Definition Extraction Feature Analysis: From Canonical to Naturally-Occurring Definitions

Mireia Roig Mirapeix **Luis Espinosa-Anke** **Jose Camacho-Collados**
School of Computer Science and Informatics
Cardiff University, United Kingdom
{roigmirapeixm,espinosa-ankel,camachocolladosj}@cardiff.ac.uk

Abstract

Textual definitions constitute a fundamental source of knowledge when seeking the meaning of words, and they are the cornerstone of lexical resources like glossaries, dictionaries, encyclopedia or thesauri. In this paper, we present an in-depth analytical study on the main features relevant to the task of definition extraction. Our main goal is to study whether linguistic structures from canonical (the Aristotelian or *genus et differentia* model) can be leveraged to retrieve definitions from corpora in different domains of knowledge and textual genres alike. To this end, we develop a simple linear classifier and analyze the contribution of several (sets of) linguistic features. Finally, as a result of our experiments, we also shed light on the particularities of existing benchmarks as well as the most challenging aspects of the task.

1 Introduction

Definition Extraction (DE) is the task to extract textual definitions from naturally occurring texts (Navigli and Velardi, 2010). The development of models able to identify definitions in freely occurring text has many applications such as the automatic generation of dictionaries, thesauri and glossaries, as well as e-learning materials and lexical taxonomies (Westerhout, 2009; Del Gaudio et al., 2014; Jurgens and Pilehvar, 2015; Espinosa-Anke et al., 2016). Moreover, definitional knowledge has proven to be a useful signal for improving language models in downstream NLP tasks (Joshi et al., 2020). The task of DE is currently approached almost unanimously as a supervised classification problem, and the latest methods have demonstrated an outstanding performance, to the point of reducing the error rate to less than 2% in some datasets (Veyseh et al., 2019). However, the high performance of these models could be mainly due to artifacts in the data, and thus they may not generalize to different domains.

The main aim of this paper is to analyze to what extent is possible to learn a universal definition extraction system from canonical definitions, and to understand the core differences that currently exist in standard evaluation testbeds. In particular, we propose experiments where we develop a machine learning model able to distinguish definitions with high accuracy in a corpus of canonical definitions, and later evaluate such model in different (pertaining to different domains and genres) datasets. Our evaluation datasets are two, namely: the Word-Class Lattices (WCL) dataset from Navigli et al. (2010), and DEFT, from the SemEval 2020 Task 6 - Subtask 1 (Spala et al., 2019). The former provides an annotated set of definitions and non-definitions with syntactic patterns similar to those of definition sentences from Wikipedia (what the authors call *syntactically plausible false definitions*). The latter presents a robust English corpus that explores the less straightforward cases of term-definition structures in free and semi-structured text from different domains (i.e., biology, history and government), and which is not limited to well-defined, structured, and narrow conditions.

We include a detailed descriptive analysis of both corpora that identifies similarities and differences between definitions and non-definitions, later used for feature selection and analysis. We come to conclusions regarding the discriminative power of certain linguistic features. Interestingly, these features alone

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 81–91
Barcelona, Spain (Online), December 12, 2020

do not have a strong effect on the results, but, the combining feature sets of different nature can improve performance, even in target corpora having heterogeneous domains and non-canonical definitions.

To train and first evaluate the model, we use the annotated WCL dataset. This dataset contains sentences from a sample of the WCL corpus that includes both definitions and non-definitions with syntactic patterns very similar to those found in definitions (e.g. "Snowcap is unmistakable"). The syntactic patterns are simple and represent what we could refer to as canonical definitions. We will test the performance of a model trained on this dataset, and evaluate on the DEFT dataset, which contains a set of definitions and non-definitions from various topics such as biology, history and government.

2 Related Work

Over the last years, DE has received notorious attention for its applications in Natural Language Processing, Computational Linguistics and Computational Lexicography (Espinosa-Anke and Saggion, 2014), as it has been proven to be applicable to glossary generation (Muresan and Klavans, 2002; Park et al., 2002), terminological databases (Nakamura and Nagao, 1988) or question answering systems (Saggion and Gaizauskas, 2004; Cui et al., 2005), among many others.

Research on DE has seen contributions where the task is typically proposed as a binary classification problem (whether a sentence is a definition or not), although with exceptions (Jin et al., 2013). DE has also been studied in languages other than English, e.g., Slavic languages (Przepiórkowski et al., 2007), Spanish (Sierra et al., 2008) or Portuguese (Del Gaudio et al., 2014). Many of these approaches use symbolic methods depending on manually crafted or semi-automatically learned lexico-syntactic patterns (Hovy et al., 2003; Westerhout and Monachesi, 2007) such as 'refers to' or 'is a'.

