Argument & Computation 9 (2018) 1–14
DOI 10.3233/AAC-170025
IOS Press

Argumentation mining: How can a machine acquire common sense and world knowledge?

Marie-Francine Moens
Department of Computer Science, KU Leuven, Belgium
E-mail: sien.moens@cs.kuleuven.be

Abstract. Argumentation mining is an advanced form of human language understanding by the machine. This is a challenging task for a machine. When sufficient explicit discourse markers are present in the language utterances, the argumentation can be interpreted by the machine with an acceptable degree of accuracy. However, in many real settings, the mining task is difficult due to the lack or ambiguity of the discourse markers, and the fact that a substantial amount of knowledge needed for the correct recognition of the argumentation, its composing elements and their relationships is not explicitly present in the text, but makes up the background knowledge that humans possess when interpreting language. In this article[1] we focus on how the machine can automatically acquire the needed common sense and world knowledge. As very few research has been done in this respect, many of the ideas proposed in this article are tentative, but start being researched.

We give an overview of the latest methods for human language understanding that map language to a formal knowledge representation that facilitates other tasks (for instance, a representation that is used to visualize the argumentation or that is easily shared in a decision or argumentation support system). Most current systems are trained on texts that are manually annotated. Then we go deeper into the new field of representation learning that nowadays is very much studied in computational linguistics. This field investigates methods for representing language as statistical concepts or as vectors, allowing straightforward methods of compositionality. The methods often use deep learning and its underlying neural network technologies to learn concepts from large text collections in an unsupervised way (i.e., without the need for manual annotations). We show how these methods can help the argumentation mining process, but also demonstrate that these methods need further research to automatically acquire the necessary background knowledge and more specifically common sense and world knowledge. We propose a number of ways to improve the learning of common sense and world knowledge by exploiting textual and visual data, and touch upon how we can integrate the learned knowledge in the argumentation mining process.

Keywords: Natural language understanding, representation learning, argumentative text processing

1. Introduction

Argumentation is an intelligent communication task that is inherent to human behaviour. It is a discourse activity that aims at increasing or decreasing the acceptability of a controversial claim or point of view [67]. Humans are very keen to convince others of their opinion and bring arguments to a discussion to support their claims. Finding arguments in an automated way in human discourse was early on discovered as a desirable characteristic of intelligent machines or agents [48], often referred to as argumentation mining or argument mining [10,46]. Throughout this paper we use the term "argumentation" as discourse phrases or sentences or larger discourse units that have an argumentative function

[1]The article is the written version of a keynote lecture given at COMMA 2016, the 6th International Conference on Computational Models of Argument, on September 13, 2016 at Potsdam University, Germany.

in this discourse. The term "argument" refers to the set of one or more phrases or sentences that function as "premises", along with another set of phrases or sentences that function as "conclusions" or "claims" and that together form the argument. The mining is usually applied on language data, although argumentation mining of non-verbal data such as video discourse is not excluded.

Argumentation mining has been defined as the automated detection of the argumentation structure and classification of both its component elements and their argumentative relationships. It is distinct from argumentative zoning, which only refers to the functional classification of discourse fragments (e.g., of scientific papers) [64], but which lacks the recognition of argumentative relationships. Over the years many argumentation models that expect to find a typical argumentation structure have been defined, the most popular being the model of Toulmin [65]. He has suggested an argumentation model composed of six component types: claim, data, qualifier, warrant, rebuttal and backing, which recently has been simplified to aid automated recognition into a claim and its premises [46] or into a claim and its premises where a distinction is made between supporting and attacking elements [29]. Recognizing argumentation structures in text is considered a difficult task both for humans and machines. Arguments are sometimes revealed by the presence of discourse markers (e.g., in English adverbs such as "because"), but such markers have often ambiguous meanings or are missing. Moreover, the premises in favor of the same claim might be far apart in a discourse, making it difficult to automatically link them together. Also argumentation structures might take the form of graphs, where nested tree structures (e.g., a full argumentation composed of a claim and its supporting premises or counter elements form the premise for a more general claim [46]). In the above cases humans recognize such argumentative discourse by relying on their world, common sense or domain specific knowledge, as is illustrated in the following example.

> "Technology negatively influences how people communicate. Some people use their cellphone constantly and do not even notice their environment."

