Decompositional Argument Mining: A General Purpose Approach for Argument Graph Construction

Debela Gemechu and Chris Reed
Centre for Argument Technology, School of Science & Engineering
University of Dundee, Dundee, UK
d.t.gemechu@dundee.ac.uk, c.a.reed@dundee.ac.uk

Abstract

This work presents an approach decomposing propositions into four functional components and identify the patterns linking those components to determine argument structure. The entities addressed by a proposition are **target concepts** and the features selected to make a point about the target concepts are **aspects**. A line of reasoning is followed by providing evidence for the points made about the target concepts via aspects. **Opinions on target concepts** and **opinions on aspects** are used to support or attack the ideas expressed by target concepts and aspects. The relations between aspects, target concepts, opinions on target concepts and aspects are used to infer the argument relations. Propositions are connected iteratively to form a graph structure. The approach is generic in that it is not tuned for a specific corpus and evaluated on three different corpora from the literature: AAEC, AMT, US2016G1tv and achieved an F score of 0.79, 0.77 and 0.64, respectively.

1 Introduction

Argument mining is the process of identifying argumentative structure contained within a text. It involves segmenting arguments into elementary discourse units (EDUs), distinguishing argumentative units from non-argumentative units, classifying argument components into classes such as premise and claim, identifying and labeling argument relations between the components, and identifying argument schemes. We are here aimed at mining argument structure from text segmented into EDUs (or, more precisely for argument mining, Argumentative Discourse Units, ADUs (Peldszus and Stede, 2015)).

Several argument mining approaches use features identified from individual EDUs and apply classifiers (Moens et al., 2007); others use features that span EDUs and apply dependency parsing (Muller et al., 2012), similarity (Lawrence et al., 2014), linguistic indicators (Villalba and Saint-Dizier, 2012) and their combinations (Lawrence and Reed, 2015). Recently, a neural end-to-end method for argument mining shows that dependency parsing outperforms an EDU-level classifier (Eger et al., 2017). Stab and Gurevych (2014b) use both EDU-level and cross-EDU features to improve performance. The EDU-spanning features used by these latter approaches include syntactic dependency and lexical overlap between the EDUs. For instance, Eger et al. (2017) applied token level syntactic dependency to learn the relations between EDUs. Even though cross-EDU tokens are used for argument mining, the nature of such tokens is not studied well.

Following the same line of reasoning, similarity approaches use EDU level similarity to determine argument structure. Lawrence et al. (2014) use Latent Dirichlet Allocation (LDA) topic modeling; Lawrence and Reed (2015) use WordNet[1] Synset hierarchy to determine similarity between propositions. Such approaches start from a conclusion and determine the most related proposition to create hierarchical graph structure based on the assumption that a conclusion is similar to a premise. Similarity, however, does not necessarily entail an argument relation and vice-versa.

In this work, we aim to detect argument relations (AR) and their category (support vs attack) based on the nature of the relations existing among the functional components of propositions. The functional components of propositions are: target concepts (***C***), aspects (***A***), opinions on aspects (***OA***) and opinions on target concepts (***OC***). In order to identify ARs and their category, we train classifiers using the relations between the four components. The classifiers provide an output pre-

[1] http://wordnet.princeton.edu/

dicting whether any pair of propositions involve an AR or not, and categorize the AR.

To the best of our knowledge there is no approach that decomposes propositions into fine-grained components and uses them to determine argument structure. Our **Decompositional Argument Mining** (DAM) identifies argument structure by exploiting similarity (between ***C*** and ***A***) and relations between the polarities of ***OC*** and ***OA***. Our first hypothesis is then the AR between EDUs is governed by the relations between their functional components. For instance, the support relation between (2) and (9) from Table 1 is a function of the similarity between ***C*** of (9) *"cooking; potato; burger"* and A of (2) *"food"* and the agreement between the polarities of their respective opinion expressions (i.e. the opinions *"have an opportunity; interesting"* and *"better"* are both positive). Similarly the support relation between (6) and (7) is the function of the similarity between ***A*** of (6) *"job"* and ***C*** of (7) *"job"* and the agreement between the polarities of their respective opinion expressions (i.e. *"are losing"* and *"are fleeing"* are both negative). The attack relation between (10) and (11) is the function of the similarity between ***C*** of (10) *"advertising"* and ***A*** of (11) *"advertising"* and the contradiction between the polarities of the opinion on ***A*** of (10) *"should be prohibited"* and the opinion on ***C*** of (11) *"needs"*.

Our second hypothesis is that automatic recognition of argument structure can be substantially enhanced by using the relations between the four functional components of propositions as compared to other features like discourse indicators which are rare to find. For instance, none of the propositions presented in the example are linked via discourse indicators, and yet the relations between the four components can be used as a basis for identifying their ARs. The third hypothesis is that fine-grained similarity is more reliable and accurate than EDU level similarity. The similarity between the entirety of propositions is not a good indicator of AR. For instance the similarity between (3) and (8) is 0.737 (as provided by ADW (Pilehvar et al., 2013)) and yet does not involve an AR, but (8) and (1) has a similarity score of 0.45 and involves an AR since there is a strong similarity between the aspect of (1) *"family"* and target concept of (8) *"family"*.