A notable contribution to DE is the Word Class Lattices model (Navigli and Velardi, 2010), which explores DE on the WCL dataset, a set of encyclopedic definitions and distractors, and which we use in this paper. In a subsequent contribution, Espinosa-Anke and Saggion (2014) present a supervised approach in which only syntactic features derived from dependency relations are used, and whose results are reported higher to the WCL method. For identifying definitions with higher linguistic variability, a weakly supervised approach is presented in Espinosa-Anke et al. (2015). And finally, models based on neural networks have been leveraged for exploiting both long and short-range dependencies, either combining CNNs and LSTMS (Espinosa-Anke and Schockaert, 2018) or BERT (Veyseh et al., 2019), and which are currently the highest performing models on WCL.

3 Data

In this section we present the datasets utilized for our analysis, namely WCL (Section 3.1) and DEFT (Section 3.2), and provide a descriptive analysis comparing both datasets (Section 3.3).

3.1 WCL dataset

The WCL dataset (Navigli et al., 2010) contains 1,772 definitions and 2,847 non-definitions. Each instance is extracted from Wikipedia, and definitions follow a canonical structure following the *genus et differentia* model (i.e., '`X is a Y which Z`'). A preliminary (and shallow) analysis that can be performed without any linguistic detail revolves around comparing the length of definitions vs non definitions. Specifically, definitions have 27.5 words on average, while non-definitions have an average length of 27.2 words. The median for definitions and non-definitions, respectively, is 25 and 24. Although the difference is quite small, it seems that encyclopedic definitions are in general slightly longer.

A particular feature of the WCL dataset is that each candidate is composed of a sentence with part-of-speech and phrase chunking annotation. For definitional sentences, an additional set of tags is provided, which identify core components in definitions such as DEFINIENDUM (term defined), DEFINITOR (definition trigger), DEFINIENS (cluster of words that define the definiendum) and REST (rest of the sentence).

Let us look now at the average length of each of these definition components (see Table 1). The DEFINIENS is typically the most important part of definition sentences (where the definition actually happens), however, it is also the shortest one, followed by the DEFINIENDUM. Moreover, REST is generally the longest but also the one with the highest variance, which fits in with the fact that it is a non-essential part of the definition that can contain varying amounts of information. These results seem to suggest

that the part of the sentence that actually makes it a definition (definiens and definiendum) is, in many occasions, quite short compared to the overall length of the sentence.

	Mean	25% Quartile	Median	75% Quartile	Standard deviation
Definiendum	7.03	2.0	4.0	9.0	7.70
Definiens	4.47	3.0	4.0	5.0	2.94
Rest	14.4	7.0	13.0	20.0	11.66

Table 1: Summary statistics of the length of definiendum, definiens and rest.

The original annotation of the WCL dataset also identifies the main verb of the definition, i.e. that are not in the REST part (Table 1(a) lists the frequent ones). As expected, the verb "to be" tops the list, with four different conjugations taking up the top 5 verbs. Note that these 5 verbs together appear in 1,670 of the 1,772 definitions in the WCL corpus, which could be a sign that the appearance of one of these verbs is a relevant feature to identify definitions. We can also find the most common hypernyms in Table 1(b), although their counts are significantly lower, matching the fact that they are related to the term defined.

Verb	Counts
is	1405
was	114
are	58
refers	58
were	35

Hypernym	Counts
instrument	28
person	22
plants	19
device	14
mammal	12

Table 2: 5 most common main verbs and hypernyms in definitions in the WCL dataset.

3.2 DEFT dataset

The DEFT dataset (Spala et al., 2019) contains 853 sentences, of which 279 are definitions and 574 are non-definitions. It presents a corpus of natural language term-definition pairs embracing different topics such as biology, history, physics, psychology, economics, sociology and government. Sentences have been classified following a new schema that explores how explicit in-text definitions and glosses work in free and semi-structured text, especially those whose term-definition pairs span crosses a sentence boundary and those lacking explicit definition phrases. Thus, they identify as definitions sentences where the relation between a term and a definition requires more deduction than finding a definition verb phrase. Their focus is to identify terms and definitions, but not necessarily the verb that may or may not connect them two, which identifies as definitions a broather variety of structures.

In this case, the average length of definitions is 27.38 and non-definitions have an average length of 23.84. The median length for definitions and non-definitions is 26 and 22 respectively.

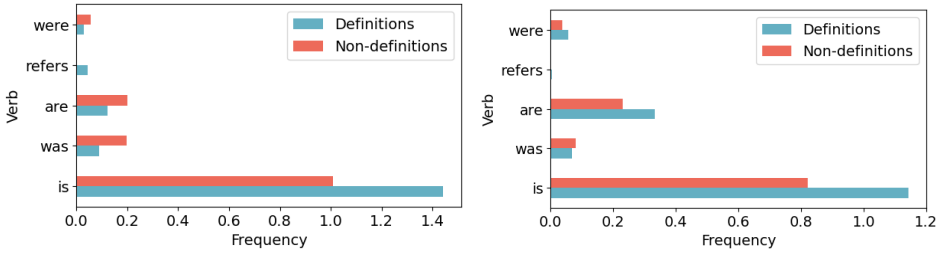
3.3 Descriptive analysis

In this section we perform a short descriptive analysis comparing the two datasets. Continuing with the instance length analysis, Table 3 shows statistics for both datasets, this time comparing length of positive (definition) and negative (non definition) sentences. As can be observed, definitions generally tend to be longer than non-definitions, although the main part of the definition is quite short compared to its overall length. Moreover, while the distribution of definitions/non-definitions is similar, the number of instances is considerably larger in the WCL corpus, which is improtant to note, as we will use it as our training set in our experiments (cf. Section 4.1.)