In the above example the second sentence functions as a supporting premise for the claim in the first sentence, but the machine has to infer such an argumentative relationship from the knowledge that using a cellphone is a way of communication. According to [59], a major stumbling block for building highly performant systems for argumentation mining is the lack of this world, common sense or domain knowledge:

> "Knowledge is a major bottleneck to argument mining, given a controversial issue and a set of texts in which arguments can be found." [59]

One could manually draft sufficient knowledge patterns or provide a multitude of manually annotated examples to train an argumentation mining system, but both options have serious limitations given the variety of possible claims and their supporting or attacking premises in an ever changing world. The aim of this paper is to give an outline of current and potential approaches to automatically acquire such knowledge in the frame of argumentation mining. This knowledge would take the form of reusable representations learned from data.

First, this article will go deeper into current methods for argumentation mining and language understanding. Special attention will go to the difficulties of current natural language understanding in general and argumentation mining in particular. In a next section we will outline solutions of representation and deep learning, but will pinpoint their bottlenecks for learning common sense and world knowledge. The focus will then be on the promising paths to learn such knowledge. This article also points to many valuable technologies that have realized or could realize the automated acquisition of common sense and world knowledge.

2. Language understanding

Automated language understanding entails the interpretation of human language utterances and discourse, and their translation into a format that is usable by the machine in the execution of specific tasks such as machine translation, text analytics, information retrieval, visualization of information, decision making, or steering of machinery such as robots. Language understanding is largely concerned with recognizing actions and states and the actors that play a role in the events, which is often referred to as semantic role labeling, i.e., recognizing "who" does "what", "where", "when" and "how", and coreferent resolution, i.e., identifying coreferring expressions in a discourse (e.g., that refer to the same entity), the recognition of temporal and spatial relations between actions and entities, and identification of other relationships [55]. In addition, it involves detecting modality, that is, the factual status of a statement and can involve negation, possibility and obligational entailments of statements.

Past research in language understanding starting in the 1960s and 1970s was largely concerned with knowledge representation, that is, how knowledge in a discourse and individual sentences is encoded, how this knowledge is organized so that it can be retrieved at the appropriate moment using the available cues, and on how the current context and the background knowledge can be used to draw plausible inferences [44,61]. In the resulting automated systems, the semantic descriptions took the form of frame and script representations. These approaches are referred to as symbolic modelling or symbolic architectures. They cope with a serious scalability problem. It is impossible to handcraft the rules and semantic representations for all possible situations. As a result, these approaches were only applied in limited domains of discourse using very restricted, if any contextual knowledge giving little anticipatory power and autonomous adaptability to the models. However, the models over the years have provided a rich history of the study of inference and understanding of language on a small scale [8].

Later in the 1990s statistical machine learning models became mainstream approaches for natural language processing. With regard to language understanding, the typical recognitions discussed above continued to be researched, now focusing on generic standard labels to be used over different discourse genres (e.g., PropBank style semantic role labeling, OntoNotes, SemEval standards among which are SpaceEval and TempEval standards for spatial and temporal processing respectively). The last decade witnessed an interest in joint learning and inference of semantic labels with structured learning models, having the advantage that certain correlations between the occurrences of output labels can be captured during training or prediction. For instance, in [34] a spatial relation, the objects involved in this relation, and the relation's attributes are jointly recognized (e.g., "the man rides a horse" where "man" is spatially and externally connected to the "horse"). Structured learning algorithms are popular for semantic recognitions in language processing, especially maximum margin (max-margin) graphical models [63] and max-margin structured support vector machines [66]. These structured learning approaches are also applied to map language utterances to symbolic or logical meaning representations such as lambda notations, first order logic, or to specifically tailored languages such as robot programming languages and database query languages (for an overview see [37]), and they have gradually replaced fully symbolic approaches that model theoretical semantics and infer meaning by the compositional interpretation of syntactic parses.

The output of a language understanding system takes the form of semantic labels, task specific commands or statements in a programming language, or numerical representations such as vectors and matrices that capture the meaning of an utterance or of a larger discourse unit [37]. Very often natural language understanding entails the translation into a format that can be used by the machine to perform another task (e.g., decision support, machine interaction).