The contribution of this work is three-fold: (a) a model to identify components linking propositions; (b) directional similarity indicating the direction of AR between propositions; (c) an approach determining the entire argument structure based on just the relations between the four functional components of proposition across three heterogeneous corpora of which two are monological and the other is dialogical (see Section 3).

2 Argument Graph Model

A proposition in the Frege's sense, is decomposed into four functional components: ***C***, ***A***, ***OC*** and ***OA***. ***C*** and ***A*** are used to link a premise and a conclusion; the polarity of ***OC*** and ***OA*** is used to identify the type of relations (inference vs conflict).

2.1 Functional Decomposition of a Proposition and their relations

We define the four functional components of a proposition before formalizing the representation of proposition in terms of the components. Examples (4) to (7) in Table 1 are taken from the first US 2016 presidential election television debate corpus (US2016G1tv) (Lawrence and Reed, 2017; Visser et al., 2019) and (1) to (3), (8) to (13) are taken from the Argument Annotated Essay Corpus (AAEC) (Stab and Gurevych, 2014a) to illustrate the components.

2.1.1 Target Concept (*C*)

A proposition makes a point about (at least one) concept: an idea, physical or abstract entity, following (Lima et al, 2010):

> "Concepts, also known as classes, are used in a broad sense. They can be abstract or concrete, elementary or composite, real or fict[it]ious. In short, a concept can be anything about which something is said, and, therefore, could also be the description of a task, function, action, strategy, reasoning process, etc." (Lima et al., 2010, p:428).

The set of concepts addressed by a proposition are referred to as target concepts, (***C***). The examples in Table 1 are annotated to show C (segmented with [], and marked by the subscript *c* and also shown in bold for convenience). (1) and (2) address the target concept (after stemming) *"camp"*, whilst the targets concepts in (3) are *"family"* and *"camp"*. The target concept is analogous to a topic of a propositions and usually presented as a subject of a proposition. Aspects specialize the topic of a proposition by providing specific angle of reasoning.

No	Example
1	**[Camping]**$_c$ [*is a great way*]$_{oc}$ to [bring]$_{oa}$ **[families]**$_a$ [together]$_{oa}$
2	**[Campers]**$_c$ [*have an opportunity*]$_{oc}$ to try some [interesting]$_{oa}$ **[food]**$_a$
3	When **[families]**$_c$ go **[camping]**$_c$, they put the **[jobs]**$_a$ and **[sporting events]**$_a$ [*on hold*]$_{oa}$
4	**[Housing]**$_c$ [*did collapse*]$_{oc}$
5	**[These countries, especially China]**$_c$, [*are taking*]$_{oc}$ **[Americans' jobs]**$_a$
6	**[We]**$_c$, [*are losing*]$_{oc}$ [our]$_a$ [*good*]$_{oa}$ **[jobs]**$_a$ so many of them
7	**[Our jobs]**$_c$, [*are fleeing*]$_{oc}$ **[the country]**$_a$
8	By putting aside these events, the **[family]**$_c$ [*has an opportunity*]$_{oc}$ to [bond]$_{oa}$ their [relationships]$_a$
9	**[Cooking]**$_c$ over a fire makes **[burgers]**$_c$ and **[potatoes]**$_c$ [*taste better*]$_{oc}$ than can be found at **[fast]**$_a$**[food]**$_a$**[place]**$_a$
10	**[Advertising]**$_c$ **[alcohol]**$_a$, **[cigarettes]**$_a$, **[goods]**$_a$ and **[services]**$_a$ with **[adult content]**$_a$ [*should be prohibited*]$_{oc}$
11	**[Modern society]**$_c$ [*needs*]$_{oc}$ **[advertising]**$_a$
12	**[Ads]**$_c$ will [*keep*] us [*well informed*]$_{oc}$ about [*new*]$_{oa}$ **[products]**$_a$ and **[services]**$_a$
13	**[advertising]**$_c$ **[cigarettes]**$_a$ and **[alcohol]**$_a$ [*will definitely affect*]$_{oc}$ our children [*in negative way*]$_{oc}$

Table 1: Examples to illustrate the four functional components of a proposition: ***C***, ***A***, ***OC*** and ***OA***. (In the online version, positive and negative polarity is indicated in blue and red, respectively).

2.1.2 Aspect (*A*)

Often, a specific angle of reasoning is selected to make a point about ***C***. The concepts providing such angles of reasoning are denoted as aspects (***A***). For instance, (1) and (2) address the target concept *"camp"* with respect to the aspects *"family"* and *"food"* (in bold) respectively. Similarly, the aspects of (3) are *"job, sporting event"*. The difference between ***C*** and ***A*** is not an ontological distinction, it is rather the syntactic and semantic role they play in the respective propositions. An aspect in one proposition can be a target concept in another (see (1) and (3)).

