

Yseop at SemEval-2020 Task 5: Cascaded BERT Language Model for Counterfactual Statement Analysis

Hanna Abi Akl

Yseop

habi-akl@yseop.com

Dominique Mariko

Yseop

dmariko@yseop.com

Estelle Labidurie

Yseop

elabidurie@yseop.com

Abstract

In this paper, we explore strategies to detect and evaluate counterfactual sentences. We describe our system for SemEval-2020 Task 5: Modeling Causal Reasoning in Language: Detecting Counterfactuals. We use a BERT base model for the classification task and build a hybrid BERT Multi-Layer Perceptron system to handle the sequence identification task. Our experiments show that while introducing syntactic and semantic features does little in improving the system in the classification task, using these types of features as cascaded linear inputs to fine-tune the sequence-delimiting ability of the model ensures it outperforms other similar-purpose complex systems like BiLSTM-CRF in the second task. Our system achieves an F1 score of 85.00% in Task 1 and 83.90% in Task 2.

1 Introduction

A counterfactual can be defined as something that is contrary to the truth or that did not actually occur. It refers to an event that did not or cannot happen, as well as the possible consequences if it had happened. In the sentence "*If dogs had no ears, they could not hear*", the statement "*if dogs had no ears*" is an example of a counterfactual because dogs *do* have ears. Task 5 of SemEval-2020 (Yang et al., 2020) focuses on identifying these specific sentence types amongst sentences delivering close semantic similarities. This implies understanding and disambiguating the causal link between two sentence fragments.

We approached this task as an opportunity to test the effectiveness of disambiguating at a grammatical level against traditional baseline systems. This paper describes a parallel approach in deriving some meaning from text to leverage the influence of context and relevance of structure in recognizing counterfactual statements. Specifically, we will explore how many expressions of such statements a mapping of grammatical types can cover before falling short to high-performing models, most notably BERT (Devlin et al., 2018), which performed well on both tasks with an F1 score of 85.00% on Task 1 and 83.90 % on Task 2.

2 Related Work

Although the task of detecting counterfactuals is relatively new, (Son et al., 2017) proposes using modal logic to form rule-based determination methods from social media posts. These methods are supplemented by a statistical classifier (Linear SVM) that is retrained to tackle more challenging counterfactual forms.

Previous work on causal identification by (Levin and Hovav, 1994) studied the contribution of verbs in the determination of causal relations. By analyzing closely different formulations, they concluded that similarities in meaning can be derived from very different syntactic structures

As for causality relation extraction, different deep learning systems built on the success of neural-based models have been proposed. The linguistically informed CNN model (Dasgupta et al., 2018) leverages the use of word embeddings and other linguistic features to detect causal patterns and outperforms rule-based classification. (Liang et al., 2019) introduces a multi-level causal detector that makes use of multi-head self-attention to capture semantic features at word level and infer causality at segment

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

level. This engineered system has rivaled state-of-the-art models in terms of performance and thorough understanding of complex semantic information such as discourse relations and transitivity rules. Finally, (Li et al., 2019) presents a self-attentive BiLSTM-CRF model that makes use of transfer learning to overcome the problem of data insufficiency and extract causal relations in natural language text. This solution transfers a trained embedding from a large corpus and uses the causality tagging scheme to identify dependencies between cause and effect. Experimental results prove the effectiveness of this model, but its major limitation is the insufficiency of high-quality annotated data to learn from.

3 Dataset

The datasets used are those provided by the shared task organisers. The data is described in (Yang et al., 2020). As per official task instructions, no additional data was used.

4 Task 1: Classification Problem

Task 1 is a classification problem which aims at recognising text sections as either *counterfactual* or *not*. Counterfactual sections are labeled **1** and non-counterfactual sections are labeled **0**.

The proposed baseline by the task organisers is a Support Vector Machine (SVM) classifier. After a preliminary phase based on exploring grammatical features engineering, we discuss two different approaches and compare them to the baseline: a classical approach using popular machine learning classifiers with semantic and additional grammatical features, a combination that has shown to perform well on information retrieval tasks relying on text understanding (Dai and Callan, 2019), and a deep learning approach using a BERT linear classifier. The objective of this competitive approach is to determine whether a supplement of linguistic features can be sufficient to correctly recognise counterfactual structures as opposed to running heavier models integrating broader contextual knowledge.

4.1 Proposed Method

4.1.1 Exploration

As the data set contains sentences with close semantics, we run a first exploratory analysis to try to establish disambiguation patterns. A manual linguistic analysis of the training dataset shows that verbs are key elements for the detection of counterfactuals. (Son et al., 2017) identify 7 characteristics related to counterfactuals, all depending on a verb feature. Verb tenses in particular are key to disambiguation, so we build a generalist grammar based on our observation to include only the verbal forms that seem relevant for our analysis. We categorize *could*, *would* and *should* as modals so they wouldn't be identified as verbs in the preterit tense. Finally, we add a pattern to disambiguate the category of *wish* (*verb*) and *wishes* (*noun*), and identify conditional statements (*If*). We eventually retain 4 main features among the 7 described by (Son et al., 2017): Verb inversion, Modal chunks, Wish verbs and If clauses. We then generate combinations of tokens based on their grammatical category (i.e., verb, pronoun) as well as their linguistic properties (i.e., tense). The full grammar is provided in Appendix A.