3. Why is natural language understanding such a difficult task for a machine?

Despite recent progress, language understanding remains a difficult task for a machine. First, meaning resides in the configuration of words, while sentence and discourse grammars remain ambiguous, so a lot of emphasis lies in the correct understanding of words in a particular context and grammatical configuration. Second, linguistic patterns, whether acquired manually or automatically, are insufficient to capture the meaning of freely uttered language. For instance, as meaning is resided in words and their contexts, given the large amount of content bearing words in a language and even a much larger number of possible contexts, an exhaustive set of patterns is difficult to acquire. This limitation is present in cases where the linguistic patterns needed to interpret language are manually acquired, as well as in cases when language corpora are manually annotated with ground truth labels to be used for training the machine learning models. Third, the interpretation of a discourse often entails making inferences with information found in the discourse (e.g., found in previously communicated utterances) as well as with background knowledge that humans in a communication setting assume that their audience possesses, or other contextual knowledge such as the physical context in which the language is interpreted. This process takes place when analyzing individual utterances or sentences and when interpreting a full discourse or parts of it. Inferencing in natural language understanding is traditionally realized by logical reasoning in case the representations are logical in nature [8], but recently statistical inference has emerged. In the latter case the representations are often numerical, and reasoning is performed in a Bayesian or algebraic framework, for instance, by applying simple mathematical operations on the numerical representations [28]. Finally, a lot of content is not expressed explicitly but resides in the mind of communicator and audience. This content includes common sense and world knowledge, domain knowledge and knowledge of the broader context in which the discourse functions (e.g., political or cultural context). For instance, if the text mentions the action of swimming, it entails that the person or animal swimming has jumped or gone into the water, even if the text would rarely mention this. Although models exist to reason with such common sense and world knowledge, the biggest problem is how to automatically acquire it.

4. What makes argumentation mining extra difficult?

Argumentation mining puts an extra dimension to the language understanding process. It is concerned about the argumentative role of language fragments in a full discourse [11,27,50,56]. As seen above meaning is resided in the configuration of words, but discourse markers that signal arguments to a proposition are often ambiguous or missing. Moreover, linguistic patterns, whether acquired manually or automatically, are insufficient to capture the meaning of freely uttered language and now we have to identify an extra layer of pragmatic understanding [7], that is, we have to discover the pragmatic function of clauses in a discourse for which we need to provide the machine with extra patterns, either by handcrafting them or by learning these based on annotated training examples. Finally but most importantly, language use is incomplete. A lot of content is not expressed explicitly but resides in the mind of communicator and audience. Reading between the lines can reveal subtle argumentation. For instance, in the text “I think people will end up concluding that at least some of the intensity of the monsoon in Queensland can be attributed to climate change. The waters off Australia are the warmest ever measured and those waters provide moisture to the atmosphere.”, the argumentative relation between the claim in the first sentence and the premise in the second, can only be detected if one knows that warm waters and moisture in the atmosphere are a sign of climate change. World, common sense, domain and contextual

knowledge play a primordial role in argumentation mining. Typically, content can function as an argument in one context, but not in another. For instance, a reference to "moisture" and "warm water" in a weather report is less likely to be used as an argument in a discussion on climate change.

5. Current proposed solutions: Representation learning and deep learning

During the last decade we have witnessed the emergence of representation learning as a major topic in machine learning, natural language processing and computer vision. The key idea is the use of low-level sensory data and the discovery of latent factors, expressed either in algebraic or probabilistic spaces, which may explain the generation of the low-level observed data [70]. One motivation is to ease feature engineering (which in semantic processing of language is often complex taking into account lexical and morpho-syntactic features) by making it as automatic as possible. A second goal is to mimic human perception and simulate concepts like distributed representations in a connectionist approach. Their success is especially apparent in signal and image processing, where neural network based deep learning models [36] were introduced to model the receptive fields of cells in the visual cortex. Deep learning has become a branch of machine learning where one attempts to model high-level abstractions by using a "deep" graph with multiple processing layers that perform linear and non-linear transformations. Deep learning models are mostly implemented as artificial neural networks or as their probabilistic interpretation in deep belief networks. These networks succeed in recognizing to a certain extent higher-level abstract concepts in visual data [70]. Such emergent architectures [23] are also used for modelling text data. For instance, neural networks or Bayesian network models are trained with a large collection of texts to learn distributional semantic models or language models that capture the contexts of a word. As a result, a word can be represented as a vector (i.e., a word embedding) the dimensions of which embed the contextual knowledge (e.g., [43]) or as a predictive language model (e.g., [19]). These works build further on older research of word representations that were learned from large text corpora that predict a word in the context of surrounding words or that learn topical concepts by considering co-occurrences in a full discourse. Examples are latent semantic indexing (LSI) [18], probabilistic graphical models such as probabilistic latent semantic indexing (pLSA) [32] and latent Dirichlet allocation (LDA) [3] and variants that learn to predict a word in a context window [20], neural network based models as in [15,16], and deep learning approaches based on recurrent networks (RNNs), convolutional neural networks (CNNs) and recursive tree RNNs, which infer latent concepts that are not apparent in the raw data (e.g., [2]). An advantage of the models is that apart from being a predictive language model [2], a word representation or embedding is created that embeds the context of the word in this representation. Word embeddings commonly take the form of word vectors, that is, each word is associated with a real valued vector in an N-dimensional space (usually $N = 50$-1000). The most famous architectures are the CBOW and Skip-gram NN-LM models [43], the latter referring to the skip-gram neural network language model, and the GloVe model [52]. These models are characterized by a simple, single-layered architecture, where their main differences regard the inputs (CBOW: words within short window; Skip-gram NNLM: the current word) and output (CBOW: the current word is predicted; Skip-gram NNLM: the context of the word is predicted). The hidden layer in these architectures implements a linear function, which guarantees an efficient computation. The trained weights of the nodes (dimensions) of the hidden (or exceptionally the output) layers of the network form the word embedding vectors. Such word embeddings may be effectively combined with visual and other perceptual input to produce cognitively plausible and multimodal embeddings (e.g., [5,14,31,62,69]).