Regarding frequency of specific POS tags, in Section 3.1 we have seen how some verbs are extremely abundant in definitions in the WCL corpus. However, these are quite common verbs in general in these datasets, as Figures 1(a) and 1(b) show. Note that, for instance, 'is' is more frequent in definitions in both datasets, with an average frequency greater than 1 in both datasets (1.4% and 1.1% in WCL and DEFT, respectively). However, 'was' is actually the opposite and is more frequent in non-definitions while the others are much less common and do not seem to be as present in both types of sentences.

		Instances	Mean Length	Median Length
WCL	Definitions	1772 (38.36%)	27.5	25
	Non-Definitions	2847 (61.64%)	27.2	24
DEFT	Definitions	279 (32.71%)	27.38	26
	Non-Definitions	574 (67.29%)	23.84	22

Table 3: Number of instances, mean and median length for definitions and non-definitions from both WCL and DEFT datasets.



(a) WCL dataset. (b) DEFT dataset.

Figure 1: Frequency of common verbs in definitions and non-definitions.

Concerning hypernyms (a.k.a `genus` in Aristotelian definitions), although the counts are much lower for hypernyms than for verbs (Table 2), in Figure 2 we illustrate how the hypernyms that appear at least 5 times in the WCL dataset are usually more common in definitions in both datasets. The presence of such hypernyms is likely to be more related to the topics defined than the structure of the sentence, but having any kind of hypernym is probably a relevant feature of definitions, as canonical or lexicographic definitions have (or should have) at least one.

We observed that definitions and non-definitions present different frequencies of POS and chunk patterns. In the WCL dataset it seems that definitions have a higher frequency of noun phrases (denoted as 'NP' or 'NP NN', for instance), while non-definitions have more prepositional phrases ('PP' or 'PP IN'). However, we do not observe these similarities in the DEFT dataset.

Finally, we computed the most PoS-based patterns structures[1] (occurring at least 5 times) in the main part of definitions from the WCL dataset. We have observed that these structures are much more common in definitions than in non-definitions in both corpora, which seems to indicate definitions tend to use a particular morphosyntactic set of structures which can be strong indicators of definitional knowledge.

4 Evaluation

In this section we explain our experiments in definition extraction. In particular, we train a supervised model on the WCL corpus of canonical definitions, and tested on the same corpus (via cross-validation) and the DEFT corpus. With this experiment we aim at understanding relevant features for definition extraction and whether features from canonical definitions can be extrapolated to other domains.

Section 4.1 describes the experimental settings and Section 4.2 presents the main results.

4.1 Experimental setting

In the following we explain the experimental setting for our definition extraction experiments. In Section 4.1.1 we explain our supervised definition extraction model and its features inspired by our descriptive analysis. Then, we explain the data preprocessing (Section 4.1.2) and training details (Section 4.1.3).

4.1.1 Model and features

As supervised model we made use of a Support Vector Machine (SVM) given its efficiency and effectiveness in handling a large set of linguistic features. The model uses an RBF kernel and a combination of different features. The main one is based on n-grams of range 1 to 3 from the tagged sentences, i.e. each word contains chunk tag, PoS tag and word separated by an underscore.

[1]PoS-based patterns are any ordered sequences of tags (PoS or chunk) such as 'NP DT' (noun phrase followed by a determiner).

