Argument Extraction from News

Christos Sardianos
Dept. of Informatics and Telematics
Harokopio University of Athens
Omirou 9, Tavros, Athens, Greece
sardianos@hua.gr

Ioannis Manousos Katakis and **Georgios Petasis** and **Vangelis Karkaletsis**
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) “Demokritos”
GR-153 10, P.O. BOX 60228, Aghia Paraskevi, Athens, Greece
{gkatakis, petasis, vangelis}@iit.demokritos.gr

Abstract

Argument extraction is the task of identifying arguments, along with their components in text. Arguments can be usually decomposed into a claim and one or more premises justifying it. The proposed approach tries to identify segments that represent argument elements (claims and premises) on social Web texts (mainly news and blogs) in the Greek language, for a small set of thematic domains, including articles on politics, economics, culture, various social issues, and sports. The proposed approach exploits distributed representations of words, extracted from a large non-annotated corpus. Among the novel aspects of this work is the thematic domain itself which relates to social Web, in contrast to traditional research in the area, which concentrates mainly on law documents and scientific publications. The huge increase of social web communities, along with their user tendency to debate, makes the identification of arguments in these texts a necessity. In addition, a new manually annotated corpus has been constructed that can be used freely for research purposes. Evaluation results are quite promising, suggesting that distributed representations can contribute positively to the task of argument extraction.

1 Introduction

Argumentation is a branch of philosophy that studies the act or process of forming reasons and of drawing conclusions in the context of a discussion, dialogue, or conversation. Being an important element of human communication, its use is very frequent in texts, as a means to convey meaning to the reader. As a result, argumentation has attracted significant research focus from many disciplines, ranging from philosophy to artificial intelligence. Central to argumentation is the notion of argument, which according to [Besnard and Hunter, 2008] is a set of assumptions (i.e. information from which conclusions can be drawn), together with a conclusion that can be obtained by one or more reasoning steps (i.e. steps of deduction). The conclusion of the argument is often called the claim, or equivalently the consequent or the conclusion of the argument, while the assumptions are called the support, or equivalently the premises of the argument, which provide the reason (or equivalently the justification) for the claim of the argument. The process of extracting conclusions/claims along with their supporting premises, both of which compose an argument, is known as argument extraction [Goudas et al., 2014] and constitutes an emerging research field.

Nowadays, people have the ability to express their opinion with many different ways, using services of the social Web, such as comments on news, fora, blogs, micro-blogs and social networks. Social Web is a domain that contains a massive volume of information on every possible subject, from religion to health and products, and it is a prosperous place for exchanging opinions. Its nature is based on debating, so there already is plenty of useful information that waits to be identified and extracted [Kiomourtzis et al., 2014].

Consequently, there is a large amount of data that can be further explored. A common form for mining useful information from these texts, is by applying sentiment analysis techniques. Sentiment analysis can be proven as a quick way to capture sentiment polarity of people about a specific topic. Two of the domains

Proceedings of the 2nd Workshop on Argumentation Mining, pages 56–66,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

where capturing public opinion is of great importance, are e-Government and policy making. In this way, politicians and policy makers can refine their plans, laws and public consultations prior to their publication or implementation. Additionally, it could help the voters in deciding which policies and political parties suit them better. However, a more fine-grained analysis is required in order to detect in which specific aspects of a policy, a citizen is in favour or against. Such analysis can be achieved through argument extraction: once a document that relates to a policy is located, it is examined in order to identify segments that contain argument elements, such as premises that are against or in support of a claim or an entity (such as nuclear energy or renewable energy sources). The main idea behind this filtering of public opinion as found on the social Web, is that citizens that try to justify their opinion with arguments may be more important or influential than the less justified ones.

Motivated by this need, in this paper we propose a supervised approach for argument extraction from relevant media, based on Conditional Random Fields [Lafferty et al., 2001]. Following the state of the art (i.e. [Goudas et al., 2014; Hou et al., 2013]), our approach studies the applicability of existing approaches on the domain of social Web, mainly news and blogs, although the evaluation focuses only on news, due to copyright issues[1]. Assuming that we know whether a sentence contains an argument element or not (i.e. by applying an approach similar to the one described in [Goudas et al., 2014]), our approach tries to detect the exact segments that represent these elements (i.e. claims and premises) through the use of a CRF classifier [Lafferty et al., 2001]. Targeting a set of thematic domains and languages as wide as possible, we have tried to minimise the use of domain and language depended resources. Thus our approach exploits features such as words, part-of-speech tags, small lists of language-dependent cue words, and distributed representations of words [Mikolov et al., 2013a,b,c], that can be easily extracted from unannotated large corpora. Our approach has been evaluated on manually annotated news in the Greek language, containing news from various thematic domains, including sports, politics, economics, culture, and various social problems, while the evaluation results are quite promising, suggesting that distributed representations can contribute positively to this task.