2.1.3 Opinion on Target Concept (*OC*)

OC is an opinion expressed on ***C*** to express positive or negative attitudes. The opinionated words in a proposition are usually ambiguous and do not fall into the conventional opinionated words category. For instance, in (5), the opinion *"are taking"*, which is expressed on the target concept *"country, china"*, does not fall into the conventional opinionated word category.

2.1.4 Opinion on Aspect (*OA*)

OA is an opinion expressed on an ***A*** to provide positive or negative attitudes. For instance, in (2) the opinion *"interesting"* is expressed on the aspect *"food"*.

Since we have defined the four components of a proposition, we can now formalise the representation of a proposition in terms of the components. Hence, a proposition, p, can be represented as a set of tuples,

$$P = \{\langle C_0, o_{C0}, \{\langle A_0, o_{A0}\rangle, \cdots, \langle A_i, o_{Ai}\rangle\}\rangle, \langle C_1, o_{C1}, \{\langle A_1, o_{A1}\rangle, ..., \langle A_j, o_{Aj}\rangle\}\rangle, \cdots \langle C_n, o_{Cn}, \{\langle A_j, o_{Aj}\rangle, \cdots, \langle A_k, o_{Ak}\rangle\}\rangle\} \quad (1)$$

Where, C_i, A_i, o_{Ci}, o_{Ai} represents ***C***, ***A***, ***OC*** and ***OA***, respectively.

2.1.5 The Relations Between the Four Functional Components

The relations between the four components fall into two categories: similarity and agreement. The relation between ***C*** and ***A*** is similarity whereas agreement (or contradiction) between ***OC*** and ***OA***. The relations between ***C*** and ***A*** are further categorized into four: (a) similarity between ***C*** of a premise and a conclusion, (b) similarity between ***A*** of a premise and a conclusion, (c) similarity between ***A*** of a premise and ***C*** of a conclusion, and (d) similarity between ***A*** of a conclusion and ***C*** of a premise. The relations between ***OC*** and ***OA*** are also categorized into four: (a) the agreement between ***OC*** of a conclusion and a premise, (b) the agreement between ***OA*** of a premise and a conclusion, (c) the agreement between ***OC*** of a conclusion and ***OA*** of a premise, and (d) the agreement between ***OA*** of a conclusion and ***OC*** of a premise.

2.2 Argument Relation

The argument relation (*AR*) between a premise and a conclusion is a function of the relations between the four components. A classifier is trained on the relations between the four components to identify the patterns encoded by the type of AR:

Inference relations: A pair of propositions involving support relation.
Conflict relations: A pair of propositions involving attack relation.

To mention, when a premise develops one or more aspects of a conclusion, the aspects of a conclusion form ***C*** of a premise (i.e are highly similar). For instance, (8) supports (1) in relation to the aspect *"family"*; (9) supports (2) in relation to the aspect *"food"*. The relation between ***OC*** and ***OA*** is identified through matching the polarity of the opinions. For instance the polarity of the opinions on (1 and 8) matches (both are positive), since the propositions involve support relation. Similarly, the attack relation between (10) and (11) is indicated by the similarity between ***C*** of (10) and ***A*** of (11) and the contradiction between the polarities of the opinions on ***OC*** of (10) and ***OC*** of (11).

Accordingly, the AR between propositions is defined by,

$$\mathcal{AR} = \begin{cases} S & \text{if } rel(C, A, OC, OA) = \theta \\ AT & \text{if } rel(C, A, OC, OA) = \beta \\ N & \text{otherwise} \end{cases} \quad (2)$$

where, S stands for support, AT for attack and N for none, while θ, β representing the result of a classifier (θ for support and β for attack).

A graph structure is formed to represent an argument by linking proposition whose components are related via the valid relations encoded by AR. Propositions and the relations between them are nodes, the connections between the nodes form the edges. Figure 1 shows an argument structure for a portion of propositions in Table 1, where (11) is attacking (10), (12) is supporting (11), (13) is attacking (11), and (13) is supporting (10) based on the similarity between ***C*** and ***A*** and the agreement between the polarities of the opinion expressions on ***C*** and ***A***.

3 Methodology

In this section, we present the data-sets and the major components of our approach.

3.1 Data

We aim to cover varieties of data-sets (though not comprehensive), annotated based on the underlying set of argumentation theory to see how our approach behaves across heterogeneous data-sets without tuning to a specific data-set. we use three

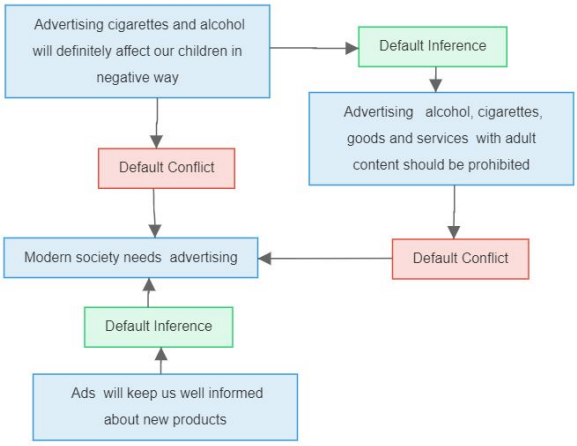

Figure 1: Argument structure for propositions (10), (11), (12), and (13) from Table 1

corpora, with different types of source material (monologue, dialogue), different creation rubrics (naturally occurring, created under direction), different argument structure conventions (recursive, limited), different notions of inference (typed, untyped) and different notions of conflict (rebut-only, rebut and undercut).

