What is the Essence of a Claim? Cross-Domain Claim Identification

Johannes Daxenberger[†], **Steffen Eger**[†‡], **Ivan Habernal**[†], **Christian Stab**[†], **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract

Argument mining has become a popular research area in NLP. It typically includes the identification of argumentative components, e.g. claims, as the central component of an argument. We perform a qualitative analysis across six different datasets and show that these appear to conceptualize claims quite differently. To learn about the consequences of such different conceptualizations of claim for practical applications, we carried out extensive experiments using state-of-the-art feature-rich and deep learning systems, to identify claims in a cross-domain fashion. While the divergent conceptualization of claims in different datasets is indeed harmful to cross-domain classification, we show that there are shared properties on the lexical level as well as system configurations that can help to overcome these gaps.

1 Introduction

The key component of an argument is the *claim*. This simple observation has not changed much since the early works on argumentation by Aristotle more than two thousand years ago, although argumentation scholars provide us with a plethora of often clashing theories and models (van Eemeren et al., 2014). Despite the lack of a precise definition in the contemporary argumentation theory, Toulmin's influential work on argumentation in the 1950's introduced a claim as an 'assertion that deserves our attention' (Toulmin, 2003, p. 11); recent works describe a claim as 'a statement that is in dispute and that we are trying to support with reasons' (Govier, 2010).

Argument mining, a computational counterpart of manual argumentation analysis, is a recent growing sub-field of NLP (Peldszus and Stede, 2013a). 'Mining' arguments usually involves several steps like separating argumentative from non-argumentative text units, parsing argument structures, and recognizing argument components such as claims—the main focus of this article. Claim identification itself is an important prerequisite for applications such as fake checking (Vlachos and Riedel, 2014), politics and legal affairs (Surdeanu et al., 2010), and science (Park and Blake, 2012).

Although claims can be identified with a promising level of accuracy in typical argumentative discourse such as persuasive essays (Stab and Gurevych, 2014; Eger et al., 2017), less homogeneous resources, for instance online discourse, pose challenges to current systems (Habernal and Gurevych, 2017). Furthermore, existing argument mining approaches are often limited to a single, specific domain like legal documents (Mochales-Palau and Moens, 2009), microtexts (Peldszus and Stede, 2015), Wikipedia articles (Levy et al., 2014; Rinott et al., 2015) or student essays (Stab and Gurevych, 2017). The problem of generalizing systems or features and their robustness across heterogeneous datasets thus remains fairly unexplored.

This situation motivated us to perform a detailed analysis of the concept of claims (as a key component of an argument) in existing argument mining datasets from different domains.[1] We first review and qualitatively analyze six existing publicly available datasets for argument mining (§3), showing that the conceptualizations of claims in these datasets differ largely. In a next step, we analyze the influence of these differences for cross-domain claim identification. We propose several computational models for claim identification,

[1]We take the machine learning perspective in which different *domains* mean data drawn from different distributions (Murphy, 2012, p. 297).

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2055–2066
Copenhagen, Denmark, September 7–11, 2017.

including systems using linguistically motivated features (§4.1) and recent deep neural networks (§4.2), and rigorously evaluate them on and across all datasets (§5). Finally, in order to better understand the factors influencing the performance in a cross-domain scenario, we perform an extensive quantitative analysis on the results (§6).

Our analysis reveals that despite obvious differences in conceptualizations of claims across datasets, there are some shared properties on the lexical level which can be useful for claim identification in heterogeneous or unknown domains. Furthermore, we found that the choice of the source (training) domain is crucial when the target domain is unknown. We release our experimental framework to help other researchers build upon our findings.[2]

2 Related Work

Existing approaches to argument mining can be roughly categorized into (*a*) *multi-document* approaches which recognize claims and evidence across several documents and (*b*) *discourse level* approaches addressing the argumentative structure within a single document. Multi-document approaches have been proposed e.g. by Levy et al. (2014) and Rinott et al. (2015) for mining claims and corresponding evidence for a predefined topic over multiple Wikipedia articles. Nevertheless, to date most approaches and datasets deal with single-document argumentative discourse. This paper takes the discourse level perspective, as we aim to assess multiple datasets from different authors and compare their notion of 'claims'.

Mochales-Palau and Moens (2009) experiment at the discourse level using feature-rich SVM and a hand-crafted context-free grammar in order to recognize claims and premises in legal decisions. Their best results for claims achieve 74.1% F_1 using domain-dependent key phrases, token counts, location features, information about verbs, and the tense of the sentence. Peldszus and Stede (2015) present an approach based on a minimum spanning tree algorithm and model the global structure of arguments considering argumentative relations, the stance and the function of argument components. Their approach yields 86.9% F_1 for recognizing claims in English 'microtexts'. Habernal and Gurevych (2017) cast argument component identification as BIO sequence labeling and jointly model separation of argumentative from non-argumentative text units and identification of argument component boundaries together with their types. They achieved 25.1% Macro-F_1 with a combination of topic, sentiment, semantic, discourse and embedding features using structural SVM. Stab and Gurevych (2014) identified claims and other argument components in student essays. They experiment with several classifiers and achieved the best performance of 53.8% F_1 score using SVM with structural, lexical, syntactic, indicator and contextual features. Although the above-mentioned approaches achieve promising results in particular domains, their ability to generalize over heterogeneous text types and domains remains unanswered.

