There and Back Again: Cross-Lingual Transfer Learning for Event Detection

Tommaso Caselli, Ahmet Üstün
Rikjuniversiteit Groningen, Groningen, The Netherlands
{t.caselli|a.ustun}@rug.nl

Abstract

English. In this contribution we investigate the generalisation abilities of a pre-trained multilingual Language Model, namely Multilingual BERT, in different transfer learning scenarios for event detection and classification for Italian and English. Our results show that zero-shot models have satisfying, although not optimal, performances in both languages (average F1 higher than 60 for event detection *vs*. average F1 ranging between 40 and 50 for event classification). We also show that adding extra fine-tuning data of the evaluation language is not simply beneficial but results in better models when compared to the corresponding non zero-shot transfer ones, achieving highly competitive results when compared to state-of-the-art systems.

1 Introduction

Recently pre-trained word representations encoded in Language Models (LM) have gained lot of popularity in Natural Language Processing (NLP) thanks to their ability to encode high level syntactic-semantic language features and produce state-of-the-art results in various tasks, such as Named Entity Recognition (Peters et al., 2018), Machine Translation (Johnson et al., 2017; Ramachandran et al., 2017), Text Classification (Eriguchi et al., 2018; Chronopoulou et al., 2019), among others. These models are pre-trained on large amounts of unannotated text and then fine-tuned using the induced LM structure to generalise over specific training data. Given their success in monolingual environments, especially for English, there has been a growing interest in the development of *cross-lingual* as well as *multilingual* representations (Vulić and Moens, 2015; Ammar et al., 2016; Conneau et al., 2018; Artetxe et al., 2018) to investigate different cross-lingual transfer learning scenarios, including zero-shot transfer, i.e. the direct application of a model fine-tuned using data in one language to a different test language.

Following the approach in Pires et al. (2019), in this paper we investigate the generalisation abilities of Multilingual BERT (Devlin et al., 2019) [1] on English (EN) and Italian (IT). Multilingual BERT is particularly well suited for this task because it easily allows the implementation of cross-lingual transfer learning, including zero-shot transfer.

We use event detection as our downstream task, a highly complex semantic task with a well established tradition in NLP (Ahn, 2006; Ji and Grishman, 2008; Ritter et al., 2012; Nguyen and Grishman, 2015; Huang et al., 2018). The goal of the task is to identify event mentions, i.e. linguistic expressions describing "things" that happen or hold as true in the world, and subsequently classify them according to a (pre-defined) taxonomy. The complexity of the task relies in its high dependence on the context of occurrence of the expressions that may trigger an event mention. Indeed, the *eventiveness* of an expression is prone to ambiguity because there exists a continuum between eventive and non-eventive readings in the space of event semantics (Araki et al., 2018). Such intrinsic ambiguity of event expressions challenges the generalisation abilities of stochastic models and allows to investigate advantages and limits of transfer learning approaches when semantics has a pivotal role in the resolution of a problem/task.

We explore different multi-lingual and cross-

[1] https://github.com/google-research/bert

lingual aspects of transfer learning with respect to event detection through a series of experiments, focusing on the following research questions:

RQ1 How well do Multilingual BERT fine-tuned models generalise in zero-shot transfer learning scenarios on both languages?

RQ2 Do we obtain more robust models by fine-tuning zero-shot models with additional (training) data of the evaluation language?

Our results show that Multilingual BERT obtains satisfying performances in zero-shot scenarios for the identification of event triggers (average F1 63.53 on Italian and 66.79 on English), while this is not the case for event classification (average F1 42.86 on Italian and 51.26 on English). We also show that extra fine-tuning the zero-shot models with data of the evaluation language is not just beneficial, but it actually gives better results than models fine-tuned on the corresponding test language only (i.e. fine-tuning and test in the same language), and achieves competitive results with state-of-the-art systems developed using dedicated architectures. Our code is available (`https://github.com/ahmetustun/BertForEvent`).

2 Data

We have used two corpora annotated with event information: the TempEval-3 corpus (TE3) for English (UzZaman et al., 2013) and the EVENTI corpus for Italian (Caselli et al., 2014). The corpora have been independently annotated with language specific annotation schemes, grounded on a shared metadata markup language for temporal information processing, ISO-TimeML (ISO, 2008), thus sharing definitions and tags' names for the markable expressions. The corpora are composed by contemporary news articles[2] and have been developed in the context of two evaluation campaigns for temporal processing, namely TempEval-3 and EVENTI@EVALITA 2014.