The rest of the paper is organized as follows: Section 2 refers to the related work on argument extraction, section 3 describes the proposed methodology and the corresponding features used for our approach. Section 4 presents the experimental results and the tools we utilized and finally, section 5 concludes the paper and proposes some future directions.

2 Related Work

A plethora of argument extraction methods consider the identification of sentences containing arguments or not as a key step of the whole process. More specifically, the above approaches face the process of argument extraction as a two-class classification problem. However, there are approaches which try to solve the argument extraction problem in a completely different way. [Lawrence et al., 2014] combined a machine learning algorithm to extract propositions from philosophical text, with a topic model to determine argument structure, without considering whether a piece of text is part of an argument. Hence, the machine learning algorithm was used in order to define the boundaries and afterwards classify each word as the beginning or end of a proposition. Once the identification of the beginning and the ending of the argument propositions has finished, the text is marked from each starting point till the next ending word. An interesting approach was proposed by [Graves et al., 2014], who explored potential sources of claims in scientific articles based on their title. They suggested that if titles contain a tensed verb, then it is most likely (actually almost certain) to announce the argument claim. In contrast, when titles do not contain tensed verbs, they have varied announcements. According to their analysis, they have identified three basic types in which articles can be classified according to genre, purpose and structure. If the title has verbs then the claim is repeated in the abstract, introduction and discussion, whereas if the title does not have verbs, then the claim does not appear in the title or introduction but appears in the abstract and discussion sections.

Another field of argument extraction that has recently attracted the attention of the research community, is the field of argument extraction from online

[1] Although we have created a manually annotated corpus concerning both news and blogs, only the corpus containing news can be redistributed for research purposes.

discourses. As in the most cases of argument extraction, the factor that makes the specific task such challenging, is the lack of annotated corpora. In that direction, [Houngbo and Mercer, 2014], [Aharoni et al., 2014] and [Green, 2014] focused on providing corpora, that could be widely used for the evaluation of the argument extraction techniques. In this context, [Boltužić and Šnajder, 2014] collected comments from online discussions about two specific topics and created a manually annotated corpus for argument extraction. Afterwards they used a supervised model to match user-created comments to a set of predefined topic-based arguments, which can be either attacked or supported in the comment. In order to achieve this, they used textual entailment (TE) features, semantic text similarity (STS) features, and one "stance alignment" (SA) feature. One step further, [Trevisan et al., 2014] described an approach for the analysis of German public discourses, exploring semi-automated argument identification by combining discourse analysis methods with Natural Language Processing methods. They focused on identifying conclusive connectors, substantially adverbs (i.e. hence, thus, therefore), using a multi-level annotation on linguistic means. Their methodological approach consists of three steps, which are performed iteratively (manual discourse linguistic argumentation analysis, semi-automatic Text Mining (PoS-tagging and linguistic multi-level annotation) and data merge) and their results show the argument-conclusion relationship is most often indicated by the conjunction because followed by since, therefore and so. [Ghosh et al., 2014] attempted to identify the argumentative segments of texts in online threads. They trained expert annotators to recognize argumentative features in full-length threads. The annotation task consisted of three subtasks. In the first subtask, annotators had to identify the Argumentative Discourse Units (ADUs) along with their starting and ending points. Secondly, they had to classify the ADUs according to the Pragmatic Argumentation Theory (PAT) into Callouts and Targets. As a final step, they indicated the link between the Callouts and Targets. Apart from that, they proposed a hierarchical clustering technique that assess how difficult it is to identify individual text segments as Callouts. [Levy et al., 2014] defined the task of automatic claim detection in a given context and outlined a preliminary solution. Their supervised learning approach relied on a cascade of classifiers designed to handle the skewed data. Defining their task, they made the assumption that the articles given are relatively small set of relevant free-text articles, provided either manually or by automatic retrieval methods. More specifically, the first step of their task was to identify sentences containing context dependent claims (CDCs) in each article. Afterwards they used a classifier in order to find the exact boundaries of the CDCs detected. As a final step, the ranked each CDC in order to isolate the most relevant to the corresponding topic CDCs. That said, their goal is to automatically pinpoint CDCs within topic-related documents.