Rosenthal and McKeown (2012) set out to explore this direction by conducting cross-domain experiments for detecting claims in blog articles from LiveJournal and discussions taken from Wikipedia. However, they focused on relatively similar datasets that both stem from the social media domain and in addition annotated the datasets themselves, leading to an identical conceptualization of the notion of claim. Although Al-Khatib et al. (2016) also deal with cross-domain experiments, they address a different task; namely identification of argumentative sentences. Further, their goals are different: they want to improve argumentation mining via distant supervision rather than detecting differences in the notions of a claim.

Domain adaptation techniques (Daume III, 2007) try to address the frequently observed drop in classifier performances entailed by a dissimilarity of training and test data distributions. Since techniques such as learning generalized cross-domain representations in an unsupervised manner (Blitzer et al., 2006; Pan et al., 2010; Glorot et al., 2011; Yang and Eisenstein, 2015) have been criticized for targeting specific source and target domains, it has alternatively been proposed to learn *universal* representations from general domains in order to render a learner robust across *all* possible domain shifts (Müller and Schütze, 2015; Schnabel and Schütze, 2013). Our approach is in a similar vein. However, rather than trying to improve classifier performances for a specific source-target domain pair, we want to detect *differences* between these pairs. Furthermore, we are looking

[2] https://github.com/UKPLab/emnlp2017-claim-identification

Corpus	*Reference*	*Genre*	*#Docs*	*#Tokens*	*#Sentences*	*#Claims*
VG	Reed et al. (2008)	various genres	507	60,383	2,842	563 (19.81%)
WD	Habernal and Gurevych (2015)	web discourse	340	84,817	3,899	211 (5.41%)
PE	Stab and Gurevych (2017)	persuasive essays	402	147,271	7,116	2,108 (29.62%)
OC	Biran and Rambow (2011a)	online comments	2,805	125,677	8,946	703 (7.86%)
WTP	Biran and Rambow (2011b)	wiki talk pages	1,985	189,140	9,140	1,138 (12.45%)
MT	Peldszus and Stede (2015)	micro texts	112	8,865	449	112 (24.94%)

Table 1: Overview of the employed corpora.

for *universal* feature sets or classifiers that perform generally well for claim identification across varying source and target domains.

3 Claim Identification in Computational Argumentation

We briefly describe six English datasets used in our empirical study; they all capture claims on the discourse level. Table 1 summarizes the dataset statistics relevant to claim identification.

3.1 Datasets

The AraucariaDB corpus (Reed et al., 2008) includes various genres (**VG**) such as newspaper editorials, parliamentary records, or judicial summaries. The annotation scheme structures arguments as trees and distinguishes between claims and premises at the clause level. Although the reliability of the annotations is unknown, the corpus has been extensively used in argument mining (Moens et al., 2007; Feng and Hirst, 2011; Rooney et al., 2012).

The corpus from Habernal and Gurevych (2017) includes user-generated web discourse (**WD**) such as blog posts, or user comments annotated with claims and premises as well as backings, rebuttals and refutations (α_U 0.48) inspired by Toulmin's model of argument (Toulmin, 2003).

The persuasive essay (**PE**) corpus (Stab and Gurevych, 2017) includes 402 student essays. The scheme comprises major claims, claims and premises at the clause level (α_U 0.77). The corpus has been extensively used in the argument mining community (Persing and Ng, 2015; Lippi and Torroni, 2015; Nguyen and Litman, 2016).

Biran and Rambow (2011a) annotated claims and premises in online comments (**OC**) from blog threads of LiveJournal (κ 0.69). In a subsequent work, Biran and Rambow (2011b) applied their annotation scheme to documents from Wikipedia talk pages (**WTP**) and annotated 118 threads. For our experiments, we consider each user comment in both corpora as a document, which yields 2,805 documents in the OC corpus and 1,985 documents in the WTP corpus.

Peldszus and Stede (2016) created a corpus of German microtexts (**MT**) of controlled linguistic and rhetoric complexity. Each document includes a single argument and does not exceed five argument components. The scheme models the argument structure and distinguishes between premises and claims, among other properties (such as proponent/opponent or normal/example). In the first annotation study, 26 untrained annotators annotated 23 microtexts in a classroom experiment (κ 0.38) (Peldszus and Stede, 2013b). In a subsequent work, the corpus was largely extended by expert annotators (κ 0.83). Recently, they translated the corpus to English, resulting in the first parallel corpus in computational argumentation; our experiments rely on the English version.

