Proceedings of the ResT-UP2 Proceedings @LREC2022, pages 1–7
Marseille, 20-25 June 2022

A First Attempt at Unreliable News Detection in Swedish

Ricardo Muñoz Sánchez[1*], **Eric Johansson**[2*], **Shakila Tayefeh**[2*], **Shreyash Kad**[2*]
[1]The University of Gothenburg, [2]Chalmers University of Technology
Gothenburg, Sweden
ricardo.munoz.sanchez@svenska.gu.se, {ericjoha, tayefeh, shreyash}@student.chalmers.se

Abstract

Throughout the COVID-19 pandemic, a parallel infodemic has also been going on such that the information has been spreading faster than the virus itself. During this time, every individual needs to access accurate news in order to take corresponding protective measures, regardless of their country of origin or the language they speak, as misinformation can cause significant loss to not only individuals but also society. In this paper we train several machine learning models (ranging from traditional machine learning to deep learning) to try to determine whether news articles come from either a reliable or an unreliable source, using just the body of the article. Moreover, we use a previously introduced corpus of news in Swedish related to the COVID-19 pandemic for the classification task. Given that our dataset is both unbalanced and small, we use subsampling and easy data augmentation (EDA) to try to solve these issues. In the end, we realize that, due to the small size of our dataset, using traditional machine learning along with data augmentation yields results that rival those of transformer models such as BERT.

Keywords: Text categorisation, Less-resourced languages, Statistical and Machine Learning Methods

1. Introduction

Even though misinformation in media has existed for a long time, the digital era has allowed for it to have a wider reach, as was seen during Brexit and the U.S. presidential elections from 2016 and 2020. This has been exacerbated during the COVID-19 pandemic amid the uncertainty and length of this event. During his speech at the Munich Security Conference 2020, Tedros Adhanom Ghebreyesus (2020), general-director of the World Health Organization (WHO), used the term "infodemic" to characterize this viral spread of misinformation in a parallel manner to the actual pandemic . This has had real world effects, such as anti-lockdown demonstrations (Wikipedia, 2022) and the rise of more wide-spread anti-vax movements (Baer, 2021). As with most countries, Sweden has been no stranger to these (Glad and Sundberg, 2021).

In this paper we work with news articles related to the COVID-19 pandemic in Swedish coming from reliable and unreliable sources. We use traditional and neural models to attempt to determine whether a given article comes from an reliable or an unreliable source. Given that our dataset is unbalanced, we also explore three different kinds of data augmentation (subsampling, backtranslation, and easy data augmentation) to attempt to solve the issues caused by this. The dataset we use was originally presented by Kokkinakis (2021) but to the best of our knowledge, it hasn't been used since its introduction. More information about the dataset itself can be found in section 3. On the other hand, we describe our data augmentation methods in section 3.1 and the models that we use for classification in section 4.

We observe that a logistic regression model with tf-idf representations performs the best at detecting previously unseen unreliable articles, while maintaining a good overall F1-score. This model outperforms BERT and other models such as LSTMs and SVMs. We also realize that easy data augmentation (EDA) tends to improve the results of the models in most cases. Due to these surprising results, we conclude that the current dataset is too small and the more complex models are probably overfitting the training data.

2. Background

Several tasks have arisen in the NLP community to try to study mis- and disinformation. In this section we will give a brief overview of them, as well as some of the methods that have been used to tackle these tasks. Even though we are studying news coming from unreliable sources, two closely related tasks exist.

In fake news detection, we try to determine whether a news article is intentionally deceptive. However, this requires us to know the intent of the person writing it, so is is often reduced to whether an article is truthful or not (Oshikawa et al., 2020). One way to do this is through simple classification of the titles and text of the articles has been attempted, both with traditional machine learning ((Shu et al., 2020) uses these as baselines) and with deep learning (see for example (Raza, 2021)) approaches. However, fact verification has also been successfully used for this task (Vijjali et al., 2020). Torabi Asr and Taboada (2018) note that it is important for fake news to have annotations the epistemological truth value of each article rather than on a source level. Because of that, we consider our task to be detection of news coming from unreliable sources rather than fake news detection, despite using similar methods and approaches.