The first is Argument Annotated Essay Corpus (AAEC) (Stab and Gurevych, 2014a) which has a total of 90 arguments. Propositions under each argument are labelled as premise, claim or major claim. The corpus has 31,194 tokens, 1,552 propositions and 1214 Argument relations (AR). The second corpus is the Argumentative Micro Text (AMT) (Peldszus and Stede, 2013) which is a collection of 112 short texts collected from human subjects in German and were translated into English. It is annotated following the argumentation structure outlined by Peldszus and Stede (2013) and attain high inter-annotator agreement score. The structure consists of a central claim, and support/attack propositions. It has a total of 8,007 tokens, 576 propositions and 272 argument relations. We have also used dialogical corpus from the first US 2016 presidential election television debate between the candidates Clinton and Trump (US2016G1tv) (Lawrence and Reed, 2017; Visser et al., 2019) which is annotated based on AIF (Chesnevar et al., 2006) using the OVA+ annotation tool (Janier et al., 2014)[2] and stored in the AIFdb database (Lawrence et al., 2015). The corpus has a total of 15,805 tokens, 1,473 propositions and 505 inferences.

In addition to the original annotation, we anno-

[2] http://ova.arg-tech.org

tate ***C***, ***A***, ***OC*** and ***OA***. We obtain the total of 3,455, 4,113, 4,359, and 2,987 ***C***, ***A***, ***OC*** and ***OA*** respectively. For the corpus evaluation, a second annotator analysed 10% of claim-premise pairs form the combined corpora. To this end, we combine the three corpora and randomly select 10% of claim-premise pairs and provide it to the second annotator after removing the annotation labels of the first annotation. The annotation of the four components is compared against the original annotation to calculate the inter-annotators agreement. This gave a Cohen's kappa score $\kappa = 0.86$, $\kappa = 0.82$, $\kappa = 0.81$, and $\kappa = 0.80$ on ***C***, ***A***, ***OC*** and ***OA***, respectively. The annotation of the second annotator is discarded after calculating the Cohen's kappa score. The description of the annotation process and guideline is available online [3].

3.2 Identifying Argument Structure

Our approach involves a pipeline of four steps: Given segmented argumentative text, the first step identifies ***C***, ***A***, ***OC*** and ***OA***. The similarity component determines the degree of similarity between C and A. The next step identifies the polarity of the opinions to determine if they contradict or agree. The last component uses the similarity between ***C*** and ***A***, and the relation between ***OC*** and ***OA*** (contradiction or agreement) to link propositions and iteratively construct a graph. The details are provided below.

3.2.1 Identifying Aspects, Target Concepts and Opinions

We formulate the task in two ways: relation extraction task adapted from information extraction, and a sequence labeling task adapted from aspect based opinion mining.

C*, *A*, *OC* and *OA* identification as a relation extraction task.** We model it as a relation extraction task since ***C, ***A***, ***OC*** and ***OA*** are syntactically interdependent. Relation extraction has been studied extensively in natural language processing using supervised methods (Kambhatla, 2004; Zhao and Grishman, 2005) and semi-supervised methods (Etzioni et al., 2005; Banko et al., 2007). Supervised methods use classification techniques: Maximum Entropy Models (Borthwick et al., 1998), Hidden Markov Models (Bikei et al., 1997), Support Vector Machines (Asahara and Matsumoto, 2003), and Conditional Random Fields (McCallum and Li, 2003).

Following the same line of reasoning, we train four classifiers (Naive Bayes, CRF, bag of features based SVM, and tree kernel based SVM) to classify the words in a proposition as ***C***, ***A***, ***OC*** or ***OA***. The first three classifiers use frequency, part of speech category and universal dependency as classification features. The tree kernel SVM is trained using the portion of the dependency tree connecting the four components as positive examples and the rest as negative examples.

C*, *A*, *OC* and *OA* identification as a sequence labeling task.** The sequence labeling model is adapted from aspect based opinion mining. Aspect based opinion mining identifies opinions expressed on a target object and specific aspects of the object (Zhang and Liu, 2014). Taking the analogy of target object:aspects in opinionated text to ***C:***A***:***OC***:***OA*** in argumentative text, we apply similar techniques for identifying ***C***, ***A***, ***OC*** and ***OA***.