3.2 Qualitative Analysis of Claims

In order to investigate how claim annotations are tackled in the chosen corpora, one co-author of this paper manually analyzed 50 randomly sampled claims from each corpus. The characteristics taken into account are drawn from argumentation theory (Schiappa and Nordin, 2013) and include among other things the claim type, signaling words and discourse markers.

Biran and Rambow (2011b) do not back-up their claim annotations by any common argumentation theory but rather state that claims are *utterances which convey subjective information and anticipate the question 'why are you telling me that?'* and need to be supported by justifications. Using this rather loose definition, a claim might be any subjective statement that is justified by the author. Detailed examination of the LiveJournal corpus (OC) revealed that sentences with claims are extremely noisy. Their content ranges from a single word, (*"Bastard."*), to emotional expressions of personal regret, (*"::hugs:: i am so sorry hon .."*)

to general Web-chat nonsense (*"W-wow... that's a wicked awesome picture... looks like something from Pirates of the Caribbean...gone Victorian ...lolz."*) or posts without any clear argumentative purpose (*"what i did with it was make this recipe for a sort of casserole/stratta (i made this up, here is the recipe) [...] butter, 4 eggs, salt, pepper, sauted onions and cabbage..add as much as you want bake for 1 hour at 350 it was seriously delicious!"*). The Wikipedia Talk Page corpus (WTP) contains claims typical to Wikipedia quality discussions (*"That is why this article has NPOV issues."*) and policy claims (Schiappa and Nordin, 2013) are present as well (*"I think the gallery should be got rid of altogether."*). However, a small number of nonsensical claims remains (*"A dot."*).

Analysis of the MT dataset revealed that about half of claim sentences contain the modal verb *'should'*, clearly indicating policy claims (*"The death penalty should be abandoned everywhere."*). Such statements also very explicitly express the stance on the controversial topic of interest. In a similar vein, claims in persuasive students' essays (PE) heavily rely on phrases signaling beliefs (*"In my opinion, although using machines have many benefits, we cannot ignore its negative effects."*) or argumentative discourse connectors whose usage is recommended in textbooks on essay writing (*"Thus, it is not need for young people to possess this ability."*). Most claims are value/policy claims written in the present tense.

The mixture of genres in the AraucariaDB corpus (VG) is reflected in the variety of claims. While some are simple statements starting with a discourse marker (*"Therefore, 10% of the students in my logic class are left-handed."*), there are many legal-specific claims requiring expert knowledge (*"In considering the intention of Parliament when passing the 1985 Act, or perhaps more properly the intention of the draftsman in settling its terms, there are [...]"*), reported and direct speech claims (*"Eight-month-old Kyle Mutch's tragic death was not an accident and he suffered injuries consistent with a punch or a kick, a court heard yesterday."*), and several nonsensical claims (*"RE: Does the Priest Scandal Reveal the Beast?"*) which undercut the consistency of this dataset.

The web-discourse (WD) claims take a clear stance to the relevant controversy (*"I regard single sex education as bad."*), yet sometimes anaphoric (*"My view on the subject is no."*). The usage of discourse markers is seldom. Habernal and Gurevych (2017) investigated hedging in claims and found out that it varies with respect to the topic being discussed (10% up to 35% of claims are hedged). Sarcasm or rhetorical question are also common (*"In 2013, is single-sex education really the way to go?"*).

These observations make clear that annotating claims—the central part of all arguments, as suggested by the majority of argumentation scholars—can be approached very differently when it comes to actual empirical, data-driven operationalization. While some traits are shared, such as that claims usually need some support to make up a 'full' argument (e.g., premises, evidence, or justifications), the exact definition of a claim can be arbitrary—depending on the domain, register, or task.

4 Methodology

Given the results from the qualitative analysis, we want to investigate whether the different conceptualizations of claims can be assessed empirically and if so, how they could be dealt with in practice. Put simply, the task we are trying to solve in the following is: *given a sentence, classify whether or not it contains a claim.* We opted to model the claim identification task on sentence level, as this is the only way to make all datasets compatible to each other. Different datasets model claim boundaries differently, e.g. MT includes discourse markers within the same sentence, whereas they are excluded in PE.