3 Proposed Approach

The work presented in this paper is motivated mainly by needs in the area of e-Government and policy making, aiming at performing argument extraction on large corpora collected from the social Web, targeting mainly on-line newspapers and blogs. Through a process that identifies segments that correspond to argument elements (claims and premises), performs aspect-based sentiment analyses, matches arguments to policy elements, and aggregates results from multiple sources, policy makers have the ability to receive the necessary feedback for ongoing public consultations, laws, issues that concern citizens, and captures the public opinion towards various issues. In this context, identified opinions are classified according to the contained argumentation that supports each opinion: Apparently, argument extraction can be a powerful tool for any decision making procedure. For example, it would be extremely useful for a government to be in position of knowing the public opinion about a law that is intended to be presented. Apart from that, it is of great value to detect the arguments against or in favour used in public discussions about the specific issue, in order to end up with a law which would be acceptable from a larger percentage of citizens.

The requirements for an argument extraction approach operating in such a context are several, including the ability to process as many thematic domains as possible, to be as accurate as possible regarding the identified argument elements, and utilise as fewer linguistic resources as possible, as it needs to operate also in less-resourced languages, such as Greek. Of course, it should be able to extract arguments from documents that influence the public opinion (such as news) or documents where citizens express their opinions and views (such as blogs). The goal of this research is to develop an approach for the task of argument extraction, based on machine learning, that will fulfill these requirements and will be applicable to the Greek language.

Our approach is based on Conditional random fields (CRFs) [Lafferty et al., 2001], a probabilistic framework for labeling and segmenting structured data such as sequences, which has been applied to a wide range of segmenting tasks, from named-entity recognition [McCallum and Li, 2003] and shallow parsing [Sha and Pereira, 2003], to aspect-based sentiment analysis [Patra et al., 2014]. Beyond features such as words and part-of-speech tags, our approach exploits a small lexicon of cue words, which usually signal the presence of a premise segment, and distributed representations of words [Mikolov et al., 2013a,b,c]. These map words to vectors of a high-dimensional space (usually more than 100 dimensions), which are created without human intervention from observing the usage of words on large (non-annotated) corpora. More specifically, our approach exploits the "word2vec"[2] tool [Mikolov et al., 2013a,b,c], which can make highly accurate guesses about a word's meaning based on its usage, provided enough data, usage and context for each word are available. The "word2vec" approach tries to arrange words with similar meaning close to each other, and interesting feature that we want to exploit in our approach in order to widen the "word space" beyond the words observed during the training phase.

3.1 Expansion of word feature space

Trying to provide an approach for argument extraction supporting multiple thematic domains, we exploit word similarities for expanding the word feature space. As already discussed, "word2vec" is a tool that computes similarities between words from large corpora and generates a real-valued feature vector for each word. It actually trains a recurrent neural network and maximizes the probability for a word to appear in a specific context.

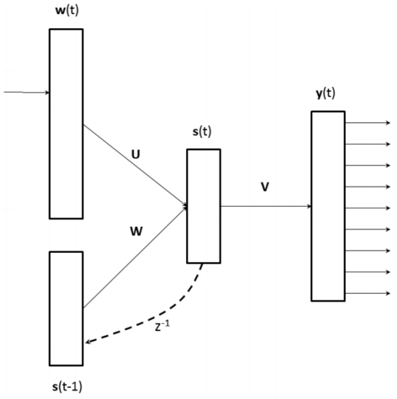

Figure 1: Recurrent Neural Network Language Model (Mikolov et al., 2013)

As shown in figure 1, each word comes as input to the first layer $w(t)$ of the recurrent neural network, representing an input word at time t. As a result, matrix u holds the word representations, with each column representing the words. The hidden layer $s(t)$ maintains a representation of the sentence history by having a recursive connection z^{-1} to the previous word $s(t-1)$. Finally, $y(t)$ produces a probability distribution over words, from which a list of similar words is generated as output. In practice, "word2vec" takes as input a continuous stream of words from a corpus and generates a ranking including the k (defined by the user) most similar words for each word appeared in the input stream. As an example, the most similar words for the word "ορειβασία" ("climbing") according to our "word2vec" generated model for Greek are shown in Table 1, while Table 2 shows the 40 most similar words to the Greek word "λιγνίτης" ("lignite"), selected from the domain of renewable energy sources. As can be seen from Table 2, all suggested words according to cosine similarity over the word feature vectors are relevant to the thematic domain where lignite belongs, with 4 most similar words being either inflected forms of lignite in Greek, or other forms of carbon-related substances.