The underlying idea behind the model is that ***C***, ***A***, ***OC*** and ***OA*** are interdependent and occur in a sequence in a sentence. The model is based on the Inside-Outside-Begin (IOB) labelling schema (Ramshaw and Marcus, 1999). Accordingly, we use the IOB labeling schema where, B-Concept denotes the beginning of a concept; I-Concept, denotes that the token is inside the concept, and O for other (non ***C***, ***A***, ***OC*** or ***OA***) tokens. Hidden Markov Models (HMM) (Jin et al., 2009), Conditional Random Fields (CRF) (Sminchisescu et al., 2006) and recently, convolutional neural networks (CNN) (Poria et al., 2016) are common techniques employed. The assumption that an observation only depends on the current state and that a given state depends on its immediate predecessor state made HMM approaches less applicable for relations involving long distance dependencies. CRF is also a linear model and suffers from the same criticism as HMM. CNN on the other hand can encode long distance relations existing between concepts. As a result, we use CNN to train the model since ***C***, ***A***, ***OC*** and ***OA*** can appear a long way away from each other.

3.2.2 Identifying the contradiction between opinions

Our aim here is to compare the polarities between ***OC*** and ***OA*** to check if they match or contradict.

The opinionated words in our case are context dependent ("are taking our jobs" vs "are taking our

[3] http://arg.tech/~debela/Guidelines.pdf

presents") and often the contexts are fine-grained (see example 5). We aim to disambiguate the sentiment orientation of the words via identifying *constrained synonyms* (CS). Constrained synonyms are subset of synonyms expressing similar sense to the current opinion word in a fine-grained context. For instance, among the synonyms of "taking" in (5), we are interested to identify synonyms like "robbing" and "stealing", constrained by a given context like "China, Americans jobs". Our hypothesis is then the use of such CS expressing a similar opinion can improve the estimation of the polarity of ambiguous opinion words by aggregating the information coming from multiple words expressing a similar opinion to the current opinion. In order to identify the CS, we enhance word embedding to enforce the encoding of fine-grained contexts.

Our Context Sensitive Polarity Prediction (CSPP) technique consists of two main components: identifying CS and predicting polarity using the CS.

To identify the CS, we extend CBOW based Word2Vec (Tomas et al., 2013) (see Equation 3). Accordingly, given a fine-grained context, the extended CBOW predict CS for an opinion word in the context. We use ***C*** and ***A*** as a fine-grained context of the opinion and encode them in the representation of words. The embedding is extended by introducing an additional output layer (called the *constrained context*, CC, output layer) to update the embedding based on the fine-grained contexts. The two output layers are connected to the previous layer in the network and the cost function is the loss of the first plus the second output. Given a sequence of words W=$\{w_1, ..., w_N\}$, the Constrained Embedding (CE) objective function is defined by the formula in Equation 4.

$$CBOW(W) = \frac{1}{N}\sum_{i=1}^{N} \log \mathrm{P}(w_i \mid gc_{wi}) \tag{3}$$

$$CE(W) = \frac{1}{N}\sum_{i=1}^{N} \log \mathrm{P}(w_i \mid gc_{wi}) + \log \mathrm{P}(w_i \mid cc_{wi}) \tag{4}$$

where d is the number of fine-grained context which is equivalent to the number of target concepts and aspects; gc_{wi} indicates the global contexts identified by taking $d/2$ words to the left and right of w_i ($d/2$ words to the left and right of the current word is taken to equalize the number of global context with the number of fine-grained context); cc_{wi} is given by the aggregation of fine-grained and global context (gc_{wi}) using Equation 5. Given an input sequence $w_i, w_{i+1}, ...w_n$, and fine-grained context $c_j, c_{j+1}, ...c_d$, the function which aggregates both contexts to produce (cc_{wi}) for the current word w_i is given by:

$$\begin{aligned} cc_{wi} = [&e\vec{w}^T_{i-d/2}([e\vec{c}^T_j, e\vec{c}^T_{j+1}, e\vec{c}^T_{j+2}..., e\vec{c}^T_d]), ... \\ &e\vec{w}^T_{i-1}([e\vec{c}^T_j, e\vec{c}^T_{j+1}, e\vec{c}^T_{j+2}..., e\vec{c}^T_d]), \\ &e\vec{w}^T_{i+1}([e\vec{c}^T_j, e\vec{c}^T_{j+1}, e\vec{c}^T_{j+2}..., e\vec{c}^T_d]), ..., \\ &e\vec{w}^T_{i+d/2}([e\vec{c}^T_j, e\vec{c}^T_{j+1}, e\vec{c}^T_{j+2}..., e\vec{c}^T_d])] \end{aligned} \tag{5}$$

where, $e\vec{w}^T_{i-d/2}, ..., e\vec{w}^T_{i-1}, e\vec{w}^T_{i+1}, ..., e\vec{w}^T_{i+d/2}$ are the transpose of pre-trained vectors of the global contexts of the current word w_i and $e\vec{c}^T_j, e\vec{c}^T_{j+1}, e\vec{c}^T_{j+2}, ..., e\vec{c}^T_d$ are the transpose of pre-trained vectors of the d sized fine-grained contexts.