Events are defined as anything that can be said to happen, or occur, or hold true, with no restriction to parts-of-speech (POS), including verbs, nouns, adjectives, and also prepositional phrases (PP). Every event mention is further assigned to one of 7 possible classes: OCCURRENCE, ASPECTUAL, PERCEPTION, REPORTING, I(NTESIONAL) STATE, I(NTENSIONAL) ACTION, and STATE, capturing the relationship the event participates (such as factual, evidential, reported, intensional). Although semantically interoperable, one of the most relevant annotation differences that may impact the evaluation of the zero-shot models concerns the marking of modal verbs and copulas introducing event nouns, adjectives or PPs. While in English these elements are never annotated as event triggers, this is done in Italian. A detailed description of additional language specific adaptations and differences between English and Italian is reported in Caselli and Sprugnoli (2017).

Tables 1 and 2 illustrate the distribution of the annotation of events for POS (token based) and classes (event based), respectively. Both corpora, when released, did not explicitly have a development section. Following previous work (Caselli, 2018), we generated development sets by excluding from the training data all the documents that composed the test data for Italian and English in the SemEval 2010 TempEval-2 campaign (Verhagen et al., 2010).

The Italian corpus is larger than the corresponding English version, although the distribution of events, both per POS and per class, is comparable. The different distribution of the REPORTING, I_STATE, I_ACTION, and STATE classes reflects differences in annotation instructions rather than language specific characteristics. For instance, in Italian, the class REPORTING is assigned only if the event mention is an instance of a speech verb/noun (*verba/nomina dicendi*), while in English this constraint is less strict.

3 Model

Multilingual BERT (Bidirectional Encoder Representations from Transformers) shares the same framework of the monolingual English BERT$_{BASE}$ (Devlin et al., 2019). BERT is a pre-trained LM that improves over existing fine-tuning approaches by jointly conditioning on both left and right contexts in all layers to generate pre-trained deep bidirectional representations. Multilingual BERT's architecture contains an encoder consisting of 12 Transformer blocks with 12 self-attention heads (Vaswani et al., 2017), and

[2]We have excluded the extra test set on historical news from the Italian data set, and the automatically annotated training set from the English one.

	TE3			EVENTI			
POS	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Examples**
Verb	8,141	393	542	11,269	193	2,426	en:*run*; it:*correre*
Noun	2,268	124	175	6,710	111	1,499	en:*attack*; it:*attacco*
Adjectives	165	8	21	610	9	118	en:*(is) dormat*; it:*(è) dormiente*
Other/PP	29	1	8	146	1	25	en:*on board*; it:*a bordo*
Total	10,603	526	746	18,735	314	4,068	

Table 1: Distribution of events per POS in each corpus per Training, Development, and Test data.

	TE3			EVENTI			
Classes	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Examples**
OCCURRENCE	6,530	302	466	9,041	162	1,949	en:*run*; it:*correre*
ASPECTUAL	264	33	35	446	14	107	en:*start*; it:*inizio*
PERCEPTION	79	4	2	162	2	37	en:*see*; it:*vedere*
REPORTING	1,544	67	92	714	8	149	en:*say*; it:*dire*
I_STATE	651	29	36	1,599	29	355	en:*like*; it:*piacere*
I_ACTION	827	57	47	1,476	25	357	en:*attempt*; it:*tentare*
STATE	708	34	68	4,090	61	843	en:*keep*; it:*tenersi*
Total	10,603	526	746	17,528	301	3,798	

Table 2: Distribution of event classes in each corpus per Training, Development, and Test data.

hidden size of 768.

Unlike the original BERT, Multilingual BERT is pre-trained on the concatenation of monolingual Wikipedia pages of 104 languages with a shared word piece vocabulary. One of the peculiar characteristics of this multilingual model is that it does not make use of any special marker to signal the input language, nor has any mechanism that explicitly indicates that translation equivalent pairs should have similar representations.

For the fine-tuning, we use a standard sequence tagging model. We apply a softmax classifier over each token by passing the token's last layer of activation to the softmax layer to make a tag prediction. Since BERT's wordpiece tokenizer can split words into multiple tokens, we take the prediction for the first token (piece) per word, ignoring the rest. No parameter tuning was performed, learning rate was set to 1e-4, and batch size to 8.