[*]All authors had equal contribution. Correspondence to ricardo.munoz.sanchez@svenska.gu.se

Dataset description			
Split	Total size	Reliable	Unreliable
Train	1399	1259	140
Validation	298	269	29
Test	296	268	28

Table 1: Number of articles in each of the splits.

On the other hand, rumour mining focuses on unproven claims, often on social media. While the task can be seen as text classification (as task 8 of SemEval-2017 (Derczynski et al., 2017)), there have been people that have studied how rumours spread (see (Ma et al., 2018) for an example).
Other related tasks include detection of hyper-partisan news, stance, clickbait, satire, and propaganda.

3. Dataset

We use a dataset that was originally introduced in (Kokkinakis, 2021). This is a dataset that contains "news" related to the pandemic coming from different sources. These vary from official announcements from the government, to blogs that usually post articles about conspiracy theories. While the text of the articles is not freely available for download due to copyright, individual sentences can be accessed in a randomized order through Korp, the corpus search interface from SpråkbankenText[1] (Borin et al., 2012).
While the original dataset has more fine-grained labels, we grouped them into reliable sources and unreliable sources. A thorough list of the reliable and of the unreliable sources can be found in Tables 2 and 3, respectively, as well as a short description of most of them.
The dataset itself consists of the titles and the texts of each article, as well as other metadata such as the date it was published on and the URL of the article. Given that our dataset does not contain a thorough compilation of all COVID-19 articles that have been published in Swedish (neither by source nor by date), we just used the text and the titles for our classification task.
There are 1796 articles coming from reliable sources and 198 coming from unreliable ones in the dataset. In order to create a train/validation/test split, we randomly selected articles such that there was a similar proportion of official to unofficial articles in each split. The actual size of the splits can be seen in Table 1. We decided against the recommendation of Zhou et al. (2021) of not letting the same source appear in more than one split, as it would have meant that the validation and test sets would have consisted mainly of a single source due to the small size of our dataset, which could have also skewed our results.

3.1. Data Augmentation

As mentioned previously, our dataset has two important limitations: it is unbalanced and has a small number of examples of unreliable articles. In order to get around these limitations, we tried three different ways of data augmentation: subsampling of the reliable class and using a combination of backtranslation and EDA.

3.1.1. Subsampling

Given that there are about ten reliable sources for each unreliable one, one of the risks when training is that the model will decide that every article is reliable and still achieve a high accuracy. This poses a problem in alignment (Ortega et al., 2018), that is, when an AI model follows the rules that we set for it but doesn't do what we expect it to do. In other words, it learns how to "cheat" in order to get better results.

3.1.2. Backtranslation

Backtranslation (Edunov et al., 2018) is a simple augmentation method where synthetic data is generated by translating the original text into another language and then back into the original language. The intention is that the generated text retains the context of the original sentence but with different words and phrases. For this we use an API to access Google Translate[2].

3.1.3. Easy Data Augmentation (EDA)

Easy data augmentation (EDA) consists of four simple operations described in the original paper (Wei and Zou, 2019) as follows

- **Synonym replacement** Randomly select *n* words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.
- **random insertion:** Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Repeat this process n times.
- **random swap:** Randomly choose two words in the sentence and swap their positions. Repeat this process n times.
- **random deletion:** For each word in the sentences, randomly remove it from the sentence with probability p.

Parameters n and p determine the amount of noise to be added to the newly generated sentences. The original paper argues that, if we have a small dataset, using EDA on 50% of the training data can outperform the results from using the whole training data. We used the code from the authors' GitHub repository[3].

3.2. Training Sets

Using the augmentation techniques in the previous section, we obtain a total of four different training sets constructed from the original one. These new sets are as follows:

[1] https://spraakbanken.gu.se/korp/#?corpus=sv-covid-19

[2] https://github.com/lushan88a/google_trans_new

[3] https://github.com/jasonwei20/eda_nlp

Source	Description
Sveriges Radio	Swedish public service radio
Socialstyrelsen	The National Board of Health and Welfare
Myndigheten för Samhällskydd och Beredskap (msb)	The Agency for Civil Protection and Emergency Planning
Folkhälsomyndigheten	Swedish public health authority
Riksdagen	Swedish parliament
Regeringen	Swedish government
Krisinformation	Crisis information
Dagens Industri (di)	Liberal-conservative financial newspaper
Ehälsomyndigheten	Swedish e-Health agency
Göteborgsposten (gp)	Liberal, daily newspaper
Dagens Nyheter (dn)	Independently liberal newspaper
Vi	Monthly magazine on culture and society
Svenska Dagbladet (svd)	Independent moderate, daily newspaper
Hälsingborgs Dagblad (hd)	Largest Swedish daily newspaper outside of the metropolitan districts of Stockholm, Gothenburg, and Malmö
Västerbottenkuriren (vk.se, blogg.vk.se)	Swedish daily newspaper published in Västerbotten