However, some cases resist disambiguation, especially the sentences containing *could/would/should have* structures, which are very common in English and in many cases do not imply a counterfactual. Moreover, the context of these structures can vary which makes them even more complex to disambiguate. Eventually, the ambiguity between adjectives and past participles can raise many mistakes in the identification of counterfactuals when using grammars. Adjectives constructed with the suffix *-ed* are also recognised as verbs in preterit tense or past participles by the grammar (i.e., *He wasn't prepared*).

By deterministically applying the grammar on the training set, we obtain an output of 2833 sentences recognised as counterfactuals (on a total of 13000). However, only 593 of them are labeled "1" in the training set. A manual analysis of the output reveals that many statements are recognised as counterfactuals because of remaining ambiguities.

We turn to classical linear and non-linear machine learning methods to explore if combining these patterns carries any prior knowledge on disambiguation. We turn all 4 features described in Appendix A into binary variables, flagging the existence (True) or non-existence (False) of the category and associated patterns in the sentence. An example of the data transformation is provided in Appendix B. We end up

with a table of 31 features that are evaluated for predicting a counterfactual statement with the following learning methods and parameters (using scikit-learn classifiers¹):

SVM: *gamma*: scale

LOGIT: *l1 ratio*: 0.5, *penalty*: elasticnet, *solver*: saga

KNN: *n neighbors*: 3

CART: *criterion*: gini

all other parameters: sklearn defaults

Classifier	True value	Precision	Recall	F1 score
SVM	0	0.89	1.00	0.94
	1	0.70	0.05	0.10
LOGIT	0	0.89	1.00	0.94
	1	0.71	0.06	0.11
KNN	0	0.90	0.98	0.94
	1	0.39	0.12	0.18
CART	0	0.89	1.00	0.94
	1	0.68	0.06	0.11

Table 1: Task 1 - Exploratory Tests

The results provided in Table 1, especially the very poor recall metrics, prove our initial set of features are incomplete and noisy. We then decide for a more holistic approach, resuming any prior knowledge on the syntax to evaluate if existing latent variables can possibly lie in the raw morpho-syntax (POS tags) and semantic of counterfactual statements.

4.1.2 Operating methods

We first try a classical approach, supplementing classifiers with word vectorisation (Bag Of Words (BOW), TF-IDF, Word2Vec and BERT vectors) and morpho-syntactic features (POS tagging). The details of these transformations are provided in Appendix C. The classifiers evaluated are: SVM, K-Nearest-Neighbor (KNN), Multinomial Naive Bayes (NB), Decision Trees (CART), Random Forest (RF) and Multi-Layer Perceptron (MLP). Finally, we add a cross validation method with a stratified fold of 3 to our models. As shown in Table 2, the best learning models seem to have strong linearisation capabilities (MLP and SVM), which is why we also tested a deep learning model which has proved similar behaviour.

For the deep learning approach, we use a model inspired from the BERT model that achieved the best result in a similar text classification task in the NLP4IF-2019 Shared Task (Da San Martino et al., 2019). This is a BERT model with a linear layer on top. This system has proven to pay special attention to adjectives and verbs, two grammatical categories that can play a role in identifying counterfactual statements (Levin and Hovav, 1994). For our implementation, we use the *BertForSequenceClassification* model from the Transformers² library, a sentence-tokenized version of the BERT Uncased model with 12 Transformer layers and 110 million parameters.

4.2 Results

The results are based on the official training set provided by the organisers. The dataset contains 13000 lines split in the following: 40% training sample, 30% validation sample and 30% test sample.

For the classical approach, we retain the best result for each classifier out of all the possible combinations of classifiers and text processing and compare with the baseline provided for the task and the deep learning model. The +/- scores are the averaged measures from the 3-fold cross-validation results for all models except the BERT Linear Model, whose results are provided with the scikit-learn default parameters.