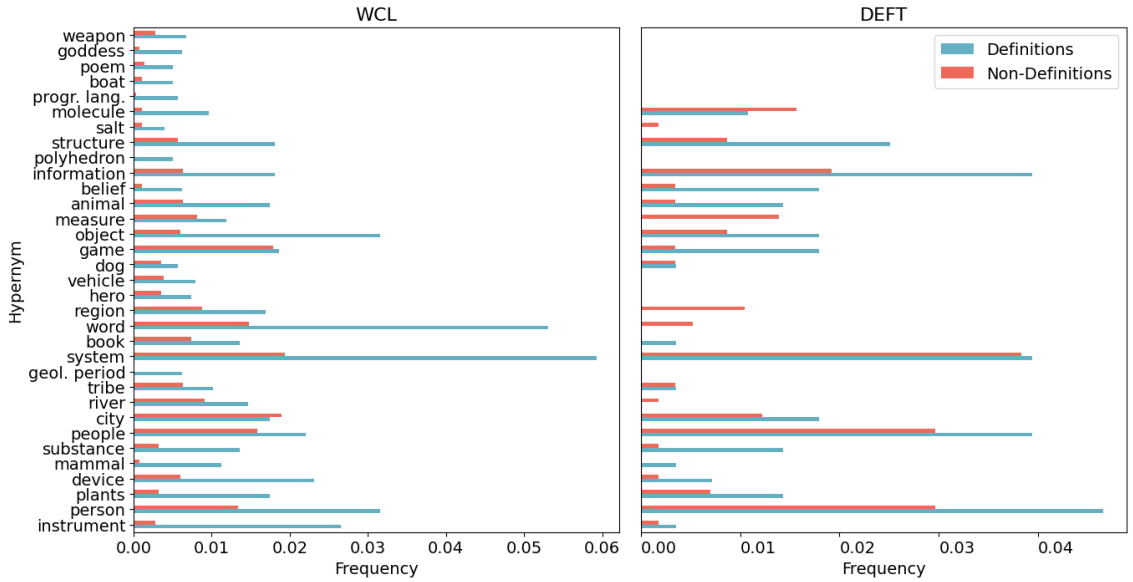
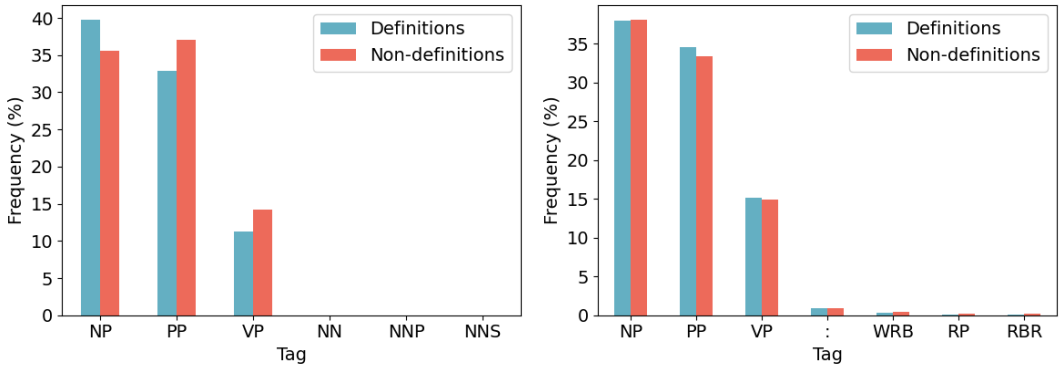


Figure 2: Average frequency of common hypernyms in definitions and non-definitions.



(a) WCL dataset.

(b) DEFT dataset.

Figure 3: Presence of chunk and PoS tags in definitions and non definitions.

The other features are based on the findings from Section 3.3. For each training set, the model computes the 5 most common definition verbs, i.e. in the main part of the definition, the 20 most common hypernyms, the 10 most common composition of chunk and PoS tags, the 6 most common chunk tags[2], the 10 most common structures of chunk and PoS tags combined, the 10 most common structures of chunk tags and the maximum length of definitions. Using this, we obtain the following new features:

- VERB: Count of common verbs present in the sentence.
- HYP: Count of common hypernyms present in the sentence.
- CT-Ch, CT-Ch&PoS: For each of the 6 most common chunk tags and the 10 most common combinations of chunk and PoS, number of occurrences divided by total number of tags in the sentence.
- STR-Ch, STR-Ch&PoS: For each of the 10 most common structures (chunk and combination of chunk and PoS respectively), a binary variable indicating if the structure is present in the sentence.
- LEN: The length of the sentence divided by the maximum length of a definition.

4.1.2 Data preprocessing

As each corpus contains different information and has a different structure, their preprocessing is slightly different, although the output has the same format: a matrix where the features are obtained from.

[2]After the 6th most common, the appearances are significantly lower and hardly relevant.