Five Most Similar Words	Cosine Similarity
ιππασία (horse-riding)	0.748
ποδηλασία (cycling)	0.721
πεζοπορία (hiking)	0.683
ιστιοπλοία (sailing)	0.681
καγιάκ (kayak)	0.674

Table 1: *"Word2vec"* sample output (most similar words to the Greek word "ορειβασία" ("climbing")).

[2] https://code.google.com/p/word2vec/

Similar Words	Cosine Similarity	Similar Words	Cosine Similarity
λιγνίτη (lignite)	0.694903	ρυπογόνο (polluting)	0.493400
λιθάνθρακας (coal)	0.665466	βιοαιθανόλη (bioethanol)	0.489851
άνθρακας (carbon)	0.644011	βιοαέριο (biogas)	0.481461
λιθάνθρακα (coal)	0.631198	ανανεώσιμα (renewable)	0.481460
ηλεκτροπαραγωγή (electricity production)	0.621633	μαζούτ (fuel)	0.478094
λιγνίτες (lignite)	0.580237	υδροηλεκτρικά (hydropower)	0.473288
ηλεκτρισμός (electricity)	0.555800	ζεόλιθος (zeolite)	0.473254
καύσιμο (fuel)	0.541152	βιομάζας (biomass)	0.473129
ορυκτά (fossil)	0.536743	ορυκτός (fossil)	0.472967
ηλεκτροπαραγωγής (electricity production)	0.532764	παραγόμενη (produced)	0.467192
βιομάζα (biomass)	0.532644	λιγνιτική (lignitic)	0.467016
γαιάνθρακες (coal)	0.509080	γεωθερμία (geothermal)	0.464868
ανανεώσιμη (reniewable)	0.508831	λιγνιτικών (lignitic)	0.464730
υδρογόνο (hydrogen)	0.503391	μεταλλεύματα (ores)	0.456796
αντλησιοταμίευση (pumped storage)	0.500784	ορυκτό (mineral)	0.456025
υ/η (hydropower)	0.499954	υδροηλεκτρική (hydropower)	0.454693
κάρβουνο (charcoal)	0.498860	ρυπογόνος (polluting)	0.451683
αιολική (wind)	0.498321	εξορύσσεται (mined)	0.450633
πλούτος (wealth)	0.496383	λιγνιτικές (lignitic)	0.449569
χάλυβας (steel)	0.494852	καυστήρας (burner, boiler)	0.447930

Table 2: *"Word2vec"* sample output (40 most similar words to the Greek word "λιγνίτης" ("lignite")). Model extracted from documents originating from news and Blogs.

Cosine similarity can also be computed at phrase-level, which means that the model tries to match words or phrases to a specific phrase. However, the size of the phrase vector file is more than twice size of the word vector file produced from the same corpus. Thus, using a phrase model requires a lot more computational resources than a word model.

3.2 Semi-supervised approach for extracting argument components

Concerning our approach for extracting argument components, we decided to extend the approach proposed by [Goudas et al., 2014], which also addressed a less-resourced language, such as Greek. [Goudas et al., 2014] suggested a two-step technique in order to extract arguments from news, blogs and social web. In the first phase of their method, they attempted to identify the argumentative sentences, employing classifiers such as Logistic Regression [Colosimo, 2006], Random Forest [Leo, 2001], Support Vector Machines [Cortes and Vapnik, 1995], Naive Bayes [Nir Friedman and Goldszmidt, 1997], etc. The features used in the classification were divided into features selected from the state of the art approaches and new features that were chosen for the domain of their application. Specifically, the state of the art features chosen, supply information about the position of the sentence inside the document as well as the number of commas and connectives inside the sentence. Moreover they examined the number of verbs (active and passive voice) in the sentence, the existence and number of cue words and entities, the number of words and adverbs in the context of a sentence, and finally the average length in characters of the words in the sentence. Regarding the new features added, this includes the number of adjectives in the sentence, the number of entities in the n^{th} previous sentence and the total number of entities from the previous n sentences. In addition to the previous features, they also examined the ratio of distributions (language models) over unigrams, bigrams, trigrams of words and POS tags.