Once the CS are identified for the current opinion word using the extended word embedding, we train a classifier to categorize the polarity, given a classification feature including the initial list of opinion words generated by Hu and Liu (Hu and Liu, 2004), the current opinion word, the CS and paragraphs containing the opinion words and the CS.

3.2.3 Computing Similarity

Similarity between C and A is used to connect propositions. In addition to aspect based, we have tried proposition level similarity for comparison:

1. **Proposition level similarity.** Computes similarity between the entirety of propositions.

2. **Aspect Based Similarity.** Computes the similarity between aspects and target concepts.

We used two state of the art similarity approaches allowing to measure the similarity between any text fragment at various linguistic levels: Align Disambiguate Walk (ADW) (Pilehvar et al., 2013) and Doc2vec (Le and Mikolov, 2014). ADW is a graph-based approach for measuring the semantic similarity of linguistic items at various levels (word senses, texts). To measure the similarity between words, ADW starts by disambiguating them using the context in which the words are used based on their WordNet representation. Doc2vec (Le and Mikolov, 2014) is an enhanced version of word2vec (Mikolov et al., 2013) that

allows for computing similarity between phrases, sentences, paragraphs or documents.

3.2.4 Identify Argument Relations and Category

A classifier is trained to learn the relations between the four components in order to link propositions. The classification features are: the similarity between ***C*** and ***A***; the relation between ***OC*** and ***OA***. To facilitate the training, we convert the continues similarity values (which ranges from 0.0 to 1.0) to a discrete value by tuning a threshold α on a development set to categorize them into two: unrelated or similar. Likewise, the relation between ***OC*** and ***OA*** holds discreet values: agreement, disagreement or neutral.

3.2.5 Iterative Graph Construction

Given a set of propositions, we build a structure consisting the valid ARs holding between the propositions. Propositions and ARs are nodes and the links between them form edges.

We start with any arbitrary proposition P_i and then identify the associated functional components. The similarity between ***C*** and ***A*** of P_i and all the other propositions ($P_{i+1...n}$); the agreement between ***OC*** and ***OA*** of P_i and all the other remaining propositions ($P_{i+1...n}$) are identified. A classifier is then used to identify the AR between the propositions based on the relations between their components. Accordingly, a proposition whose functional components are related with the functional components of P_i is connected to P_i to form an edge ($P_{i+1} \rightarrow P_i$). Once all the child nodes (all the premises) are connected, the proposition is marked as visited. Continuing with the next unvisited proposition, the same procedure is applied until all the propositions in the entire argument are visited.

4 Experiments

Four machine learning approaches are trained to detect ***C***, ***A***, ***OC*** and ***OA***. Two similarity approaches are tried to identify similarity between ***C*** and ***A***. CSPP is tried to identify the polarity of ***OC*** and ***OA***. Our DAM combines the best performing component identifier, similarity and the CSPP to train a classifier in order to identify AR existing between proposition. The implementation of our approach is available online [4]. It takes argumentative text as an input and returns the argument structure using AIF-JSON (Chesnevar et al., 2006) format.

[4] http://ws.arg.tech/

4.1 Evaluation technique and setup

We use ten-fold cross-validation, where the dataset is randomly divided into ten groups. Arguments are randomly split into 80% training and 20% test sets with the same class distribution. To balance the class distribution (composition of premise, conclusion, attack relation, and support relation), we follow the unitization in the respective corpus. For instance, AAEC is originally presented as 90 self contained essays consisting of conclusions, premises and the associated argument relations. Hence, we consider an argument as a unit to take all the constituted elements at a time. We report average precision, recall and F-measure computed by ten-fold cross-validation over these units.

4.2 Results and Discussions

We present the results of the individual components separately:
C, ***A***, ***OC*** and ***OA*** extraction. The four classifiers are evaluated on the three corpora as presented in Table 2. We use the class distribution of the components as a baseline. We divide the number of ***C*** and ***A*** by the total number of concepts (***C*** and ***A***) to obtain the class distribution for ***C*** and ***A***. The same procedure is followed for the opinions (***OC*** and ***OA***). The sequential labeling approach out-performed all the classifiers and the baseline across the corpora. The syntactic dependency existing between ***C***, ***A***, ***OC*** and ***OA***, regardless of the distance existing between them, is recognized by the CNN more reliably than the other classifiers. The kernel-based SVM outperformed the feature based SVM which is again attributed to its ability of encoding the syntactic dependency linking the target concepts and the aspects.

CSPP. We use SemEval data-sets (Rosenthal et al., 2017) to evaluate CSPP. We compare the result against an implementation using conventional word embedding as a baseline. CSPP achieves an overall F-measure of 0.79 while the baseline achieves 0.71. The strength of CSPP is founded on its use of multiple words expressing similar senses as the current opinion (in similar context) to gather several instances of the current ambiguous words to increases the chance of prediction.