Table 2: A list of the reliable sources, as well as a short description for most of them.

Source	Description
Anthropocene	'A politically independent, liberal forum for debate and opinion formation'
Det Goda Samhället	Online publication for which the financing takes place with the help of grants from private individuals and companies
Fria Tider	Immigration-critical online newspaper
Nyadagbladet	A Swedish online newspaper founded in 2012 which is nationalist, science-skeptical, and non-partisan.
Swebbtv	Swedish media channel, the channel describes itself as being politically independent and critical of Sweden's immigration policy
kavlaner.se	Anti vaccination campaign
humanismkunskap.org	(No description available)
sv.technocracy.news	Proponents of technocracy, tend to be very conspiratorial regarding COVID-19
frihetsportalen.se	'This site is produced by Mats Jangdal in Sweden and mainly in Swedish. Occasionally I publish in English. The site is devoted to topics like freedom, property rights and the UN climate fraud, also politics in general.'
static.bloggproffs.se	(No description available)
cornucopia.cornubot.se	'... The blog's ambition is to be wrong in everything [sic]. By writing about potential problems before they arise or worsen, we may be able to avoid them or reduce their consequences ...'
newsvoice.se	' ... NewsVoice does not shy away from exposing corruption and abuse of power and is therefore not politically correct ...'
trovetandeochvetenskap.se	Blog with the subheading: 'Only those who swim against the current reach the source'

Table 3: A list of the unreliable sources, as well as a short description for most of them.

1. The unchanged original training set.
2. For each unreliable data point in training set 1, one new data point is generated by backtranslation and five new data points through a combination of backtranslation and EDA.
3. A balanced training set extracted from the original one by subsampling the reliable articles.
4. Data augmentation as in training set 2 is performed to all data points from training set 3.

The original validation and test sets are used in their

original forms throughout the training and evaluation of all models.

4. Models for Text Classification

We compare several kinds of models to determine which one has the best performance.

4.1. Logistic Regression

In order to establish a baseline, we used a logistic regression model for binary classification. We use stemming and stopword removal to clean our text and then use tf-idf to obtain numerical features that are then fed to the logistic regression model.

4.2. Support Vector Machine (SVM)

Another traditional machine learning method we use was a support vector machine (SVM), as they tend to work well in classification tasks (Meyer et al., 2003). For this method we use the same preprocessing as with the logistic regression model. The only difference being that we feed the tf-idf features to an SVM with a linear kernel rather than to a logistic regression model.

4.3. biLSTM

One of our neural models was a bidirectional LSTM. These are neural networks that use two LSTM (Hochreiter and Schmidhuber, 1997) layers, one in each direction, and then concatenate the hidden states of each direction to feed them to a linear layer for classification.
For this model, we use only the first 300 tokens of each article in order to avoid disappearing gradients. We also use word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) in order to obtain intermediate representations of the text. More specifically, we use the Swedish embeddings trained on the CoNLL17 corpus[4] (Zeman et al., 2017) found at the NLPL word embeddings repository[5] (Fares et al., 2017).

4.4. BERT

The other neural model that we use is the Swedish verison of BERT released by the National Library of Sweden (Malmsten et al., 2020), available in the Hugging Face repository[6]. The first token of BERT's output, *[CLS]*, is then fed to a linear layer for classification.
In terms of specific implementation, we fine-tune the BERT model using our training data to obtain better representation of the text using this special token. We also use only the first 300 tokens of the text of each article to maintain consistency across the two neural models. Moreover, we use the BERT tokenizer in order to preprocess the text.

