thamme Gowda, Ulf Hermjakob, Nelson Liu, Jonathan May, Alexandra Mayn, Nima Pourdamghani, Michael Pust, Kevin Knight, Nikolaos Malandrakis, et al., "Elisa system description for lorehlt 2017," 2017.

[16] Rada Mihalcea, Carmen Banea, and Janyce Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 976–983.

[17] Lei Shi, Rada Mihalcea, and Mingjun Tian, "Cross language text classification by model translation and semi-supervised learning," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1057–1067.

[18] Nikolaos Malandrakis, Anil Ramakrishna, Victor Martinez, Tanner Sorensen, Dogan Can, and Shrikanth Narayanan, "The elisa situation frame extraction for low resource languages pipeline for lorehlt'2016," *Machine Translation*, vol. 32, no. 1-2, pp. 127–142, 2018.

[19] Matthew Wiesner, Chunxi Liu, Lucas Ondel, Craig Harman, Vimal Manohar, Jan Trmal, Zhongqiang Huang, Najim Dehak, and Sanjeev Khudanpur, "Automatic speech recognition and topic identification for almost-zero-resource languages," in *Proc. Interspeech*, 2018.

[20] Yuhong Guo and Min Xiao, "Cross language text classification via subspace co-regularized multi-view learning," *arXiv preprint arXiv:1206.6481*, 2012.

[21] Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi, "Cross-lingual text classification via model translation with limited dictionaries," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 95–104.

[22] Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy, "Low-resource cross-lingual event type detection via distant supervision with minimal effort," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 70–82.

[23] Nikolaos Malandrakis, Ondrej Glembek, and Shrikanth S Narayanan, "Extracting situation frames from non-english speech: Evaluation framework and pilot results.," in *INTERSPEECH*, 2017, pp. 2123–2127.

[24] Pavlos Papadopoulos, Ruchir Travadi, Colin Vaz, Nikolaos Malandrakis, Ulf Hermjakob, Nima Pourdamghani, Michael Pust, Boliang Zhang, Xiaoman Pan, Di Lu, et al., "Team elisa system for darpa lorelei speech evaluation 2016.," in *INTERSPEECH*, 2017, pp. 2053–2057.

[25] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[26] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*. Citeseer, 2005, vol. 5, pp. 79–86.