After the extraction of the argumentative sentences, they proceeded to the process of argument components (claims and premises) identification. In this stage, they applied a CRF classifier on a manually corpus. The features required for this task were the words of the sentences, gazetteer lists of known entities for the thematic domain, gazetteer lists of cue words and lexica of verbs and adjectives that appear most frequently in argumentative sentences of the training data.

In the approach proposed by [Goudas et al., 2014], gazetteers are core features of the argument extraction process. In our approach, we want to reduce this dependency on gazetteers, by exploiting distributed representation for words, using the proposed method described in subsection 3.1. This will help us widen the spectrum of words that can be handled by our classifier and thus, manage to create a more fine-grained CRF model.

4 Empirical Evaluation

In this section the performance of the proposed approach will be examined. The performance metrics that will be used in order to evaluate our approach is accuracy, precision, recall and F1-measure. The empirical evaluation involves two experiments: The first experiment concerns that use of the "word2vec" tool, in order to obtain a suitable model for Greek, while the second experiment involves the evaluation of our approach for argument extraction on a manually annotated corpus.

4.1 Obtaining a "word2vec" model for Greek

In this section the steps performed for acquiring a "word2vec" model for the Greek language will be described, while the performance of the acquired model regarding word similarities will be examined. The performance metric that will be used in order to evaluate our word similarity model is accuracy. Accuracy denotes the number of words that are strictly related to the word given divided by the total number of words suggested as similar.

4.1.1 Experimental Setup

Dealing with semantic similarities, requires large volumes of data. As a result, in order to extract the distributed representation of words with the "word2vec" tool, we used a corpus that included around 77 million documents. These documents were written in Greek, and originated from news, blogs, Facebook[3] and Twitter[4] postings. Table 3 presents some properties of the utilised corpus. All documents were converted to lower-case before processed with the "word2vec" tool.

The evaluation task for this experiment related to the ability to extend a gazetteer (lexicon) of cue words or domain-specific entities with new entries, by exploiting the "word2vec" generated models to detect similar words. In order to evaluate this task, a seed list of cue words/entities was manually constructed. For each word in the seed list, the five more similar words were identified with the obtained "word2vec" model, and used to augment the list. Then, these new entries to the lists were manually examined, in order to identify which of these additions were correct or not (i.e. new entries were also cue words or entities from the same thematic domain).

	News	Blogs	Facebook	Twitter
Sentences	23.4	42.9	17.6	166
Words	492.8	853.2	197.3	1400

Table 3: Corpus Properties (in millions of documents).

4.1.2 Evaluation Results

Since the documents in our corpus were divided in four large categories (according to their source of origin), we started with the creation of four different "word2vec" models. Evaluation of the acquired models showed that news and blogs provide more fine-grained models in comparison to the models obtained from Facebook and Twitter. This happens because the Facebook and Twitter postings are usually less formal, many words are used with different senses than in news/blogs, postings may not have proper syntax or spelling and often contain abbreviations. As a result, a lot of noise has been inserted in the corresponding output models.

The authors of [Goudas et al., 2014] have made available to us the cue word and entity lists they have used in their experiments, which concern the thematic domain of renewable energy sources. Their list of cue words was manually extracted from their corpus by the researches, while the list of entities was provided by domain experts and policy makers.

Trying to expand these lists, we randomly selected twenty cue words and twenty entities from these, as a seed. For each seed word, the five more similar words were examined. Evaluation results suggest that there was a large variation on the similarities drawn for the same words from the news/blogs corpora and the Facebook/Twitter corpora. As it was expected, the models produced from the Facebook and Twitter corpora were worse than the others.

[3] http://www.facebook.com/

[4] http://www.twitter.com/. Each "tweet" was considered as a document.

Table 4 shows sample results for the word "λιγνίτης" ("lignite"), from the "word2vec" models of the news and blogs corpora. As we can see, the obtained similar words both for news and blogs corpora belong to the same domain, thus they can all be used to expand our word feature space and gazetteers for this specific domain.