¹<https://github.com/scikit-learn/scikit-learn>

²<https://github.com/huggingface/transformers>

Classifier	Processing	Precision	Recall	F1 score
Baseline SVM	TF-IDF	72.72 %	8.73 %	15.59 %
MLP	BERT Sentence Version	80 +/- 1 %	80 +/- 1 %	80 +/- 1 %
SVM	BERT Sentence Version	79 +/- 0 %	81 +/- 1 %	80 +/- 0 %
CART	TF-IDF with Stop Words	69 +/- 1 %	70 +/- 1 %	70 +/- 1 %
NB	BERT Sentence Version	64 +/- 1 %	75 +/- 2 %	66 +/- 1 %
KNN	TF-IDF with Stop Words	71 +/- 1 %	64 +/- 1 %	66 +/- 1 %
RF	TF-IDF with Stop Words	91 +/- 1 %	61 +/- 1 %	65 +/- 1 %
BERT Linear	BERT Uncased	87 +/- 1 %	83 +/- 1 %	86 +/- 1 %

Table 2: Task 1 - Benchmark Tests

The blind test set consists of 7000 unlabeled lines of text. Our best model, the BERT Linear, achieves an F1 score of **85.00%**, a Precision score of **84.20%** and a Recall score of **85.90%** on this set.

5 Task 2: Sequence Delimitation

Task 2 is a sequence delimitation problem. The text sections are similar to the ones labeled "1" in the Task 1 dataset. The purpose of this task is to extract, in a text section identified as counterfactual, the sub-strings identifying the antecedent and the consequent elements (Yang et al., 2020). We use the following split sampling for the training data (3551 individuals): **1740 sentences** for the training sample, **746 sentences** for the validation sample and **1065 sentences** for the test sample.

5.1 Proposed Method

Our approach for this task is also comparative and evaluates two deep learning models. It consists in testing whether we can supply enough linguistic knowledge to determine all causal forms in counterfactual statements to challenge the breadth and depth of a model that leverages the power of BERT.

For both systems, we designed each provided statement in the training set as chunks. These chunks are composed of tokens labelled as *C* when the token belongs to a sub-string *Consequent* and *A* when the token belongs to a sub-string *Antecedent*. Tokens belonging to neither are marked *I*. These transformations of the target sequences, additional transformations and their input levels are described in Appendix D (D.0.1 and D.0.2). These target features are identified as **CHUNKS** in our results Table 3 and 4. Since we can tackle this task through token classification, we build a first Sequence Extractor system using the BERT model from task 1 and a wrapper³ to concatenate a Multi-Layer Perceptron classifier layer on top of it, as demonstrated in (Dai and Callan, 2019).

The second system is inspired by discriminative models and Conditional Random Fields (CRF) in particular. We design the system as a Named Entity tagger that takes sentence tokens and a set of morpho-syntactic token-level features as input and predicts the target class of each token.

We enhance the discriminative properties of the CRF by working on some additional layers and modeling a Deep Learning CRF. We experiment with linguistic embeddings, features and regularisation methods as enhancements for the final BiLSTM-CRF Neural Model. A high-level diagram of the system architecture is presented in Appendix E.

The full configurations for both systems are detailed in Appendix D (D.0.3 and D.0.4).

5.2 Results

Since we are working with a complex model with high tuning capabilities, the most relevant results for the BiLSTM-CRF system on the training dataset are retained and presented in Table 4. The BERT-MLP model results are shown in Table 5. Here again the +/- measures are averaged from the 3-fold cross validation scores.

The results show that generalization might be a problem for both systems as discussed in (Zeyer et al., 2019). The gap in performance between the BiLSTM-CRF and BERT-MLP is due to the superior

³<https://github.com/charles9n/bert-sklearn>

Model	Embedding	Features	Regularisation	Precision	Recall	F1 score
BiLSTM-CRF	FastText + C2IDX	POS + CHUNKS	B	77 +/- 1 %	76 +/- 2 %	77 +/- 2 %
BiLSTM-CRF	Stacked + C2IDX	POS + CHUNKS + SG	C	76 +/- 1 %	75 +/- 1 %	77 +/- 2 %
BiLSTM-CRF	GloVe + C2IDX	POS + CHUNKS	C	78 +/- 1 %	78 +/- 1 %	79 +/- 1 %
BiLSTM-CRF	GloVe + C2IDX	POS + CHUNKS + SG	A	76 +/- 1 %	75 +/- 1 %	77 +/- 1 %
BiLSTM-CRF	C2IDX	POS + CHUNKS + BERTvec + SG	B	76 +/- 1 %	76 +/- 1 %	77 +/- 1 %
BiLSTM-CRF	Stacked + C2IDX	POS + CHUNKS	C	76 +/- 1 %	76 +/- 1 %	76 +/- 1 %

Table 3: Task 2 - BiLSTM-CRF Benchmark Tests

Model	Features	Regularisation	Precision	Recall	F1 score
BERT-MLP	POS + CHUNKS	A	81 +/- 1 %	82 +/- 1 %	81 +/- 1 %
BERT-MLP	CHUNKS + SG	A	82 +/- 1 %	83 +/- 1 %	82 +/- 1 %
BERT-MLP	CHUNKS	A	85 +/- 1 %	85 +/- 1 %	85 +/- 1 %
BERT-MLP	CHUNKS + BIO NER	A	83 +/- 1 %	84 +/- 1 %	83 +/- 1 %
BERT-MLP	CHUNKS + SG + BIO NER	A	83 +/- 1 %	83 +/- 1 %	83 +/- 1 %
BERT-MLP	POS + CHUNKS + SG + BIO NER	A	84 +/- 1 %	84 +/- 1 %	84 +/- 1 %