4 Experiments

Event detection is best described as composed by two sub-tasks: first, identify if a word, *w*, in a given sentence *S* is an instance of an event mention, ev_w; and subsequently, assign it to a class *C*, $ev_w \in C$. We break the experiments in two blocks: in the first block, we investigate the quality of the fine-tuned Multilingual BERT models on the identification of the event mentions only. This is an easier task with respect to classification, as it can be framed as a binary classification task. In this way, we can actually have a sort of maximal threshold of the performance of the zero-shot cross-lingual transfer learning models. In the second block of experiments, we investigate the ability of the models in performing the two sub-tasks "at once", i.e. identifying and classifying an event mention. This is a more complex task, especially in zero-shot transfer learning scenarios, because the ISO-TimeML classes are assigned following syntactic-semantic criteria: the same word can be assigned to different classes according to the specific syntactic context in which it occurs. For each language pair and direction of the transfer (i.e. EN_{train}–IT_{test} *vs.* IT_{train}–EN_{test}), we also benchmark the performance in monolingual fine-tuned transfer scenarios (i.e. IT_{train}–IT_{test} *vs.* EN_{train}–EN_{test}), to have an upper-bound limit of Multilingual BERT and an indirect evidence of the intrinsic quality of the proposed multilingual model. For the English data, we also test the performance using English BERT_{BASE}, so to better understand limits of the multilingual model.

Finally, we compare our results to the best systems that participated in the corresponding evaluation campaigns in each language, as well as to state-of-the-art systems. In particular, we selected:

- HLT-FBK (Mirza and Minard, 2014), a feature-based SVM model for Italian (best system at EVENTI@EVALITA);

- ATT1 (Jung and Stent, 2013), a feature-based MaxEnt model for English (best system for event detection and classification at TempEval-3);

- CRF4TimeML (Caselli and Morante, 2018),

a feature-based CRF model for English that has obtained state-of-the-art results on event classification;

- Bi-LSTM-CRF (Reimers and Gurevych, 2017; Caselli, 2018), a neural network model based on a Bi-LSTM using a CRF classifer as final layer. The architecture has been originally developed and tested on English (Reimers and Gurevych, 2017), and subsequently adapted to Italian (Caselli, 2018). The English version of the system reports state-of-the-art scores for the event detection task only, while the Italian version obtained state-of-the-art results for detection and classification.

5 Results

All scores for the Multilingual BERT models have been averaged against 5 runs (Reimers and Gurevych, 2017). Subscript numbers correspond to standard deviation scores. Tables 3 and 4 illustrate the results on the Italian test data for the event detection and the event detection and classification sub-tasks, respectively. Results on the English test are illustrated in Table 5 for event detection and in Table 6 for event detection and classification. For each experiment, we also report the number of fine-tuning epochs.

The main take-away is that the portability of the zero-shot models is not the same for the two sub-tasks: for the event detection sub-task, both models obtain close results (average F1 63.53 on Italian *vs*. average F1 66.79 on English), while this is not the case for the event detection and classification sub-task (average F1 42.86 on Italian *vs*. average F1 51.26 on English), suggesting this sub-task as being intrinsically more difficult. We also observe that the zero-shot models have different behaviors with respect to Precision and Recall: the zero-shot transfer on Italian has a high Precision and a low Recall, while the opposite happens on English. [4] The stability of the zero-shot models seems to be influenced by the size of the fine-tuning training data. In particular, zero-shot transfer learning on English consistently results in more stable models, as the lower scores for the standard deviation show when compared to the Italian counterpart (+/- 2.04 for EVENTI$_{train}$ on the TE3 test data *vs*. +/- 7.45 for TE3$_{train}$ on the EVENTI test data for the event detection sub-task; +/- 2.67 for EVENTI$_{train}$ on the TE3 test data *vs*. +/- 3.15 for TE3$_{train}$ on the EVENTI test data for the event detection and classification sub-task).

Annotation differences in the two languages have an impact in the evaluation of the zero-shot models. To measure this, we excluded all modal and copula verbs both as predictions on the English test by the zero-shot Italian model, and as gold labels from the Italian test, when applying the zero-shot English model. In both cases we observe an improvement, with an increase of the average F1 to 72.26 on English and 66.01 on Italian. Although other language specific annotations may be at play, the Italian zero-shot model appears to be more powerful than the English one.