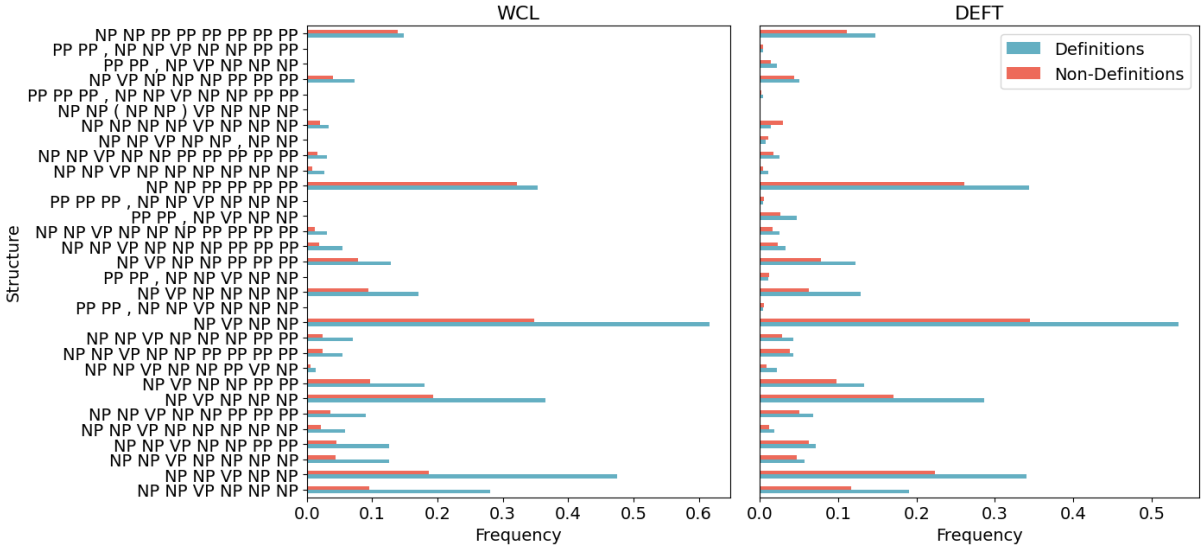


Figure 4: Presence of structures of chunk tags in definitions and non-definitions.

As the WCL dataset contains all the definitions' annotations, we classify each part in a different column and also the verbs and hypernyms annotated. We later retag the sentences with PoS tags and chunk, using, respectively, the NLTK[3] pos_tag function and the RegexpParser with the following grammar:

```
parser = RegexpParser('''
    VP: {<MD>?<RB.?>?<V.*>+(<CC><V.*>+)*} {<TO>?<V.*>+(<CC><V.*>+)*}
    PP: {<TO|IN>+<DT>*<CD>*<JJ.?>*<N.*>*<CC>*<N.*>*} {<TO|IN>+<WDT|EX|WP.?|PP.?|RB>*}
    NP: {<DT>*<CD>*<JJ.?>*<N.*|P.*>*<CC>*<N.*|PP.?>*} {<WDT|EX|WP.?|P.*|RB>*}''')
```

It distinguishes 3 phrases: verb phrases (contain a verb sometimes preceded by a modal or 'to', with possible adverbs and another verb after a coordinating conjunction), prepositional phrases (starting with a preposition and followed by determinants, cardinal numbers, nouns or pronouns) and noun phrases (including a noun or pronoun sometimes preceded by determiners, cardinal numbers or adjectives).

The output is a matrix where each row corresponds to a sentence and each column has different information such as the sentence (tagged and not), the term being defined, the hypernyms annotated in the sentence, the main verb of the definition, the label and different columns that contain the tags (both PoS tags and chunk or only chunk) for the whole sentence and for the main part of the definition (definiendum and definiens). For non-definitions, some columns such as the verb, hypernym and tags of the main part of the definition contain NaN values, as they only exist for definitions.

The preprocessing for the DEFT corpus is simpler: we tag the sentences using the same rules and save the sentences, tags and label in different columns. Numbers at the beginning of sentences have removed.

4.1.3 Training procedure

As mentioned earlier, the model was trained on the WCL dataset. We used sklearn[4] for training and evaluating the SVM model. For the experiments, the SVM hyperparameters were chosen after testing the following values: $[0.0001, 0.001, 0.1, 1, 5, 10, 50, 100]$ for C, and $[0.0001, 0.001, 0.1, 1, 5, 10, 50, 100]$ for gamma, both in a validation set. Finally, the evaluation on the WCL dataset is performed through 10-fold cross-validation, with 10% of the corpus used for validation in each fold. Then, the model is trained on the whole WCL corpus and evaluated on the DEFT corpus. The final hyperparameters of the SVM were $C = 5$ and gamma$= 0.1$. In addition to the SVM model, as a baseline we trained a Naive Bayes with the same features. This model was trained with its standard implementation in sklearn.

4.2 Results

The results on the WCL dataset are displayed in Table 4. As a naive baseline we include the results of a system that would identify all sentences as definitions (referred to as *Naive(all defs)* in the table).

[3] https://www.nltk.org/

[4] https://scikit-learn.org/stable/

As can be observed, all metrics are above 0.97 and the average metrics are all close to 0.98. This proves the reliability of the SVM model with all our proposed linguistic features, which attains the highest performance of any non-linear model in the task. As a point of comparison, recent works have reported slightly worse results using highly parametrized models such as convolutional and recurrent neural networks (Espinosa-Anke and Schockaert, 2018).