Table 4: Task 2 - BERT-MLP Benchmark Tests

performance of Transformer models over LSTM encoders. When coupling the BERT-MLP model with one set of features, we observe steep learning, but cascading multiple feature layers results in a smoother and more accurate learning process. Ultimately, since the accuracy of the other linguistic features is closely tied to the library that generates them (particularly the POS tagger), they do not add any enhancement to the chunking feature. On the blind test dataset, our best model achieves a Precision score of **82.30%**, a Recall score of **88.80%**, an F1 score of **83.90%** and an Exact Match of **25.90%**. The Exact Match measure is severely affected by chunking limits which explains why the score for this metric is particularly poor.

6 Conclusion

In this paper, we presented our experiments for identifying counterfactual statements. Modeling features derived from a linguistic analysis such as specific grammar structures for counterfactual statements and coupling them with established machine learning or deep learning models did not perform as well as context-learning models, as our hybrid BERT-MLP solution outperforms even complex combinations of deep learners and displays a better level of understanding and handling challenging counterfactual forms. Future work could explore the impact of graph knowledge in accommodating systems and rendering them more perceptive of implicit and ambiguous textual meanings.

References

- Ashgar, Nabiha. 2016. *Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey*. Arxiv 2016.
- Cheng, Gaofeng and Peddinti, Vijayaditya and Povey, Daniel and Manohar, Vimal and Khudanpur, Sanjeev and Yan, Yonghong. 2017. *An Exploration of Dropout with LSTMs*. Interspeech 2017.
- Da San Martino, Giovanni and Barron-Cedeno, Alberto and Nakov, Preslav. 2019. *Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection*. Arxiv 2019.
- Dai, Zhuyun and Callan, Jamie. 2019. *Deeper Text Understanding for IR with Contextual Neural Language Modeling*. Arxiv 2019.
- Dasgupta, Tirthankar and Saha, Rupsa and Dey, Lipika and Naskar, Abir. 2018. *Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks*. Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pages 306–316.
- Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. CoRR, abs/1810.04805.

- Dunietz, Jesse. 2018. *Annotating and Automatically Tagging Constructions of Causal Language*. Carnegie Mellon University.
- Dunietz, Jesse and Levin, Lori and Carbonell, Jaime. 2017. *The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations*. Proceedings of the 11th Linguistic Annotation Workshop, ACL Anthology 2017.
- Levin, Beth and Hovav, Malka R.. 1994. *A Preliminary Analysis of Causative Verbs in English*. *Lingua* 92, 35-77.
- Li, Zhaoning and Li, Qi and Zou, Xiaotian and Ren, Jiangtao. 2019. *Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings*. Arxiv 2019.
- Liang, Shining and Zuo, Wanli and Shi, Zhenkun and Wang, Sen. 2019. *A Multi-level Neural Network for Implicit Causality Detection in Web Texts*. Arxiv 2019.
- Ma, Xuezhe and Hovy, Eduard H.. 2016. *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*. Arxiv 2016.
- Nazaruka, Erika. 2019. *Identification of Causal Dependencies by using Natural Language Processing: A Survey*. ENASE 2019.
- Reimers, Nils and Gurevych, Iryna. 2017. *Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks*. Arxiv 2017.
- Son, Youngseo and Buffone, Anneke and Raso, Joe and Larche, Allegra and Janocko, Anthony and Zembroski, Kevin and Schwartz, H. and Ungar, Lyle. 2017. *Recognizing Counterfactual Thinking in Social Media Texts*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 654–658.
- Yang, Xiaoyu and Obadinma, Stephen and Zhao, Huasha and Zhang, Qiong and Matwin, Stan and Zhu, Xiaodan. 2020. *SemEval-2020 Task 5: Counterfactual Recognition*. Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020).
- Zeyer, Albert and Bahar, Parnia and Irie, Kazuki and Schluter, Ralf and Ney, Hermann. 2019. *A Comparison of Transformer and LSTM Encoder Decoder Models for ASR*. Aachen University.

Appendix A. List of disambiguation features and associated grammars.

We qualify the 4 identified features using structures which are sequences of Part-Of-Speech tags. The Part-of-Speech tags follow the Penn Treebank convention and are extracted using Stanford CoreNLP⁴. Features are listed along with their corresponding sequences and example sentences. The grammar is implemented with NLTK.