The addition of extra fine-tuning with data from the evaluation language results in a positive outcome, improving performances in both sub-tasks. In three out of the four cases (event detection on English, and event detection and classification on English and Italian) the extra-fine tuning with the full training set of the evaluation language results in better models than the corresponding non zero-shot ones. Adding training material targeting the evaluation test is a well know technique in domain adaptation (Daumé III, 2007). Quite surprisingly with respect to previous work that used this approach, we observe an improvement also with respect to fine-tuned transfer scenarios, i.e. models tuned and tested on the same language, suggesting that the multilingual model is actually learning from both languages.

In terms of absolute scores, our results for the zero-shot scenarios are in line with the findings reported in Pires et al. (2019) for typologically related languages, such as English and Italian. However, limits of zero-shot transfer scenarios seem more evident in semantic tasks when compared to morpho-synatactic ones. For instance, Pires et al. (2019) reports absolute F1 scores comparable to ours on Named Entity Recognition on 4 language pairs, while results on POS tagging achieve an accuracy above 80% on all language pairs. More recently, Wu and Dredze (2019) have shown a similar behavior to our zero-shot scenarios of Multilingual BERT in a text classification task.

[4] For instance, average Precision for event detection is 93.11 on Italian *vs*. 53.19 on English, while average Recall is 51.71 on Italian and 89.92 on English, respectively. A similar pattern is observed for the detection and classification sub-task.

Fine Tuning	Epochs	EVENTI F1
$TE3_{train}$ - *zero-shot*	1	$63.53_{7.45}$
$TE3_{train}$ + $EVENTI_{dev}$	1 + 2	$77.57_{1.73}$
$TE3_{train}$ + $EVENTI_{train}$	1 + 1	$87.17_{0.56}$
$EVENTI_{train}$	1	$87.36_{1.16}$
(Caselli, 2018)	n/a	**87.79**
HLT-FBK	n/a	86.68

Table 3: Event mention detection - test on Italian. Best scores in bold.

Fine Tuning	Epochs	EVENTI F1
$TE3_{train}$ - *zero-shot*	2	$42.86_{3.15}$
$TE3_{train}$ + $EVENTI_{dev}$	1 + 2	$55.38_{1.34}$
$TE3_{train}$ + $EVENTI_{train}$	1 + 3	$\mathbf{73.90}_{0.45}$
$EVENTI_{train}$	2	$73.69_{0.80}$
(Caselli, 2018)	n/a	72.97
HLT-FBK	n/a	67.14

Table 4: Event detection and classification - test on Italian. Best scores in bold.

Fine Tuning	Epochs	TE3 F1
$EVENTI_{train}$ - *zero-shot*	1	$66.79_{2.04}$
$EVENTI_{train}$ + $TE3_{dev}$	1 + 2	$80.67_{1.11}$
$EVENTI_{train}$ + $TE3_{train}$	1 + 1	$81.87_{0.13}$
$TE3_{train}$	1	$81.39_{1.23}$
(Reimers and Gurevych, 2017)[3]	n/a	**83.45**
ATT1	n/a	81.05

Table 5: Event mention detection - test on English. Best scores in bold.

Fine Tuning	Epochs	TE3 F1
$EVENTI_{train}$ - *zero-shot*	2	$51.26_{2.67}$
$EVENTI_{train}$ + $TE3_{dev}$	1 + 2	$64.16_{2.82}$
$EVENTI_{train}$ + $TE3_{train}$	1 + 3	$68.97_{0.94}$
$TE3_{train}$	2	$63.36_{1.47}$
CRF4TimeML	n/a	**72.24**
ATT1		71.88

Table 6: Event detection and classification - test on English. Best scores in bold.

6 Discussion

Extra fine-tuning Extra fine-tuning, even with a minimal amount of data as shown by the results using the development sets, shifts the model's predictions to be more in-line with the corresponding language specific annotations. Furthermore, it reduces the effects of cross-lingual transfer based on the presence of the same word pieces between the fine-tuned and the evaluation languages due to the single multilingual vocabulary of Multilingual BERT (Pires et al., 2019). This also results in an increasing stability of the models and a reduction of the differences in the average scores for Precision and Recall with respect to the zero-shot models.