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9805	0.9820	0.9776	0.9797
2	0.9762	0.9755	0.9761	0.9758
3	0.9827	0.9838	0.9796	0.9816
4	0.9870	0.9872	0.9847	0.9859
5	0.9740	0.9756	0.9667	0.9710
6	0.9740	0.9748	0.9717	0.9731
7	0.9892	0.9901	0.9876	0.9888
8	0.9784	0.9767	0.9732	0.9750
9	0.9827	0.9828	0.9810	0.9819
10	0.9805	0.9784	0.9796	0.9790
Average	0.9805	0.9807	0.9778	0.9792
Naive Bayes	0.8837	0.8849	0.8686	0.8743
Naive(all defs)	0.3836	0.1918	0.5000	0.2768

Table 4: Results of the SVM model on the WCL dataset using 10-fold cross validation. Precision, recall and F1 are macro metrics. The last two rows include the average results of the two baselines considered.

When testing the model on the DEFT corpus, the results are not close to being as satisfactory as they are in the WCL dataset, as we can see in Table 5. The model trained on the WCL dataset performs significantly worse than other recent models (Spala et al., 2020), which could be expected given the different nature of the definitions. In the following section we provide a more extensive analysis that also attempts at explaining the performance difference between the two datasets.

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.7011	0.6573	0.5900	0.5872
Naive Bayes	0.5909	0.5626	0.5689	0.5611
Naive(all defs)	0.3271	0.1635	0.5000	0.2465

Table 5: DEFT results of the SVM and baselines trained on the WCL corpus.

5 Analysis

5.1 Feature analysis

Figure 5 shows the features of the model with highest χ^2. Some of them are compositions extremely common in definitions such as 'is a', 'is an' or 'refers', but we also find others more topic related such as 'mythology' or 'greek', which would probably be artifacts from the WCL dataset.

For a detailed view of each additional feature's significance, we ran the model removing one or more features at a time. Morever, we also ran the model using the n-gram features only, with different combinations of words and tags. We can find this feature analysis in Table 6. Although the accuracy in the 10-fold cross-validation setting does not change significantly when removing only one feature, and even improves slightly in the case of hypernyms, they do show changes when evaluating on the DEFT corpus. We observe significantly lower accuracy when removing more than one feature at a time (last two rows), decreasing regularly when removing more features and obtaining between $0.93 - 0.94$ using only n-gram features, which indicates that these features rely on and interact with each other to improve accuracy. The differences are more significant when evaluating the model on the DEFT corpus, the accuracy goes from around 0.70 when using all features to 0.55 when removing some of them. This proves that the

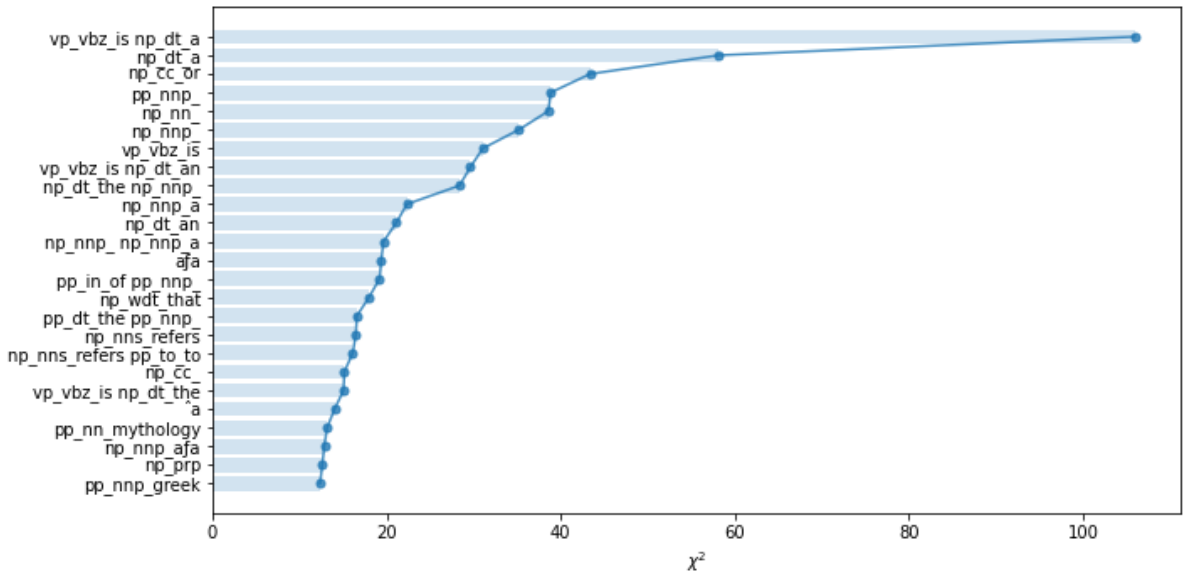


Figure 5: Features from the SVM model trained on WCL with highest χ^2.

additional features are relevant to identify definitions and improve the metrics significantly, especially in unseen corpora. In fact, the performance of using n-grams features only achieves a performance of 0.934 F1 in the in-domain WCL corpus, and a significantly lower 0.575 performance on the DEFT corpus.