All six datasets described in the previous section have been preprocessed by first segmenting documents into sentences using Stanford CoreNLP (Manning et al., 2014) and then annotating every sentence as claim, if one or more tokens within the sentence were labeled as claim (or major claim in PE). Analogously, each sentence is annotated as non-claim, if none of its tokens were labeled as claim (or major claim). Although our basic units of interest are sentences, we keep the content of the entire document to be able to retrieve information about the context of (non-)claims.[3]

We are not interested in optimizing the properties of a certain learner for this task, but rather

[3]This is true only for the feature-based learners. The neural networks do not have access to information beyond individual sentences.

want to compare the influence of different types of lexical, syntactical, and other kinds of information across datasets.[4] Thus, we used a limited set of learners for our task: a) a standard L2-regularized logistic regression approach with manually defined feature sets[5], which is a simple yet robust and established technique for many text classification problems (Plank et al., 2014; He et al., 2015; Zhang et al., 2016a; Ferreira and Vlachos, 2016); and b) several deep learning approaches, using state-of-the-art neural network architectures.

The **in-domain experiments** were carried out in a 10-fold cross-validation setup with fixed splits into training and test data. As for the **cross-domain experiments**, we train on the entire data of the source domain and test on the entire data of the target domain. In the domain adaptation terminology, this corresponds to an unsupervised setting.

To address class-imbalance in our datasets (see Table 1), we downsample the negative class (non-claim) both in-domain and cross-domain, so that positive and negative class occur approximately in an 1:1 ratio in the training data. Since this means that we discard a lot of useful information (many negative instances), we repeat this procedure 20 times, in each case randomly discarding instances of the negative class such that the required ratio is obtained. At test time, we use the majority prediction of this ensemble of 20 trained models. With the exception of very few cases, this led to consistent performance improvements across all experiments. The systems are described in more detail in the following subsections. Additionally, we report the results of two baselines. The majority baseline labels all sentences as non-claims (predominant class in all datasets), the random baseline labels sentences as claims with 0.5 probability.

4.1 Linguistically Motivated Features

For the logistic regression-based experiments (LR) we employed the following feature groups. *Structure Features* capture the position, the length and the punctuation of a sentence. *Lexical Features* are lowercased unigrams. *Syntax Features* account for grammatical information at the sentence level. We include information about the part-of-speech and parse tree for each sentence. *Discourse Features* encode information extracted with help of the Penn Discourse Treebank (PDTB) styled end-to-end discourse parser as presented by Lin et al. (2014). *Embedding Features* represent each sentence as a summation of its word embeddings (Guo et al., 2014). We further experimented with sentiment features (Habernal and Gurevych, 2015; Anand et al., 2011) and dictionary features (Misra et al., 2015; Rosenthal and McKeown, 2015) but these delivered very poor results and are not reported in this article. The full set of features and their parameters are described in the supplementary material to this article. We experiment with the full feature set, individual feature groups, and feature ablation (all features except for one group).

4.2 Deep Learning Approaches

As alternatives to our feature-based systems, we consider three deep learning approaches. The first is the Convolutional Neural Net of Kim (Kim, 2014) which has shown to perform excellently on many diverse classification tasks such as sentiment analysis and question classification and is still a strong competitor among neural techniques focusing on sentence classification (Komninos and Manandhar, 2016; Zhang et al., 2016b,c). We consider two variants of Kim's CNN, one in which words' vectors are initialized with pre-trained GoogleNews word embeddings (CNN:w2vec) and one in which the vectors are randomly initialized and updated during training (CNN:rand). Our second model is an LSTM (long short-term memory) neural net for sentence classification (LSTM) and our third model is a bidirectional LSTM (BiLSTM).

For all neural network classifiers, we use default hyperparameters concerning hidden dimensionalities (for the two LSTM models), number of filters (for the convolutional neural net), and others. We train each of the three neural networks for 15 iterations and choose in each case the learned model that performs best on a held-out development set of roughly 10% of the training data as the model to apply to unseen test data. This corresponds to an early stopping regularization scheme.

5 Results

In the following, we summarize the results of the various learners described above. Obtaining all results required heavy computation, e.g. the cross-

[4] For the same reason, we do not optimize any hyperparameters for individual learners, unless explicitly stated.

[5] Using the liblinear library (Fan et al., 2008).