5. Experimental Results

Somewhat surprisingly, the logistic regression model outperformed all the others. Even though the one trained on the original training set fared poorly, when using EDA and subsampling the performance soars, achieving a F1-score of 0.759 on the unreliable articles and an overall F1-score of 0.866. This greatly outperforms the second best model, which is BERT using both EDA and subsampling with an F1-score of 0.709 for the unreliable for the unreliable articles and an overal F1-score of 0.837. The full results of our experiments can be seen in table 4 and are reported in terms of test set accuracy and F1-scores for the test set and for each class.
Regarding the traditional machine learning models, we can observe that with the logistic regression models any kind of augmentation improves the results. Meanwhile, EDA has a marked improvement for the SVM. Similarly, using subsampling improves both the overall F1-score and the F1-score for the unreliable class, even though we obtain a slightly worse F1-score for the reliable one.
With the LSTM models, we clearly note that subsampling leads to a worse perfomance of the models. Moreover, while we get mixed results with EDA, the F1-scores both overall and of the unreliable class are higher when using the unaltered training set. This is most likely due to having a small dataset to begin with, an issue made worse due to the data-hungry nature of LSTMs.
Finally, our BERT model performed the best when using EDA and subsampling.

6. Discussion

As mentioned before, we found it somewhat surprising that the best performing model was a variation of the baseline one. However, when looking at the representations we used, as well as how EDA works, it starts making more sense.
The idea of EDA is that we generate datapoints from random changes in the text. Even though in paper this is a good idea, it can have a noticeable impact on the more complex methods. For example, random swapping of words will wreck havoc in a sequential model such as LSTMs, while random swapping and insertion of synonyms can change the BERT models in unexpected ways. However, it is important to note that the BERT models that we used are pre-trained, so they can also better harness synonyms and similar changes.
On the other hand, tf-idf is a bag-of-words approach. This means that random insertions and swaps do not affect it at all. On the other hand, both backtranslation and synonym replacement should enhance the representations obtained through this method. Despite this, we wouldn't expect such improved results when compared to the neural network approaches.
It is important to note that the original EDA paper uses data from Twitter, which is limited to 140 characters. Even if we cropped the text of the articles to this length, the differences in information density would probably mean that the results would probably not be as good

[4] http://universaldependencies.org/conll17/
[5] http://vectors.nlpl.eu/repository/
[6] https://huggingface.co/KB/bert-base-swedish-cased

Model	Balanced	EDA	Acc.	F1-score		
				overall	reliable	unreliable
LogReg	No	No	0.922	0.631	0.959	0.303
LogReg	No	Yes	0.943	0.767	0.969	0.564
LogReg	Yes	No	0.926	0.822	0.958	0.686
LogReg	**Yes**	**Yes**	**0.953**	**0.866**	**0.974**	**0.759**
SVM	No	No	0.949	0.803	0.973	0.634
SVM	No	Yes	**0.956**	**0.837**	**0.976**	**0.698**
SVM	Yes	No	0.929	0.828	0.960	0.696
SVM	Yes	Yes	0.929	0.828	0.960	0.696
LSTM	No	No	**0.943**	**0.824**	**0.968**	**0.679**
LSTM	No	Yes	0.939	0.810	0.967	0.654
LSTM	Yes	No	0.912	0.772	0.951	0.594
LSTM	Yes	Yes	0.885	0.731	0.935	0.528
LSTM + sent.	No	No	0.905	**0.755**	0.947	**0.563**
LSTM + sent.	No	Yes	**0.912**	0.752	**0.951**	0.552
LSTM + sent.	Yes	No	0.892	0.712	0.940	0.484
LSTM + sent.	Yes	Yes	0.889	0.715	0.937	0.492
BERT	No	No	0.939	0.746	0.968	0.679
BERT	No	Yes	0.945	0.785	**0.971**	0.600
BERT	Yes	Yes	**0.952**	**0.837**	0.966	**0.709**

Table 4: Result from evaluating the models on the test set. We report test set accuracy and F1-score for the full test set as well as for each class. For each kind of model we bolden the best result for the scores we report.

as with text that is naturally shorter. An interesting follow-up would be to test the effectiveness of EDA on datasets with lengthier tests to see whether there is any improvement in the results. It would also be interesting to change the implementation of EDA such that it is applied to each sentence in the input text independently rather than to the full input text.