⁴<https://stanfordnlp.github.io/CoreNLP/index.html>

Feature	Pattern	Example
Verb inversion	<VBD> <PRP> <VBN> <VBD> <PRP> <RP> <VBN> <VBD> <PRP> <PRP> <VBD> <VBD> <PRP> <RP> <VBD>	(Had/VBD you/PRP provided/VBN) (Had/VBD you/PRP not/RP provided/VBN) (Had/VBD she/PRP asked/VBD) (Had/VBD she/PRP not/RP asked/VBD)
Modal chunks	<MD> <VB> <VBN> <MD> <VB> <VB> <MD> <RP> <VB> <VBD> <MD> <VB> <MD> <RP> <VB> <MD> <VB> <VBD> <MD> <VBP*> <VBN> <MD> <RP> <VBP*> <VBN> <MD> <RP> <VB> <VBN> <MD> <VBP*> <VBN> <VBN> <MD> <RP> <VBP*> <VBN> <VBN>	(Would/MD have/VB provided/VBN) (Would/MD go/VB check/VB) (Would/MD not/RP have/VB died/VBD) (Would/MD be/VB) (Should/MD not/RP have/VB) (Would/MD have/VB asked/VBD) (Would/MD have/VBP bought/VBN) (Would/MD not/RP have/VBP bought/VBN) (Would/MD not/RP have/VB provided/VBN) (Would/MD have/VBP been/VBN bought/VBN) (Would/MD not/RP have/VBP been/VBN bought/VBN)
Wish verbs	<PRP> <VBP> <PRP*> <VBD> <PRP> <VBP> <IN> <PRP> <VBD> <PRP> <VBP> <IN> <PRP> <VBD> <VBN> <MD> <VB> <IN> <PRP> <VBD> <VBN> <MD> <VB> <IN> <NN*> <VBD> <VBN> <MD> <VB> <IN> <DT> <NN*> <VBD> <VBN> <PRP> <VBP> <PRP> <VBD> <PRP> <DT> <VBP> <IN> <PRP> <MD> <VB> <VBN> <PRP> <DT> <VBP> <PRP> <MD> <VB> <VBN> <PRP> <VBP> <DT> <VBD> <PRP> <RB> <VBP> <PRP> <MD> <VB> <PRP> <RB> <VBP> <PRP> <MD> <VBD> <PRP> <RB> <VBP> <PRP> <MD> <VB> <PRP> <VBP> <PRP*> <MD> <VB> <PRP> <VBP> <PRP*> <NN*> <MD> <VB> <PRP> <VBP> <PRP*> <NN*> <MD> <VB> <PRP> <VBP> <PRP*> <NN*> <VBD> <VBN> <PRP> <VBP> <PRP*> <VBD> <VBN> <PRP> <VBP> <NN*> <VBD> <PRP> <VBP> <DT> <NN*> <VBD> <PRP> <VBP> <NN*> <VBD> <VBN> <PRP> <VBP> <NN*> <VBD> <PRP> <VBP> <PRP> <VBD> <RP> <PRP> <VBP> <DT> <VBD> <JJ>	(I/PRP wish/VBP I/PRP held/VBD) (I/PRP wish/VBP that/IN you/PRP had/VBD) (I/PRP wish/VBP that/IN I/PRP had/VBD bought/VBN) (May/MD wish/VB that/IN he/PRP had/VBD kept/VBN) (May/MD wish/VB that/IN households/NNS had/VBD kept/VBN) (May/MD wish/VB that/IN these/DT households/NNS had/VBD kept/VBN) (I/PRP wish/VBP you/PRP had/VBD) (We/PRP both/DT wish/VBP that/IN we/PRP could/MD have/VB helped/VBN) (We/PRP both/DT wish/VBP we/PRP could/MD have/VB helped/VBN) (I/PRP wish/VBP this/DT were/VBD) (I/PRP just/RB wish/VBP we/PRP could/MD have/VB) (I/PRP just/RB wish/VBP we/PRP had/VBD) (I/PRP just/RB wish/VBP we/PRP could/MD have/VB) (I/PRP wish/VBP we/PRP could/MD have/VB) (I/PRP wish/VBP my/PRP\$ parents/NNS could/MD have/VB) (I/PRP wish/VBP institutions/NNS could/MD have/VB) (I/PRP wish/VBP my/PRP\$ parents/NNS had/VBD known/VBN) (I/PRP wish/VBP they/PRP had/VBD known/VBN) (I/PRP wish/VBP articles/NNS addressed/VBD) (I/PRP wish/VBP this/DT article/NN addressed/VBD) (I/PRP wish/VBP President/NNP Obama/NNP had/VBD succeeded/VBN) (I/PRP wish/VBP President/NNP Obama/NNP had/VBD) (You/PRP wish/VBP you/PRP had/VBD n't/RP) (I/PRP wish/VBP this/DT were/VBD true/JJ)
If clauses	<IN> <PRP> <MD> <VB> <IN> <PRP> <VBD> <JJ> <IN> <EX> <VBD> <IN> <NN*> <VBD> <VBP> <PRP> <VBN> <IN> <RP> <IN> <DT> <IN> <DT> <NN*> <MD> <VB> <IN> <DT> <JJ*> <NN*> <MD> <VB> <IN> <DT> <NN*> <VBD> <IN> <DT> <NN*> <VBD>	(If/IN I/PRP could/MD go/VB) (If/IN it/PRP were/VBD true/JJ) (If/IN there/EX were/VBD) (If/IN money/NN was/VBD) (Have/VBP you/PRP been/VBN) (If/IN not/RP for/IN that/DT) (If/IN the/DT boy/NN could/MD go/VB) (If/IN the/DT young/JJ boy/NN could/MD go/VB) (If/IN the/DT boy/NN was/VBD) (If/IN his/DT money/NN was/VBD)