Features Removed (Number of features)	Average of 10-fold cross validation				Evaluation of model on DEFT corpus			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
None (0)	0.9805	0.9807	0.9778	0.9792	0.7011	0.6573	0.5900	0.5872
VERB (1)	0.9805	0.9807	0.9778	0.9792	0.7011	0.6573	0.5900	0.5872
HYP (1)	0.9812	0.9818	0.9781	0.9799	0.7046	0.6626	0.5972	0.5968
LEN (1)	0.9801	0.9803	0.9773	0.9787	0.7022	0.6599	0.5909	0.5881
CT-Ch&PoS (10)	0.9803	0.9803	0.9777	0.979	0.6928	0.6395	0.5894	0.5889
CT-Ch (5)	0.9799	0.9795	0.9775	0.9785	0.6928	0.6398	0.6088	0.6132
STR-Ch&PoS (10)	0.9805	0.9807	0.9778	0.9792	0.7011	0.6573	0.5900	0.5872
STR-Ch (10)	0.9805	0.9807	0.9778	0.9792	0.7011	0.6573	0.5900	0.5872
CT (15)	0.9706*	0.9697	0.9673	0.9684	0.5557*	0.5842	0.5934	0.5515
STR (20)	0.9719	0.9707	0.9691	0.9699	0.6131	0.5951	0.6066	0.5918
CT and STR (35)	0.9706	0.9697	0.9673	0.9684	0.5662	0.5907	0.6013	0.5612
All except n-grams (38)	0.9379	0.9349	0.9336	0.934	0.5885	0.5875	0.5994	0.5754

Table 6: Results of the SVM model (trained on the WCL dataset) using different sets of features. For accuracy, * indicates when the results start to show differences that are statistically significant (p-value$<$ 0.05 according to a t-test) with respect to the model using all features (first row).

Furthermore, we can see in Table 7 how the n-gram model is significantly more accurate when using both PoS and chunk tags and words rather than only some of them, which indicates that both words and structure of the sentence determine whether it is a definition or not.

5.2 Error analysis

In Table 8 we can see some examples of predictions from the model that provide a more in-depth view. We observe how the model is successful in correctly predicting sentences with unorthodox structures, such as non-definitions using the verb "is", and syntactically complex definitions. Moreover, some of sentences that have been predicted wrongly as definitions could be considered as definitions, but they are not defining the target word. The false negatives present complex structures probably unseen for the model. Thus, evidence suggests the model succeeds most of the times at identifying definitions and non-definitions, and has incorporated satisfactorily the distinctive characteristics of each kind of sentence.

As for the DEFT dataset, as expected from the obtained accuracy, the model makes numerous mistakes.

Terms used for n-gram	Average of 10-fold cross validation				Evaluation of model on DEFT corpus			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Chunk, PoS, words	0.9379	0.9349	0.9336	0.934	0.5885	0.5875	0.5994	0.5754
PoS tags, words	0.8809	0.8841	0.8624	0.8705	0.687	0.6383	0.551	0.5251
Words	0.8326	0.8249	0.8182	0.8206	0.6729	0.6009	0.5562	0.5452
Chunk, PoS tags	0.8465	0.8459	0.8259	0.8329	0.6917	0.6433	0.6433	0.5674

Table 7: Results of the SVM model using different types of n-gram features only.

It has a large number of false negatives (23.9 %), making its predictions less reliable in this setting. The model does a good job at detecting true negatives (91.1% of all negative instances), also due to the fact that most sentences are predicted as non-definitions. However, some false negatives do not seem to contain definitional information. Something similar happens with false positives, as some of them would most likely be considered definitions under more flexible criteria. Thus, although the performance of the model on this data set seems to be relatively low overall, this is probably because of the different tagging criteria, as many sentences that appeared as incorrectly predicted, could be labelled correctly under the annotation criteria used in the WCL dataset. For instance, the sentence "Elimination blackjack is a tournament format of blackjack." could be considered a definition with the criteria used in the DEFT dataset as it presents a *direct-defines* relation, while "It carries the correct amino acid to the site of protein synthesis" would not be considered a definition in the WCL corpus as it is not an actual textual definition.