Target → System ↓	MT		OC		PE		VG		WD		WTP		Average	
neural network models														
BiLSTM	68.8	41.8	58.0	22.4	73.0	62.0	60.9	37.7	60.0	24.5	57.9	28.5	63.1	36.1
CNN:rand	78.6	67.3	**60.5**	**25.6**	**73.6**	61.1	**65.9**	**45.0**	61.1	25.8	58.6	28.9	66.4	42.3
CNN:w2vec	73.7	60.9	58.2	23.7	74.0	61.7	63.8	33.5	62.6	**28.9**	57.3	24.3	64.9	38.8
LSTM	65.2	48.3	58.5	22.3	71.8	60.7	61.3	40.1	61.6	25.9	58.0	28.4	62.7	37.6
LR	*feature ablation and combination*													
-Discourse	73.0	60.8	59.9	22.9	70.6	60.6	62.5	42.6	63.7	23.2	59.7	30.2	64.9	40.0
-Embeddings	74.6	62.9	59.6	22.6	70.4	60.4	62.9	43.1	63.9	23.5	59.4	29.9	65.1	40.4
-Lexical	72.1	59.5	59.6	22.5	65.9	55.1	60.8	40.5	60.1	18.5	57.7	27.8	62.7	37.3
-Structure	74.4	62.6	60.0	23.0	70.4	60.4	62.0	41.8	64.2	23.4	59.5	30.0	65.1	40.2
-Syntax	**79.8**	**70.3**	59.8	22.9	72.1	**62.5**	63.4	43.8	**65.1**	25.5	**60.1**	**30.5**	**66.7**	**42.6**
All Features	74.4	62.7	59.9	22.9	70.6	60.6	62.5	42.6	63.8	23.3	59.7	30.2	65.1	40.4
LR	*single feature groups*													
+Discourse	70.0	56.7	49.4	13.8	50.1	41.7	49.6	30.6	57.6	14.9	49.5	18.4	54.4	29.3
+Embeddings	72.4	59.8	58.8	20.8	68.2	57.7	59.7	39.3	64.2	23.8	59.0	28.9	63.7	38.4
+Lexical	75.9	64.7	59.5	21.4	71.8	62.1	61.1	40.5	64.0	22.2	59.0	27.7	65.2	39.8
+Structure	57.1	42.0	56.5	20.0	54.2	39.5	55.4	33.3	48.4	9.0	55.4	25.2	54.5	28.2
+Syntax	66.7	52.5	58.1	21.0	64.1	52.9	60.7	40.4	57.6	15.5	57.0	27.0	60.7	34.9
baselines														
Majority bsl	42.9	0.0	48.0	0.0	41.3	0.0	44.5	0.0	48.6	0.0	46.7	0.0	45.3	0.0
Random bsl	50.7	33.2	49.9	13.5	50.8	38.0	50.4	28.8	51.6	10.8	48.9	18.8	50.4	23.9

Table 2: In-domain experiments, best values per column are highlighted. For each dataset (column head) we show two scores: *Macro-*F_1 score (left-hand column) and F_1 score for claims (right-hand column).

validation experiments for feature-based systems took 56 days of computing. We intentionally do not list the results of previous work on those datasets. The scores are not comparable since we strictly work on sentence level (rather than e.g. clause level) and applied downsampling to the training data. All reported significance tests were conduced using two-tailed Wilcoxon Signed-Rank Test for matched pairs, i.e. paired scores of F_1 scores from two compared systems (Japkowicz and Shah, 2014).

5.1 In-Domain Experiments

The performance of the learners is quite divergent across datasets, with Macro-F_1 scores[6] ranging from 60% (WTP) to 80% (MT), average 67% (see Table 2). On all datasets, our best systems clearly outperform both baselines. In isolation, lexical, embedding, and syntax features are most helpful, whereas structural features did not help in most cases. Discourse features only contribute significantly on MT. When looking at the performance of the feature-based approaches, the most striking finding is the importance of lexical (in our setup, unigram) information.

The average performances of $\text{LR}_{-syntax}$ and CNN:rand are virtually identical, both for Macro-F_1 and Claim-F_1, with a slight advantage for the feature-based approach, but their difference is not statistically significant ($p \leq 0.05$). Altogether, these two systems exhibit significantly better average performances than all other models surveyed here, both those relying on and those not relying on hand-crafted features ($p \leq 0.05$). The absence or the different nature of inter-annotator agreement measures for all datasets prevent us from searching for correlations between agreement and performance. But we observed that the systems yield better results on PE and MT, both datasets with good inter-annotator agreement ($\alpha_u = 0.77$ for PE and $\kappa = 0.83$ for MT).

5.2 Cross-Domain Experiments

For all six datasets, training on different sources resulted in a performance drop. Table 3 lists the results of the best feature-based (*LR All features*) and deep learning (*CNN:rand*) systems, as well as single feature groups (averages over all source domains, results for individual source domains can be found in the supplementary material to this article). We note the biggest performance drops on the datasets which performed best in the in-domain setting (MT and PE). For the lowest scoring datasets, OC and WTP, the differences are only marginal when trained on a suitable dataset (VG

[6]Described as $Fscore_M$ in Sokolova and Lapalme (2009).