Table 5: Task 1 - Exhaustive feature grammar

Appendix B. Feature engineering from Appendix A grammar.

We transform the different representations of the 4 grammatical features listed in Appendix A into categorical variables. Table 6 illustrates an example of one feature representation applied to input sentences.

sentenceID	gold_label	IF _{IN DT NN VBD}	IF _{IN DT NNP VBD}	IF _{IN DT NPS VBD}	IF _{IN DT NNS VBD}	IF _{IN EX VBD}	...
100000	1	0	0	0	1	0	...
100001	1	0	0	0	0	1	...

Table 6: Task 1 - Feature engineering examples for If clauses

Appendix C. Data transformations for Task 1.

We perform a two-level processing on the input: vectorial and morpho-syntactic. The vectorisation methods used are: Bag Of Words (BOW), TF-IDF, Word2Vec and BERT vectors. We perform a series of operations on the raw input text like removing numbers, punctuation and stop words, replacing negative contraction verbs with their complete forms (i.e., *won't*), splitting compound forms (i.e., *state-of-the-art*)

and transforming text to lowercase. We also replace multiple white spaces with a single space. Examples of these transformations are provided in Table 7. For TF-IDF and Word2Vec, we experiment with and without stop word removal. We use the NLTK⁵ stop words set and increase it with contraction patterns like *'re* or *'m*. For the BERT and Word2Vec vectors, we refrain from applying these cleaning operations to maintain more semantic freedom.

At the end of this phase, we generate a list of all the unique words in the training data called the vocabulary.

For the morpho-syntactic phase, we apply POS-based stemming and lemmatization for the BOW and TF-IDF embeddings. We also remove words with frequency less than 5 for these embeddings. This effectively decreases the dimensions of BOW and TF-IDF vectors.

Our final feature list consists of 2000 features that are the unique lemmatized vocabulary words and word groups curated from the input text. However, despite their simplicity and low time complexity, BOW and TF-IDF have two major drawbacks. First, as the size of the data and the number of unique words in the training text increase, the length of vectors becomes much larger. Moreover, in these two approaches, only words and their repetitions are important and the order of the words in the text is not considered in the model. This is why we also consider both Word2Vec and BERT embedding approaches in our experiments.

Transformation	Example
No transformation	If the lawsuit can be another means of focusing attention on these fundamental issues, then optimistically the lawsuit can provide a larger benefit.
Cleaning	if the lawsuit can be another means of focusing attention on these fundamental issues then optimistically the lawsuit can provide larger benefit
Cleaning + Normalization	If the lawsuit can be anoth mean of focu attent on these fundament issu , then optimist the lawsuit can provide a larg benefit .
Cleaning (Word2Vec + No Stop Words)	lawsuit another means focusing attention fundamental issues optimistically lawsuit provide larger benefit
Cleaning (BERT)	if the lawsuit can be another means of focusing attention on these fundamental issues, then optimistically the lawsuit can provide a larger benefit.

Table 7: Task 1 - Text cleaning examples

Appendix D. Task 2 systems configuration.

D.0.1. Features.

We apply the same feature sets to our two systems. The **bold** mentions refer to their description in the results Table 3 and Table 4.

For the BERT-MLP system, the features are introduced as additional input layers and calculated in the embedding layer. For the BiLSTM-CRF system, these features are declared in the CRF layer.

For **POS** tags, we use the Stanford⁶ Part-Of-Speech Tagger to determine the role of each word in the discourse.

For chunking the target sequences (**CHUNKS**), we tokenize each sequences and label each token of the chunk with its segment label (i.e., *A* for Antecedent and *C* for Consequent).

In order to generate BERT vector features (**BERTvec**), we use the BERT-as-service⁷ library (version 1.10.0) as a sentence-encoder to map variable-length sentences to fixed-length feature embeddings.

We also experiment with Syntactic Grammars (**SG**) by using the Stanford⁸ Parser to generate syntactic dependency relations between words.

Since our task requires labeling tokens, we could use a feature that references structure. As the results from (Reimers, 2017) show, the BIO tagging scheme performs consistently well for this type of task, which makes it the ideal candidate to add robustness to our feature set. Using the Stanford⁹ Named Entity Recognizer, which is trained on BIO entity tagged data, we generate the BIO representation of sentence entities, or BIO Named Entity Recognition (**BIO NER**) tags.