News Corpus	Blogs Corpus
υγροποιημένο (liquefied)	λιγνίτη (lignite)
γαιάνθρακας (coal)	ηλεκτρισμός (electricity)
αέριο (gas)	ηλεκτρισμός (electricity)
σχιστολιθικό (shale)	ηλεκτροπαραγωγή (electricity production)
λιγνίτη (lignite)	λιθάνθρακα (bituminous coal)
ηλεκτρισμός (electricity)	βιοαέριο (biogas)
Σχιστολιθικό (Shale)	υδροηλεκτρικά (hydropower)
σχιστών (slit)	λιθάνθρακας (bituminous coal)
ηλεκτροπαραγωγής (electricity production's)	υδροηλεκτρισμό (hydroelectricity)
ηλεκτροπαραγωγή (electricity production)	βιομάζα (biomass)

Table 4: Similar words according to the News/Blogs "word2vec" model.

On the other hand, as shown in Table 5, the results from Facebook and Twitter for the same word ("λιγνίτης") are completely irrelevant. After examining the results, we observed that the sense of many words varies between news/blogs and facebook/twitter corpora. For example, the word "attractive", in Twitter and Facebook is used in most cases as "handsome" (i.e. attractive person), while in news and blogs is usually referred as "interesting" (i.e. attractive investment). One reason for this, is clearly the irrelevance of the topics discussed in social media and the use of language used in these discussion. In addition, the vocabulary normally used in social media is not as specialized as in news sites. This means that the similarity results from social media are not expected to be efficient for using in domain independent models. A noted fact that supports the above findings is the frequency of appearance of the word "λιγνίτης" ("lignite") in the corpora. Specifically, the word "λιγνίτης", appeared 5087 times in the news/blogs corpora, unlike the Facebook/Twitter corpora that appeared 1615 times.

Facebook Corpus	Twitter Corpus
φόρτος (load)	αντιευρωπαϊσμός (anti-Europeanism)
δανειστής (loaner)	αριθμητής (numerator)
κιτρινισμός (yellowing)	εθνικισμός (nationalism)
εκτιμώμενος (estimated)	ιχνηλάτης (tracker)
αποκαθήλωση (pieta)	τ'αγοράζει (buys)
εισέπρατε (received)	εφοπλισμός (fitting)
τερματοφύλακας (goalkeeper)	Μπερλουσκονισμός (Berlusconism)
ψυχισμός (psyche)	περιπατητικός (ambulatory)
πεισμωμένος (stubborn)	κορπορατισμός (corporatism)
δανειολήπτης (borrower)	μονοπωλιακός (monopolistic)

Table 5: Similar words according to the Facebook/Twitter "word2vec" model.

Even the union of Facebook/Twitter corpora did not improve the performance of the generated model. On the other hand, the merge of the blogs and news corpora showed a significant increase on the performance of the "word2vec" model produced. The final evaluation of the "word2vec" models was conducted by two human annotators. Annotators were supplemented with a set of 20 randomly selected words which did not belong to a specific domain. The analogy between entities and cue words remained the same. Along with each word, a list with the five most similar words, as produced from the "word2vec" model, was provided. The evaluation results are shown in Table 6. According to these results, we can conclude to the fact that "word2vec" can be used for the expansion of the cue word lexicons. In addition, it can be proven a valuable resource as regards to the enrichment of the entities provided by the policy makers.

4.2 CRFs for argument extraction

In this section, the proposed approach based on CRFs and distributed representations of words will be evaluated, with the help of a manually annotated corpus, containing annotated segments that correspond to argument elements (claims and premises).

4.2.1 Experimental Setup

Unfortunately, the corpus used in [Goudas et al., 2014] was not available due to licensing limitations. As a result, we had to create a new manually annotated corpus in order to evaluate our approach. We collected 300 news articles written in Greek from

	Entities			Cue Words		
	Annot. A	**Annot. B**	**A+B**	**Annot. A**	**Annot. B**	**A+B**
Five most similar	0.810	0.840	0.825	0.830	0.870	0.850

Table 6: Evaluation Results: Accuracy of 5 most similar words.

the Greek newspaper "Αυγή"[5]. According to their site, articles can be used without restriction for non-commercial purposes. The thematic domain of the articles varies from politics and economics to culture, various social issues and sports. The documents were manually annotated by two post-graduate students with moderate experience on the annotation process. Prior to the beginning of the annotation task, the annotators were supplied with guidelines describing the identification of arguments, while a QA session was carried out afterwards. The guidelines contained text examples of premises *in favor* or *against* the central claim stated by the articles' author. In these terms, the annotators were initially called to identify the central claims stated from the author of each article. Subsequently, they looked for text segments attacking or supporting every claim respectively. These segments may sometimes start with cue words such as "διότι" ("because"), "για να" ("in order to"), "αλλά" ("but"), or may just follow the usual sentence structure. Each annotator annotated 150 documents with argument components (premises and claims).