		Predicted nodef*	Predicted def*
WCL	nodef	His death is deeply mourned by Alleycats fans as seen in the press and media. Covering the head is respectful in Sikhism and if a man is not wearing a turban, then a rumāl must be worn before entering the Gurdwara. The following are links to pictures of Myddfai taken by the club.	The term "carbonate" is also commonly used to refer to one of these salts or carbonate minerals. Elimination blackjack is a tournament format of blackjack. Balderton Old Boys also are a local football team.
	def	The Callitrichinae form one of the four families of New World monkeys now recognised In everyday usage, risk is often used synonymously with the probability of a known loss. Both equivocation and amphiboly are fallacies arising from ambiguity.	The Aurochs or urus (Bos primigenius) was a very large type of cattle that was prevalent in Europe until its extinction in 1627. In the 19th century the term anglicanism was coined to describe the common religious tradition of these churches. The term biotic refers to the condition of living organisms.
		Predicted nodef*	Predicted def*
DEFT	nodef	Living things are highly organized and structured , following a hierarchy that can be examined on a scale from small to large. At its most fundamental level , life is made up of matter. It consists of a nucleus surrounded by electrons.	Transfer RNA (tRNA) is one of the smallest of the four types of RNA , usually 70 – 90 nucleotides long. A microphyll is small and has a simple vascular system. An individual with dyslexia exhibits an inability to correctly process letters.
	def	It carries the correct amino acid to the site of protein synthesis. The rays themselves are called nuclear radiation. Herbivores eat plant material , and planktivores eat plankton.	The atom is the smallest and most fundamental unit of matter. A prokaryote is a simple, mostly single-celled (unicellular) organism that lacks a nucleus, or any other membrane-bound organelle. Matter is any substance that occupies space and has mass.

Table 8: Definition (def*) and non-definition (nodef*) predictions on both WCL and DEFT ground truth (for def and nodef classes).

6 Conclusion and Future Work

In conclusion, extracting definitions from texts is a challenging research task, which is highly dependant on the distribution and scope of the application. Nonetheless, in this paper we have shown that a simple SVM model trained on a dataset with canonical definitions using linguistic features can provide high performance while helping us understand the task better. This model has also been evaluated on a corpus with heterogeneous domains, which also provided us with insights on the qualitative difference among definitions in each setting.

Our descriptive analysis discovered interesting differences and similarities between definitions and non-definitions that can be used to differentiate them automatically. The inclusion of linguistic features based on our analysis improved significantly the performance of the model. As future work it would be interesting to extend the analysis to corpora of different characteristics and languages. As an straightforward application, a model with accurate performance across corpora would allow the automatic creation of dictionaries from general or specialized domains, as well as to better understand certain topics.

Acknowledgements

We thank the reviewers for their feedback and Emrah Ozturk for his help in the early stages of this paper.

References

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 384–391, New York, NY, USA. Association for Computing Machinery.

Rosa Del Gaudio, Gustavo Batista, and António Branco. 2014. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, 20(3):327–359.

Luis Espinosa-Anke and Horacio Saggion. 2014. Applying dependency relations to definition extraction. In *Natural Language Processing and Information Systems, NLDB 2014*, pages 63–74. Springer International Publishing Switzerland 2014, Montpellier, France, 06.

Luis Espinosa-Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385.

Luis Espinosa-Anke, Horacio Saggion, and Francesco Ronzano. 2015. Weakly supervised definition extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 176–185, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. Extasem! extending, taxonomizing and semantifying domain terminologies. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Eduard Hovy, Andrew Philpot, Judith Klavans, Ulrich Germann, and Peter T. Davis. 2003. Extending metadata definitions by automatically extracting and organizing glossary definitions. In *Proceedings of the 2003 Annual National Conference on Digital Government Research*, dg.o '03, page 1. Digital Government Society of North America.

Yiping Jin, Min-Yen Kan, Jun Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the acl anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

David Jurgens and Mohammad Taher Pilehvar. 2015. Reserating the awesometastic: An automatic extension of the wordnet taxonomy for novel terms. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1459–1465.

Smaranda Muresan and Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Jun-ichi Nakamura and Makoto Nagao. 1988. Extraction of semantic information from an ordinary english dictionary and its evaluation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, page 459–464, USA. Association for Computational Linguistics.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden, July. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, and Juana Maria Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Youngja Park, Roy J Byrd, and Branimir K Boguraev. 2002. Automatic glossary extraction: Beyond terminology identification. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, page 1–7, USA. Association for Computational Linguistics.

Adam Przepiórkowski, Łukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kubon, and Beata Wójtowicz. 2007. Towards the automatic extraction of definitions in slavic. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 43–50.

Horacio Saggion and Rob Gaizauskas. 2004. Mining on-line sources for definition knowledge. In *In Proceedings of the 17th FLAIRS 2004, Miami Bearch*, 01.

Gerardo Sierra, Rodrigo Alarcón, César Aguilar, and Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 14(1):74–98.

Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. DEFT: A corpus for definition extraction in free- and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy, August. Association for Computational Linguistics.

Sasha Spala, Nicholas A Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. Semeval-2020 task 6: Definition extraction from free text with the deft corpus.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2019. A joint model for definition extraction with syntactic connection and semantic consistency. *arXiv preprint arXiv:1911.01678*.

Eline Westerhout and Paola Monachesi. 2007. Extraction of dutch definitory contexts for elearning purpose. In Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde, editors, *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN 2007)*, pages 219–34. CLIN, Nijmegen, Netherlands.

Eline Westerhout. 2009. Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 61–67.