⁵<https://github.com/nltk/nltk>

⁶<https://nlp.stanford.edu/software/tagger.html>

⁷<https://github.com/hanxiao/bert-as-service>

⁸<https://nlp.stanford.edu/software/lex-parser.html>

⁹<https://nlp.stanford.edu/software/CRF-NER.html>

D.0.2. Embeddings.

Embeddings are computed from convolution of the different D.0.1 features (plus additional embeddings for BiLSTM-CRF) stacked in the input layer. The **bold** mentions refer to their description in the results Table 3 and Table 4.

For the **BERT-MLP** system, we add the cascaded layers of token-features referenced in section D.0.1 to the existing embeddings architecture . Our final embeddings layer is a concatenation of the 3 BERT embeddings (positional, segment, token) and the generated feature layers.

For feeding the BiLSTM-CRF system, we evaluate several embedding methods .

The first is Pre-trained Word Embeddings. This widely used technique helps tackle the problem of generalising unseen words, since word embeddings are good at capturing general syntactic as well as semantic properties of words (Reimers, 2017). For our experiments, we focus on two different approaches: the **GloVe** embeddings trained on Common Crawl (about 840 billion tokens) and the **FastText** approach trained on Common Crawl (600 billion tokens) which also extracts subword information.

As per the recommended approach of (Ma and Hovy, 2016), we also compound Character Embeddings (**C2IDX**) to take into account character-level representation of words as an additional layer to the word embedding layer.

Finally, we also consider Stacked Embeddings (**Stacked**) (i.e., a combination of existing embedding techniques designed to act in succession in a single pipeline to generate a refined embedding for our input). We use the Flair¹⁰ library (version 0.4.5) to achieve this. Our embeddings pipeline is composed of: a GloVe model, a Flair-forward model, a Flair-backward model and a BERT embedding model. By targeting these models, we cover different aspects of semantic representation for our input: the GloVe module targets word representation, the Flair modules are for character contextualisation, and the BERT layer is for sentence-level information extraction.

D.0.3. Regularisation.

Given the danger of over-parameterisation that neural networks present, we introduce some regularisation techniques.

For the BiLSTM-CRF model, we perform a **K-Fold Cross Validation** with $K = 3$. The data is not shuffled before splitting into batches. We also add **Dropout**. Results from (Reimers, 2017) show that variational dropout performs best when it comes to BiLSTM networks. Furthermore, it can be shown in (Cheng et al., 2017) that relatively smaller dropout tends to yield better results for LSTM networks. For our experiments, we implement a variational dropout on all layers with the fraction p of dropped values from the set $\{0.1, 0.3, 0.5\}$. The value of $p = 0.1$ performs best after empirical testing and is retained for our final round of benchmark tests. We couple our system with an **Elasticnet** method (i.e., a linear regression model with combined L1 and L2 priors). The tested combinations of regularisation methods are the following (displayed as A, B, C in result Table 3) :

- **Cross Validation (A)**
- **Cross Validation + Dropout (B)**
- **Cross Validation + Dropout + Elasticnet (C)**

The reference labels A, B, C are used in Table 3.

For the BERT-MLP system, we apply the default regularisation A.

D.0.4. Hyperparameters.

We also evaluate the effects of different hyperparameters on the performance of our models. For the BiLSTM-CRF system, the tested configurations in Table 8 are limited to the hardware we use. The hardware specifications are: i7 CPU processor, 16 GB of RAM with a GPU of 8 GB.

¹⁰<https://github.com/flairNLP/flair>

Character Embedding Dimension: We select the CNN approach recommended for deep LSTM networks (Reimers, 2017) and test configurations on the embedding size. Larger embedding dimensions cause redundancy while capturing character-level information and perform worse than smaller sizes.

Word Embedding Dimension: For word-level embeddings, we also test different configurations and their impact on the model. Since we also make use of pre-trained embedding models, we observe that setting the word dimension to a size close to the usual pre-trained word vectors (100 - 300) yields the best performing models.

BiLSTM Layer Dimension: We evaluate 1, 2 and 3 stacked BiLSTM layers.

Recurrent Units: The number of recurrent units u is selected from a range of sizes with $32 \leq u \leq 100$. The forward and reverse running LSTM networks have the same number of recurrent units. Multiple BiLSTM layers also have the same number of recurrent units.

Optimizer: We experiment with two commonly selected optimizers, namely stochastic gradient descent (SGD) and Adamax. SGD is highly sensitive to the learning rate, meaning choosing a too high rate can cause the system to diverge in terms of the objective function, whereas a too low rate results in a slow learning process. Adamax is selected to bypass the shortcomings of SGD.

Learning Rate: We tune the learning rate by hand and observe in many instances that it fails to converge to a minimum. We select a range of viable rates to test for best performance.