	Data-Sets											
	AAEC				AMT				US2016G1tv			
Approaches	C	A	OC	OA	C	A	OC	OA	C	A	OC	OA
Baseline	0.45	0.55	0.57	0.43	0.48	0.52	0.6	0.4	0.43	0.57	0.61	0.39
SVM-kernel	0.82	0.71	0.81	0.62	**0.78**	0.65	0.69	0.65	0.77	0.69	**0.69**	0.66
SVM-feature	0.81	0.70	0.81	0.65	0.75	0.68	0.67	0.66	0.76	0.69	0.67	0.66
CNN-Sequence	**0.83**	**0.72**	**0.82**	**0.7**	0.77	**0.69**	**0.7**	0.67	**0.78**	**0.71**	0.68	**0.67**
CRF	0.80	0.69	0.72	0.65	**0.78**	0.67	0.66	**0.69**	0.76	0.67	0.67	**0.67**
Naive Bayes	0.79	0.69	0.76	0.66	0.75	0.62	0.62	0.62	0.75	0.66	0.65	0.64

Table 2: The performance (F-measure) of ***C***, ***A***, ***OC*** and ***OA*** extraction on AAEC, AMT and US2016G1tv corpus

		Approaches											
		S&G2014b			P&S2016			PLS			DAM		
Data-Sets	Components	P	R	F	P	R	F	P	R	F	P	R	F
AAEC Para	Propositions	0.77	0.68	0.73	n/a			n/a			**0.81**	**0.77**	**0.79**
	AR	0.74	0.71	0.72				0.62	0.67	0.64	**0.82**	**0.76**	**0.79**
	ARC	0.74	0.71	0.72				n/a			**0.81**	**0.74**	**0.77**
AAEC Essay	Propositions	n/a			n/a			n/a			**0.76**	**0.73**	**0.74**
	AR							0.58	0.7	0.63	**0.73**	**0.75**	**0.74**
	ARC							n/a			**0.73**	**0.74**	**0.74**
AMT	Propositions	n/a			n/a			n/a			**0.9**	**0.67**	**0.77**
	AR				n/a	n/a	0.76	0.61	0.64	0.62	**0.91**	**0.66**	**0.77**
	ARC				n/a			n/a			**0.88**	**0.66**	**0.75**
US2016G1tv	Propositions	n/a			n/a			n/a			**0.66**	**0.62**	**0.64**
	Inference							0.51	0.62	0.56	**0.65**	**0.63**	**0.64**
	ARC							n/a			**0.63**	**0.61**	**0.62**

Table 3: The performance of Stab and Gurevyech's technique (2014b) (SG2014b), Peldszus and Stede's technique (2016) (PS2016), PLS and DAM in extracting the components of an argument, AR and the category of AR (ARC) (inference vs conflict) on AAEC (paragraph and essay level), AMT and US2016G1tv.

AR identification. The performance of our approach in identifying premises, conclusions, AR and the category of AR (inference vs conflict) is presented in Table 3. Since the AR between a premise and conclusion depends on the similarity between the ***C*** and ***A***, we tune the value of α to 0.4 on a development set (similar components have a similarity measure greater than 0.4).

Following the evaluation strategy of Stab and Gurevych (2014b), we first evaluate our approach on AAEC at paragraph and essay levels where we achieve F measures of 0.79 and 0.74, respectively. We have also achieved an F measure of 0.77 on the AMT corpus and 0.64 on US2016G1tv corpus. The performance of our approach tends to confirm our initial hypothesis: the AR between propositions is indeed governed by the relation between their functional components. The performance varies across the three corpora with the lowest performance observed on the US2016G1tv corpus. We have inspected the three corpora to identify the possible factors and identified three issues: (a) similarity is dependent on the information presented in the propositions alone, yet US2016G1tv is particularly demanding in that understanding many of the utterances depends upon (external) context in addition to what is present in the discourse; (b) since US2016G1tv corpus is dialogical, unlike the others, it includes the speakers' text in the construction of propositions and hence their representation is more complex than the monological corpora. The complex representations of propositions make the formalization and the extraction of target concepts and aspects difficult; (c) the AMT corpus has a high proportion of co-reference to represent ***C*** and ***A*** resulting in poor similarity, since the similarity between a word and its co-reference is low.

4.3 Error Analysis

Two major error types are observed. The first is related to propagation of the errors encountered during ***C*** and ***A*** extraction to the similarity identifier and AR identifier affecting the overall performance. Specifically, when a word is incorrectly identified as part of ***C*** or ***A***, their similarity measure is affected and then the decision about the AR.

The second error type is related to the similarity module which provides incorrect result for certain words. For instance, ADW provides comparable similarity values between *"food"* and *"meal"*,

and between *"food"* and *"family"*. Yet the first pair is more closely related as compared to the later. Moreover, propositions involving two or more categories of aspects (where each category is supported or attacked by different propositions) present a challenge, since it requires grouping of the aspects and consider each group as a unit to compute similarity.