Target → Source/Sys. ↓	MT		OC		PE		VG		WD		WTP		Average	
	CNN:rand													
MT	*78.6*	*67.3*	51.0	7.4	56.9	22.1	57.2	15.7	52.4	9.4	49.4	10.9	53.4	13.1
OC	57.1	39.7	*60.5*	*25.6*	56.4	42.8	58.9	37.3	54.6	13.2	**58.4**	**28.9**	57.1	32.4
PE	59.8	18.0	54.2	9.5	*73.6*	*61.1*	57.5	18.7	**55.5**	**15.9**	54.7	16.0	56.3	15.6
VG	**68.7**	**51.5**	55.8	19.2	**57.0**	32.0	*65.9*	*45.0*	51.7	10.5	54.7	22.0	57.6	27.0
WD	64.4	3.5	51.3	1.3	41.3	0.0	44.5	0.0	*61.1*	*25.8*	46.7	0.0	49.6	1.0
WTP	58.5	26.6	56.8	15.4	56.0	18.5	55.3	19.4	52.9	11.6	*58.6*	*28.9*	55.9	18.3
Average	61.7	27.9	53.8	10.6	53.5	23.1	54.7	18.2	53.4	12.1	52.8	15.6	55.0	17.9
	LR All features													
MT	*74.4*	*62.7*	53.9	17.0	51.9	29.5	56.1	34.2	55.1	14.5	52.5	21.2	53.9	23.3
OC	60.0	45.1	*59.9*	*22.9*	56.7	**47.0**	58.6	**38.0**	54.1	12.2	57.7	27.5	57.4	**34.0**
PE	58.1	36.3	54.6	17.3	*70.6*	*60.6*	54.1	21.4	54.0	13.5	54.4	20.4	55.0	21.8
VG	65.8	51.4	**57.3**	**21.7**	**57.0**	45.1	*62.5*	*42.6*	54.5	13.1	55.1	24.8	**57.9**	31.2
WD	62.6	38.5	55.4	19.0	56.0	30.1	55.1	23.3	*63.8*	*23.3*	53.6	20.9	56.5	26.3
WTP	58.0	41.7	56.1	20.3	56.8	42.6	**59.1**	**38.0**	52.2	11.2	*59.7*	*30.2*	56.5	30.8
Average	60.9	42.6	55.5	19.1	55.7	38.9	56.6	31.0	54.0	12.9	54.7	23.0	56.2	27.9
LR	*single feature groups (averages across all source domains)*													
+Discourse	40.2	15.0	31.7	5.8	30.3	27.4	27.7	19.9	40.9	4.5	25.3	13.3	32.7	14.3
+Embeddings	56.6	35.2	51.4	12.8	53.6	30.7	53.3	24.3	54.2	13.2	52.9	19.0	53.7	22.5
+Lexical	61.0	42.2	55.2	18.3	56.2	38.6	54.7	29.1	53.1	11.9	54.9	23.4	55.9	27.2
+Structure	44.2	22.9	53.6	18.5	52.5	38.4	53.6	32.1	49.1	9.0	53.4	23.3	51.1	24.0
+Syntax	54.8	37.0	54.2	17.5	54.3	40.6	55.7	32.0	53.0	11.8	53.8	22.5	54.3	26.9
	baselines													
Majority bsl	42.9	0.0	48.0	0.0	41.3	0.0	44.5	0.0	48.6	0.0	46.7	0.0	45.3	0.0
Random bsl	47.5	30.6	50.5	14.0	51.0	38.4	51.0	29.3	49.3	9.3	50.3	20.2	49.9	23.6

Table 3: Cross-domain experiments, best values per column are highlighted, in-domain results (for comparison) in italics; results only for selected systems. For each source/target combination we show two scores: *Macro-*F_1 score (left-hand column) and F_1 score for claims (right-hand column).

and OC, respectively). The best feature-based approach outperforms the best deep learning approach in most scenarios. In particular, as opposed to the in-domain experiments, the difference of the Claim-F_1 measure between the feature-based approaches and the deep learning approaches is striking. In the feature-based approaches, on average, a combination of all features yields the best results for both Macro-F_1 and Claim-F_1. When comparing single features, lexical ones do the best job.

Looking at the best overall system (LR with all features), the average test results when training on different source datasets are between 54% Macro-F_1 resp. 23% Claim-F_1 (both MT) and 58% (VG) resp. 34% (OC). Depending on the goal that should be achieved, training on VG (highest average Macro-F_1) or OC (highest average Claim-F_1) seems to be the best choice when the domain of test data is unknown (we analyze this finding in more depth in §6). MT clearly gives the best results as target domain, followed by PE and VG.

We also performed experiments with mixed sources, the results are shown in Table 4. We did this in a leave-one-domain-out fashion, in particular we trained on all but one datasets and tested on the remaining one. In this scenario, the neural network systems seem to benefit from the increased amount of training data and thus gave the best results. Overall, the mixed sources approach works better than many of the single-source cross-domain systems – yet, the differences were not found to be significant, but as good as training on suitable single sources (see above).

6 Further Analysis and Discussion

To better understand which factors influence cross-domain performance of the systems we tested, we considered the following variables as potential determinants of outcome: similarity between source and target domain, the source domain itself, training data size, and the ratio between claims and non-claims.