Weight Decay (L2 Penalty): We couple weight decay with our optimizer to add regularisation and evaluate the best value for our model.

Parameter	Tested Configurations	Best Configuration
Character Embedding Dimension	25, 50, 75, 100	25
Word Embedding Dimension	50, 200, 300, 500	300
BiLSTM Layer Dimension	1, 2, 3	3
Recurrent Units	32, 50, 64, 75, 100	100
Optimizer	Adamax, SGD	Adamax
Learning Rate	0.0015, 0.002, 0.015, 0.02	0.0015
Weight Decay (L2 Penalty)	$15e-10$, $15e-12$, $15e-14$	$15e-12$

Table 8: Task 2 - Hyperparameter Evaluation

The best configuration for each parameter in Table 8 is considered the optimal one for our use case and our model is tuned to those values for the benchmark tests.

For parameter tuning in the BERT-MLP case, we use a Random Search to facilitate multi-parameter testing. The optimal parameter selection is represented below:

Number of MLP Layers: 3

Number of MLP Hidden Neurons: 500

Max Sequence Length: 173

Number of Epochs: 3

Learning Rate: $5e-5$

Batch Size: 16

Gradient Accumulation Steps: 2

Appendix E. Architecture of Task 2 BiLSTM system.

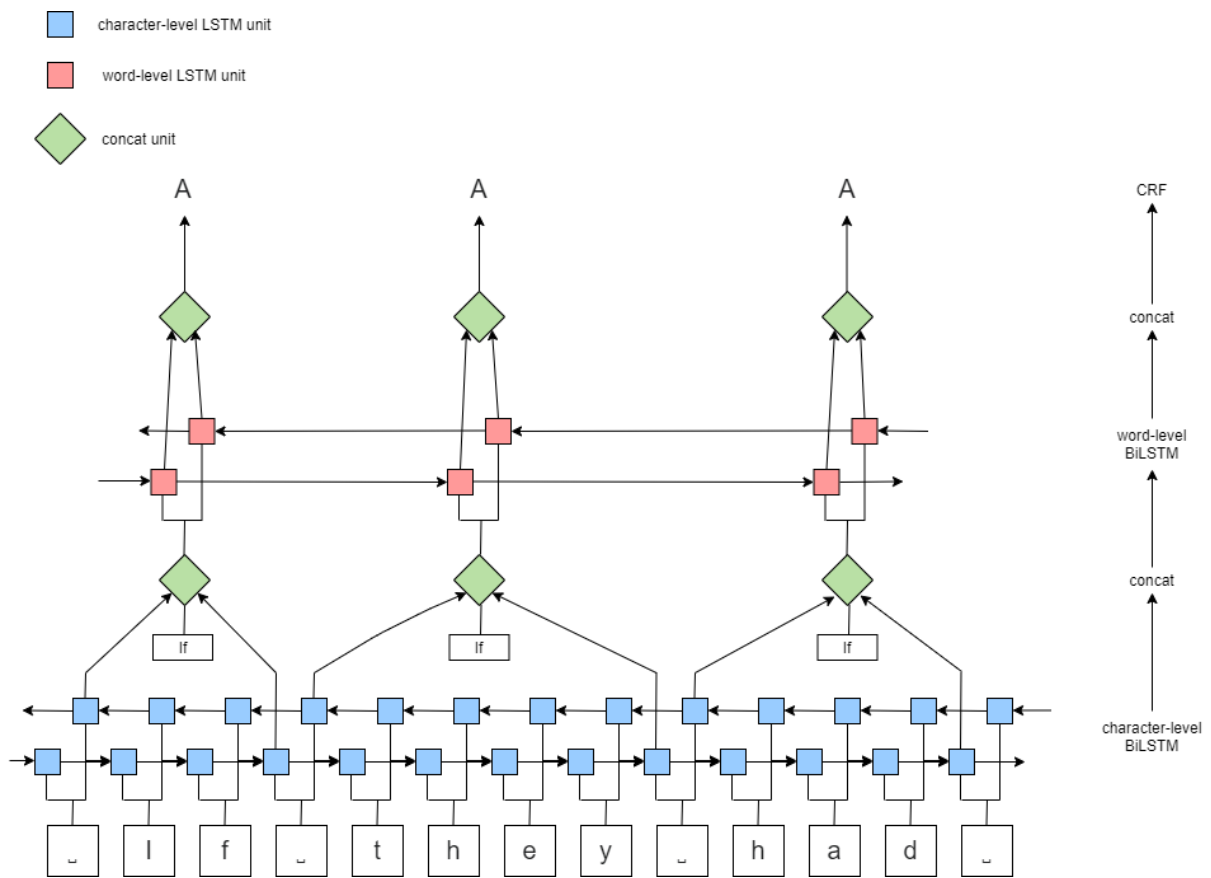


Figure 1: Task 2 - BiLSTM-CRF architecture