Once each annotator has annotated half of the corpus, pre-annotation has been applied, as a proven way to obtain significant gains in both annotation time and quality of annotation [Fort and Sagot, 2010; Marcus et al., 1993; Rehbein et al., 2009]. Since we were targeting errors of omission (segments missed by the annotators), an "overly-general" CRF model was trained on all 300 documents, and applied on the corpus. The CRF model is characterised as "overly-general", as it was derived only from sentences that contained claims and premises. Sentences not containing argument elements were omitted from training. The CRF model detected 4524 segments, significantly more than the 1172 segments annotated by the two annotators. A second round of annotation was performed, where both layers of annotations were visible (both the manual and the segments obtained through machine learning), and each annotator was asked to revise his own annotations, having two goals: *a*) examine whether any of the segments detected by the CRF model is either a claim or a premise, and *b*) exploit their experience from annotating 150 documents, to revise their annotations, especially the ones done during the early stages of annotation. During this second annotation step, a small number of errors was corrected and 19 new segments were added as argument elements, producing the "final" version of the manually annotated corpus[6], which has been used for evaluating our approach. The final version of the corpus contains 1191 segments annotated as argument elements.

Although the production of the corpus is still an ongoing process, we measured the inter-annotation agreement between of the two annotators over a fraction of the entire corpus. For this reason, we asked each annotator to annotate eighty articles already annotated by the other annotator, leading to 170 documents (out of 300) annotated by both annotators. Annotator A has annotated 918 argument elements, while annotator B has annotated 735 argument elements, out of which 624 were common between the two annotators, leading to a precision of 84.90%, a recall of 67.97%, with an F1 measure of 75.50%.

The manually annotated corpus containing 300 documents was used in order to evaluate our approach. For all evaluations, 10-fold cross validation was used, along with precision, recall, and F1 measure as the evaluation metrics. In order to measure the increase in performance, we have used a base case. Our base case was a CRF model, using as features the words and pos tags.

Our approach for argument extraction seeks to detect the boundaries of a text fragment that encloses a claim or a premise of an argument. One way to achieve this task, is to classify each word (token) of a sentence as a "boundary" token, i.e. as a token that "starts" or "ends" an argumentative segment. Using such a representation, the task can be converted into a

[5] `http://www.avgi.gr`

[6] The corpus that has been used in this evaluation is publicly available for research purposes from the authors. A revised (second) version of the corpus may be also available in the future.

classification task on each token. The "BILOU" representation seeks to classify each token with a single tag, which can be any tag from the following set: *a*) **B**: This tag represents the start/begin of a segment. It must be applied on the first token of a segment. *b*) **I**: This tag marks a token as being inside a segment. It must be applied on any token inside a segment, except the first and last ones. *c*) **L**: This tag represents the end of a segment. It must be applied on the last token of a segment. *d*) **O**: This tag marks a token as being outside a segment. It must be applied on any token that is not contained inside a segment. *e*) **U**: This tag correspond to "unit" segments, which are segments that contain a single token. It is a special case that marks a token that is the beginning and end of a segment simultaneously. For example the BILOU representation of the sentence "Wind turbines generate noise in the summer" is presented in Table 7.

BILOU tag	**word**	**prev. word**	**next word**	**...**
B-premise	Wind	-	turbines	...
I-premise	turbines	Wind	generate	...
I-premise	generate	turbines	noise	...
L-premise	noise	generate	in	...
O	in	noise	the	...
O	the	in	summer	...
O	summer	the	-	...

Table 7: Example of the BILOU representation of a sentence.

4.2.2 Results

The base case evaluation is shown in table 8. The features utilized in the base-case evaluation are: a) the words in these sentences, b) the part of speech of the words. We have performed evaluation with various words as context (0, ±2, and ±5 words before and after the word in concern). As seen from the results, the experiment which the context-5 was applied shows a slight improvement from the context-2 experiment, while the difference is larger in the case of zero context.

Context	Precision	Recall	F1
0	16.80% ±5.52	7.55% ±2.80	10.39% ±3.69
±2	34.00% ±3.19	22.33% ±2.73	26.93% ±2.85
±5	33.08% ±3.45	22.92% ±3.99	27.04% ±3.89

Table 8: CRF base case evaluation: words + pos tags.