We calculated the Spearman correlation of the top-500 lemmas between the datasets in each direction, see results in Table 5. The most similar domains are OC (source *s*) and WTP (target *t*), coming from the same authors. OC (*s*) and WD (*t*) as well OC (*s*) and VG (*t*) are also highly cor-

Target → System ↓	MT		OC		PE		VG		WD		WTP		Avg	
CNN:rand	62.8	41.4	**57.8**	**22.4**	**59.7**	36.2	**58.6**	28.1	**54.2**	**14.1**	**56.8**	25.6	**58.3**	28.0
All features	**64.7**	**49.5**	56.4	20.6	57.8	**45.8**	58.2	**36.4**	52.3	11.3	56.0	**26.0**	57.6	**31.6**
Majority bsl	42.9	0.0	48.0	0.0	41.3	0.0	44.5	0.0	48.6	0.0	46.7	0.0	45.3	0.0
Random bsl	47.5	30.6	50.5	14.0	51.0	38.4	51.0	29.3	49.3	9.3	50.3	20.2	49.9	23.6

Table 4: Leave-one-domain-out experiments, best values per column are highlighted. For each test dataset (column head) we show two scores: *Macro*-F_1 score (left-hand column) and F_1 score for claims (right-hand column).

	MT	OC	PE	VG	WD	WTP
MT	100	47	51	52	49	48
OC	56	100	55	68	71	71
PE	59	58	100	66	67	57
VG	51	58	52	100	59	62
WD	54	61	61	62	100	55
WTP	49	59	49	57	57	100

Table 5: Heatmap of Spearman correlations in % based on most frequent 500 lemmas for each dataset. Source domain: rows, target domain: columns.

related. For a statistical test of potential correlations between cross-domain performances and the introduced variables, we regress the cross-domain results (Table 3) on Table 5 (T4 in the following equation), on the number of claims #C (directly related to training data size in our experiments, effect of downsampling), and on the ratio of claims to non-claims R.[7] More precisely, given source/training data and target data pairs (s,t) in Table 3, we estimate the linear regression model

$$y_{st} = \alpha \cdot \mathrm{T4}_{st} + \beta \cdot \log(\#\mathrm{C}_s) + \gamma \cdot \mathrm{R}_t + \varepsilon_{st}, \quad (1)$$

where y_{st} denotes the Macro-F_1 score when training on s and testing on t. In the regression, we also include binary dummy variables $\mathbb{1}_\sigma = \mathbb{1}_{s,\sigma}$ for each domain σ whose value is 1 if $s = \sigma$ (and 0 otherwise). These help us identify "good" source domains.

The coefficient α for Table 5 is not statistically significantly different from zero in any case. Ultimately, this means that it is difficult to predict cross-domain performance from lexical similarity of the datasets. This is in contrast to e.g., POS tagging, where lexical similarity has been reported to predict cross-domain performance very well (Van Asch and Daelemans, 2010). The coefficient for training data size β is statistically significantly different from zero in three out of 15 cases. In particular, it is significantly positive in two (CNN:rand, CNN:w2vec) out of four cases for the neural networks. This indicates that the neural networks would have particularly benefited from more training data, which is confirmed by the improved performance of the neural networks in the mixed sources experiments (cf. §5.2). The ratio of claims to non-claims in t is among the best predictors for the variables considered here (coefficient γ is significant in three out of 15 cases, but consistently positive). This is probably due to our decision to balance training data (downsampling non-claims) to keep the assessment of claim identification realistic for real-world applications, where the class ratio of t is unknown. Our systems are thus inherently biased towards a higher claim ratio.

Finally, the dummy variables for OC and VG are three times significantly positive, but consistently positive overall. Their average coefficient is 2.31 and 1.90, respectively, while the average coefficients for all other source datasets is negative, and not significant in most cases. Thus, even when controlling for all other factors such as training data size and the different claim ratios of target domains, OC and VG are the best source domains for cross-domain claim classification in our experiments. OC and VG are particularly good training sources for the detection of claims (as opposed to non-claims)—the minority class in all datasets—as indicated by the average Claim-F_1 scores in Table 3.

One finding that was confirmed both in-domain as well as cross-domain was the importance of lexical features as compared to other feature groups. As mere lexical similarity between domains does not explain performance (cf. coefficient α above), this finding indicated that the learners relied on

[7] Overall, we had 15 different systems, see upper 15 rows in Table 2. Therefore, we had 15 different regression models.

a few, but important lexical clues. To go more into depth, we carried out error analysis on the CNN:rand cross-domain results. We used OC, VG and PE as source domains, and MT and WTP as target domains. By examining examples in which a model trained on OC and VG made correct predictions as opposed to a model trained on PE, we quickly noticed that lexical indicators indeed played a crucial role. In particular, the occurrence of the word "should" (and to a lower degree: "would", "article", "one") are helpful for the detection of claims across various datasets. In MT, a simple baseline labeling every sentence containing "should" as claim achieves 76.1 Macro-F_1 (just slightly below the best in-domain system on this dataset). In the other datasets, this phenomenon is far less dominant, but still observable. We conclude that a few rather simple rules (learned by models trained on OC and VG, but not by potentially more complex models trained on PE) make a big difference in the cross-domain setting.