Comparison to other systems Zero-shot models obtain satisfying, though not optimal, results as they fall far from both the state-of-the-art models and the best performing systems in the corresponding evaluation exercises (i.e. HLT-FBK for Italian and ATT1 for English). Extra fine-tuning with the development data provides competitive models against the best systems in the evaluation exercises only. When the full training data is used for extra fine-tuning in the target evaluation language, results are very close to the state of the art, although only in one case the Multilingual BERT model is actually outperforming it (namely, on event detection and classification for Italian). These models also obtain very competitive results with respect to state-of-the-art systems, indicating that multilinguality does not seem to negatively affect the quality of the pre-trained LM. However, results on English using English BERT_{BASE} appears to be partially in line with this observation. By applying the same settings, we obtain an average F1 on event detection of 82.85,[5] and an average F1 for event detection and classification of 71.09. Although results of the monolingual model are expected to be higher in general, in this case, we observe that the differences in performance between the two tasks are not in the same range. BERT_{BASE} obtains an increase of 2% on event detection but it reaches almost 11% on event detection and classification. Differences in class labelling between English and Italian (see Section 2) can partially explain this behaviour. However, given the sensitivity of event classification to the syntactic context, these results call for further investigation on the encoding of syntactic information between the monolingual and the multilingual BERT models.

Errors Comparing the errors of the zero-shot models is not an easy task mainly because of the language specific annotations in the two corpora. However, focusing on the three major POS, i.e. nouns, verbs, and adjectives, and on the False Negatives only, both models present a similar proportions of errors, with nouns representing the hardest case (53.84% on Italian *vs*. 54.90% on English), followed by verbs (30.29% on Italian *vs*. 17.64% on English), and by adjectives (7.51% on Italian *vs*. 5.88% on English). When observing the classification mismatches (i.e. correct event mention but

[5] Precision: 81.26; Recall: 84.70

wrong class), both models overgeneralise the OCCURRENCE class in the majority of cases. However, zero-shot transfer on English actually extends mis-classification errors mirroring the distribution of the classes of the Italian training data. In particular, it wrongly classifies English REPORTING events as I_ACTION (33.33%), and OCCURRENCE as STATE (15.51%) or I_ACTION (34.48%). Although the syntactic context may have influenced the classification errors, these patterns further highlight the differences in annotations between the two languages.

7 Conclusion

In this contribution we investigated the generalisation abilities of Multilingual BERT on Italian and English using event detection as a downstream task. The results show that Multilingual BERT seems to handle cross-lingual generalisation between Italian and English in a satisfying way, although with some limitations. Limitations in this case come from two sources: annotation differences in the two languages and, partially, the shared multilingual vocabulary. Zero-shot systems appears to be particularly sensitive to the fine-tuning data, and, in these experiments, they provide empirical evidence of the impact of different annotation decisions for events in English and Italian.

We have shown that extra fine-tuning with data of the evaluation language not only is beneficial but it may lead to better systems, suggesting that the multilingual model may be combining information from the two languages, and thus obtaining competitive results with respect to task-specific architectures. This opens up to new strategies for the development of systems by using interoperable annotated data in different languages to improve performances and possibly obtain more robust and portable models across different data distributions.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings.

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable annotation of events and event relations across domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Tommaso Caselli and Roser Morante. 2018. Systems Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Tommaso Caselli and Rachele Sprugnoli. 2017. It-TimeML and the Ita-TimeBank: Language Specific Adaptations for Temporal Annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation - Volume II*, pages 969–988. Springer.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 27–34. Pisa University Press.

Tommaso Caselli. 2018. Italian Event Detection Goes Deep Learning. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods*

in Natural Language Processing, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. *ACL 2007*, page 256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July. Association for Computational Linguistics.

SemAf/Time Working Group ISO, 2008. *ISO DIS 24617-1: 2008 Language resource management - Semantic annotation framework - Part 1: Time and events*. ISO Central Secretariat, Geneva.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. *Proceedings of ACL-08: HLT*, pages 254–262.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Hyuckchul Jung and Amanda Stent. 2013. Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 20–24, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Paramita Mirza and Anne-Lyse Minard. 2014. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Fourth International Workshop EVALITA 2014*, pages 44–49.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. 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: Short Papers)*, volume 2, pages 365–371.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. 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: Short Papers), volume 2, pages 719–725.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.