After the evaluation of the base case, we examined the impact of our gazetteer on the results. As seen in the table 9, the addition of the gazetteer provides a slight boost in out results. The most important difference in relationship with the performance of the base case is shown when no context words were used. Unlike to the previous experimental setup, when two words were used as context has better performance results instead of using five.

Context	Precision	Recall	F1
0	20.22% ±4.43	11.95% ±3.32	14.90% ±3.65
±2	35.61% ±3.75	24.36% ±3.34	28.85% ±3.19
±5	34.06% ±3.85	24.96% ±4.18	28.76% ±4.06

Table 9: CRF base case evaluation: words + pos tags + context 2/5.

Afterwards, we examined the case in which word embeddings were used for the expansion of our gazetteer. In this case, we measured in what manner the extended gazetteer created using "word2vec" could affect the performance of our model. Table 10 shows the evaluation results according to the different number of words used as context. The overall performance of our model was improved when two or five words were used as context, whereas the performance of our model decreased in the zero context configuration. As seen below the best result performed by the configuration of two word context.

Context	Precision	Recall	F1
0	20.74% ±2.63	11.29% ±1.88	14.60% ±2.20
±2	39.70% ±4.55	27.59% ±3.54	32.53% ±3.90
±5	38.72% ±5.29	27.60% ±3.36	32.21% ±4.06

Table 10: CRF base case evaluation: words + pos tags + context 2/5.

5 Conclusion

In this research paper we propose an approach for argument extraction that exploits distributed representations of words in order to be applicable on multiple thematic domains, without requiring any other linguistic resource beyond a part-of-speech tagger and a small list of cue words. Our goal was to suggest a semi-supervised method, applicable from traditional news and blogs documents to corpora from social web, mainly written in the Greek language. The proposed approach is based on previous research performed on this domain and attempts to extend its existing functionality. As gazetteer lists of entities and cue words play an important role to the argument extraction process, we suggest the expansion of the above gazetteer list which are usually provided by domain experts (in our case policy makers), using semantic similarities.

Regarding the future work of this research, we are going to examine the impact of applying bootstrapping techniques on the development of CRF models for the identification of argument components. In addition, it would be interesting to explore different classification algorithms for the extraction of premises and claims on argumentative sentences. Moreover, we would like to extract patterns based on verbs and POS and to examine if these patterns can be generalized through a grammatical inference algorithm.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 288513, and from Greek national funding (CLARIN-EL). For more details, please see the NOMAD project's website, `http://www.nomad-project.eu`, and CLARIN-EL project's website, `http://www.clarin.gr/`.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Philippe Besnard and Anthony Hunter. *Elements of argumentation*, volume 47. MIT press Cambridge, 2008.

Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

M. Strano; B.M. Colosimo. Logistic regression analysis for experimental determination of forming limit diagrams. *International Journal of Machine Tools and Manufacture*, 46(6):673–682, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125.

Karën Fort and Benoît Sagot. Influence of pre-annotation on pos-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 56–63, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-72-5.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 287–299. Springer, 2014.

Heather Graves, Roger Graves, Robert Mercer, and Mahzereen Akter. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Nancy Green. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Libin Hou, Peifeng Li, Qiaoming Zhu, and Yuan Cao. Event argument extraction based on crf. In Donghong Ji and Guozheng Xiao, editors, *Chinese Lexical Semantics*, volume 7717 of *Lecture Notes in Computer Science*, pages 32–39. Springer Berlin Heidelberg, 2013.

Hospice Houngbo and Robert Mercer. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of* *the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

George Kiomourtzis, George Giannakopoulos, Georgios Petasis, Pythagoras Karampiperis, and Vangelis Karkaletsis. Nomad: Linguistic resources and tools aimed at policy formulation and validation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

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

Breiman Leo. Random forests. *Machine Learning*, 45 (1):5–32, 2001.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics, 2014.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017.

Andrew McCallum and Wei Li. 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 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013c.

Dan Geiger Nir Friedman and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

Braja Gopal Patra, Soumik Mandal, Dipankar Das, and Sivaji Bandyopadhyay. Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 370–374, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 19–26, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-52-7.

Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Bianka Trevisan, Eva Dickmeis, Eva-Maria Jakobs, and Thomas Niehr. Indicators of argument-conclusion relationships. an approach for argumentation mining in german discourses. In *Proceedings of the First Workshop on Argumentation Mining*, pages 104–105, Baltimore, Maryland, June 2014. Association for Computational Linguistics.