Another possible follow-up experiment would be to use linguistic features rather than whole-document representations. This has proven to be a successful approach both with larger datasets (Horne et al., 2019) as well as with smaller ones (Pérez-Rosas et al., 2018). Moreover, a deeper error analysis could be done on these models.

7. Conclusions

Even though most studies on misinformation have focused on the English language, it is important to also study what happens in other languages. Different cultures react differently to global events and it is important to recognize that.

One of the main challenges we faced was a lack of annotated data on which to train our models. As far as we know, the only existing dataset so far is the one we used, introduced by Kokkinakis (2021). Even though news from unreliable sources are overtly abundant on social media and the rest of the web, it can be expensive or time-consuming to identify and label them. Moreover, Juneström (2021) note that the best known fact-checking website for Swedish news is no longer updated as of 2019. This makes it harder to gather fact-checked data on the COVID-19 pandemic in Sweden, both if we were to use annotations at source or at article levels.

In order to gather our own data, we would require the help from health and disinformation experts that are fluent in the language

We also realized that the use of EDA lead to surprisingly good results when using simple machine learning methods, especially when compared to deep learning approaches. As noted during the discussion, this might be due to the nature of the representations used for these models. These greater gains when comparing the two kinds of approaches might point out at EDA working better with either shorter texts or with non-serialized data.

It is only through assured access to the most updated information about the COVID-19 pandemic that we will be able to go through it sooner rather than later. The increasing spread of misinformation render healthcare measures less effective, allowing the virus to spread more widely.

8. Bibliographical References

Adhanom Ghebreyesus, T. (2020). Munich security conference. https://www.who.int/director-general/speeches/detail/munich-security-conference. World Health Organization. [Online; accessed 15-January-2022].

Baer, S. K. (2021). Thousands of anti-vax protesters marched in europe even though COVID deaths are rising. https://www.buzzfeednews.com/article/skbaer/antivax-europe-covid-mandates. BuzzFeed News. [Online; accessed 15-January-2022].

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., and Zubiaga, A. (2017). SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Glad, E. and Sundberg, S. (2021). Tusenmannamarschen och coronaförnekandet. https://sverigesradio.se/avsnitt/1701155. In *P3 Nyheter Dokumentär*. Sveriges Radio [Online; accessed 16-January-2022].

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780.

Horne, B. D., Nørregaard, J., and Adali, S. (2019). Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology*, 11(1):7:1–7:23.

Juneström, A. (2021). Discourses of fact-checking in swedish news media. *Journal of Documentation*, 78(7):125–140. Publisher: Emerald Publishing Limited.

Ma, J., Gao, W., and Wong, K.-F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Meyer, D., Leisch, F., and Hornik, K. (2003). The support vector machine under test. In *Neurocomputing*, volume 55, pages 169–186.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *Arxiv preprint*. arXiv:1301.3781 [cs].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc.

Ortega, P. A., Maini, V., and DeepMind Safety Team. (2018). Building safe artificial intelligence: specification, robustness, and assurance. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1. Medium [Online; accessed 14-January-2022].

Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

Raza, S. (2021). Automatic fake news detection in political platforms - a transformer-based approach. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. In *Big Data*, volume 8, pages 171–188. Mary Ann Liebert, Inc.

Torabi Asr, F. and Taboada, M. (2018). The data challenge in misinformation detection: Source reputation vs. content veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 10–15. Association for Computational Linguistics.

Vijjali, R., Potluri, P., Kumar, S., and Teki, S. (2020). Two stage transformer model for COVID-19 fake news detection and fact checking. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. International Committee on Computational Linguistics (ICCL).

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388. Association for Computational Linguistics.

Wikipedia. (2022). Protests over responses to the covid-19 pandemic — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Protests_over_responses_to_the_COVID-19_pandemic. [Online; accessed 14-January-2022].

Zhou, X., Elfardy, H., Christodoulopoulos, C., Butler, T., and Bansal, M. (2021). Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

9. Language Resource References

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicabil-

ity: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet.

Kokkinakis, D. (2021). Insights on a swedish covid-19 corpus. In Monica Monachini et al., editors, *Proceedings of the CLARIN Annual Conference 2021*, pages 31–34.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden – making a swedish BERT. *Arxiv preprint*. arXiv:2007.01658 [cs].

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, C., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.