4.4 Comparison Systems

We have compared our approach against the leading techniques in the field including Stab and Gurevych's work (2014b), Peldszus and Stede's (2016) work, and proposition level similarity. We re-implement proposition level similarity and use the results reported by the authors for the remaining approaches.

Stab and Gurevych (2014b) propose a classifier which identify argument components and AR category using a multiclass classification on (AAEC) (Stab and Gurevych, 2014a). Instead of considering the entirety of essay, they connect propositions within the same paragraph. They use Weka implementation of four different classifiers: SVM, Naive Bayes, C4.5 Decision Tree and Random Forest (Hall et al., 2009). SVM scored the best result with an overall accuracy of 0.73 and 0.72 in identifying argument components and AR respectively on AAEC (Stab and Gurevych, 2014a) at paragraph level.

Peldszus and Stede (2016) aim to map RST trees to argumentation structures (Taboada and Mann, 2006) using subgraph matching and an evidence graph model. They evaluate several features of their system on AMT (Peldszus and Stede, 2013). We are concerned with one of the features in order to make direct comparison: identifying if two EDUs are connected on which they achieve an overall F-measure of 0.76.

Most related to our work is an approach using proposition level similarity (PLS) as an integral component to determine argument structure (Lawrence and Reed, 2015). They use similarity to indicate the AR existing between EDUs and supplement other features to identify the entire argument structure. Since the similarity component alone can not induce the direction of the relation between the EDUs, we compared its performance in terms of detecting the existence of AR between EDUs. PLS provides a challenge to identify among different relations, since a pair of propositions in a given argument can score strong similarity without involving AR. PLS does not identify the direction of relation (claim vs premise) and hence these values are listed as n/a in Tables 3. We also use n/a to indicate that the evaluation result for the respective evaluation criteria (identifying premise, conclusion and AR) is not available for the comparison approaches.

Table 3 shows the performance of DAM, PLS, Stab and Gurevych's approach (2014b), and Peldszus and Stede's (2016) approach on the three data-sets. DAM outperformed all the approaches across the three corpora achieving the highest precision, recall and F-measure. The decrease in recall on AMT is attributed to the fact that co-references are productive in the corpus affecting similarity output, since similarity techniques are dependent on the lexicon choice (i.e the similarity between a word and its co-reference is low).

5 Conclusion

In this work, we have presented an approach for linking premises and conclusions that uses the similarity of target concepts and aspects, and the agreement between the opinions on target concepts and aspects of EDUs. We have demonstrated that the argument relations existing between propositions are largely dependent on the relations existing between the individual components (target concepts, aspects, opinions on target concepts and opinions on aspects) of the propositions. It would also be nice to explore about more fine-grained functional components and grammatical entities in the future works. Not only does our DAM approach outperform the current state of the art, most importantly, it is shown to work without modification across heterogeneous corpora (AAEC, AMT and US2016G1tv) which are substantially different in kind. This generality is an important milestone in the development of argument mining techniques and suggests that a combination of structural and distributional techniques, as employed here, offers the potential for robust, domain-independent performance in this extremely demanding task.

Acknowledgments

This research was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) in the United Kingdom under grant EP/N014871/1.

References

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 8–15. Association for Computational Linguistics.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, pages 2670–2676.

Daniel M. Bikei, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*, pages 152–160.

Carlos Chesnevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The knowledge engineering review*, 21(4):293–316.

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*, pages 11–22. Association for Computational Linguistics.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: an argument analysis interface. In *Computational Models of Argument: Proceedings of COMMA*, volume 266, page 463.

Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. Opinionminer: A novel machine learn-ing system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1195–1204. ACM.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics.

John Lawrence, Mathilde Janier, and Chris Reed. 2015. Working with open argument corpora. In *European Conference on Argumentation*, pages 367–380.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

John Lawrence and Chris Reed. 2017. Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In *Proceedings of the 4th International ACL/EMNLP Workshop on Argument Mining*, pages 108–117.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Joselice Ferreira Lima, Clia M. Gomes Amaral, and Lus Fernando R. Molinaro. 2010. Alternation. *CENTERIS*, 2(11):426–435.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, pages 188–191. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marie-Francine Moens, Eric 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. ACM.

Philippe Muller, Stergos D. Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. *Proceedings of COLING 2012*, pages 1883–1900.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 196–204.

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.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining*, pages 103–112.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate andwalk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation,SemEval '17,Association for Computational Linguistics*.

Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. 2006. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. pages 1501–1510.

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

Maite Taboada and William Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Mikolov Tomas, Chen Kai, Corrado Greg, and Dean Jeffrey. 2013. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv*, pages 1301–3781.

Maria P.G. Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. *COMMA*, 245:23–34.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*.

Lei Zhang and Bing Liu. 2014. Aspect and entity extraction for opinion mining. In *Data mining and knowledge discovery for big data*, pages 1–40. Springer.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 419–426. Association for Computational Linguistics.