7 Conclusion

In a rigorous empirical assessment of different machine learning systems, we compared how six datasets model claims as the fundamental component of an argument. The varying performance of the tested in-domain systems reflects different notions of claims also observed in a qualitative study of claims across the domains. Our results reveal that the best in-domain system is not necessarily the best system in environments where the target domain is unknown. Particularly, we found that mixing source domains and training on two rather noisy datasets (OC and VG) gave the best results in the cross-domain setup. The reason for this seem to be a few important lexical indicators (like the word "should") which are learned easier under these circumstances. In summary, as for the six datasets we analyzed here, our analysis shows that the essence of a claim is not much more than a few lexical clues.

From this, we conclude that future work should address the problem of vague conceptualization of claims as central components of arguments. A more consistent notion of claims, which also holds across domains, would potentially not just benefit cross-domain claim identification, but also higher-level applications relying on argumentation mining (Wachsmuth et al., 2017). To further overcome the problem of domain dependence, multi-task learning is a framework that could be explored (Søgaard and Goldberg, 2016) for different conceptualizations of claims.

Acknowledgments

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText), by the GRK 1994/1 AIPHES (DFG), and by the ArguAna Project GU 798/20-1 (DFG).

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, Berlin, Germany.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis WASSA 2011*, pages 1–9, Portland, OR, USA.

Or Biran and Owen Rambow. 2011a. Identifying justifications in written dialogs. In *Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 162–168, Palo Alto, CA, USA.

Or Biran and Owen Rambow. 2011b. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer, Berlin/Heidelberg.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, page to appear, Vancouver, Canada.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, OR, USA.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, CA, USA.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520, Bellevue, WA, USA.

Trudy Govier. 2010. *A Practical Study of Argument*, 7th edition. Wadsworth, Cengage Learning.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting Embedding Features for Simple Semi-supervised Learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 110–120, Doha, Qatar.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal.

Nathalie Japkowicz and Mohak Shah. 2014. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, Cambridge, UK.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1490–1500, San Diego, California.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1489–1500, Dublin, Ireland.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2):151–184.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 185–191, Buenos Aires, Argentina.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MA, USA.

Amita Misra, Pranav Anand, Jean Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 430–440, Denver, CO, USA.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107, Barcelona, Spain.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230, Stanford, CA, USA.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536, Denver, CO, USA.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, USA.

Huy Nguyen and Diane Litman. 2016. Improving argument mining in student essays by learning and

exploiting argument indicators versus essay topics. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pages 485–490, Key Largo, FL, USA.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760, Raleigh, NC, USA.

Dae Hoon Park and Catherine Blake. 2012. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 1–9, Jeju, Republic of Korea.

Andreas Peldszus and Manfred Stede. 2013a. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pages 801–815, Lisbon, Portugal.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 543–552, Beijing, China.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 507–511, Baltimore, MA, USA.

Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2613–2618, Marrakech, Morocco.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, pages 272–275, Marco Island, FL, USA.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, Washington, DC, USA.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic.

Edward Schiappa and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*, 1st edition. Pearson UK.

Tobias Schnabel and Hinrich Schütze. 2013. Towards robust cross-domain domain adaptation for part-of-speech tagging. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 198–206, Nagoya, Japan.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 231–235, Berlin, Germany.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, pages in press, preprint available at arXiv:1604.07370.

Mihai Surdeanu, Ramesh Nallapati, and Christopher D. Manning. 2010. Legal claim identification: Information extraction with hierarchically labeled data. In *Workshop on the Semantic Processing of Legal Texts at LREC*, pages 22–29, Valletta, Malta.

Stephen E. Toulmin. 2003. *The Uses of Argument, Updated Edition*. Cambridge University Press, New York.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017. "PageRank" for Argument Relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 672–682, Denver, CO, USA.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016a. ArgRewrite: A Web-based Revision Assistant for Argumentative Writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 37–41, San Diego, CA, USA.

Rui Zhang, Honglak Lee, and Dragomir R. Radev. 2016b. Dependency sensitive convolutional neural networks for modeling sentences and documents. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1512–1521, San Diego, CA, USA.

Ye Zhang, Stephen Roller, and Byron C. Wallace. 2016c. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1522–1527, San Diego, CA, USA.