Up until now, both language based and multimodal word embeddings were successful in simple text processing tasks such as word sense disambiguation, part-of-speech tagging, chunking and computation of semantic similarity between words (e.g., [14]), but they fail to predict semantic knowledge [58], characterized by a relational structure when trained in an unsupervised way with large collections of textual and visual data. Recently, we witness a growing interest in language models that predict content using neural networks that train on large text data sets. For instance, [53] predict the likely order of events or actions in a discourse possibly enriched with coreference chains [51]. Although results are preliminary, the above authors foresee a large applicability especially in automated understanding of narrative discourse. The grounding of language relations by visual data only starts to emerge. For instance, [26] ground causality of action verbs by using video data. We know very little on how we could use visual and multimodal representations to improve the accuracy and efficiency of language understanding by machines.

Another interesting development is the composition of the statistical representations into more abstract concepts. The principle of compositionality in language states that the meaning of a complex phrase is a function of the meanings of its parts and their mode of combination. For instance, in case of vector representations of individual words, recent work shows that the meaning of phrases or sentences could be inferred by mathematical operations on the word embedding vectors [28]. Such models often keep the syntax-driven compositionality of formal semantics while retaining the empirical nature of distributional models [45]. Current composition models rely on vector, matrix or tensor additions and multiplications, Frobenius algebras, packed dependency trees or fully supervised deep neural networks (for overviews see: [4,28,49]), and even provide logical entailment operations over vectors [30]. As a simple example, given the semantic space of word embeddings trained on a large corpus, if you know that Rome is the capital of Italy and want to infer the capital of China, then the equation Rome $-$ Italy $+$ China will return Beijing [43]. Compositionality is often linked to human's ability to produce and interpret novel utterances. However, we need to better understand how to efficiently and in a scalable way compose the meanings of larger pieces of text in a discourse by, for instance, exploiting the structure of language [41]. This might entail novel representation models and corresponding models of compositionality.

How well do these novel evolutions contribute to automated language understanding? Will these models of distributional semantics solve the problems of natural language understanding in general and argumentation mining in particular?

Current word representations and language models are restricted to capturing word meaning in a very rudimentary form by modelling contextual words and are currently used only in simple language processing tasks such as word sense disambiguation and word similarity computation [41]. It is true that the meaning of a word is defined by its local context, which is captured by the embeddings. Moreover, to a certain degree the embeddings replace tedious feature engineering in tasks of supervised recognition in language, e.g., part-of-speech tagging, sentence parsing, etc. The embeddings usually do not exploit the structure of language, and certainly not the structure of the perceptual reality, resulting in representations that are not specific enough for effective language understanding, and most importantly they lack real anticipatory power to predict semantics. In addition, given the huge number of possible combinations of words and their possible interpretations, training recognition models for language understanding that jointly learn the most plausible interpretation is time-wise very complex, a complexity that grows exponentially with the length of the sentence and discourse. The word embeddings do not contribute much to the problem of porting models trained in one subject domain to another subject domain. Our own research during the MUSE project on machine understanding of language in the context of interactive storytelling showed that the word embeddings used as features in a semantic role labeling task only could

slightly improve the results, or needed extra resources that are manually built to largely improve the performance [22,39]. Third and most importantly for the theme of this paper, current word embeddings and language models are successful representations in simple language processing tasks. However, they do not give a clear answer to the problem of the lack of common sense and world knowledge, which is one of the most important bottlenecks for language understanding and argumentation mining.

Could more sophisticated deep learning techniques help in language understanding and argumentation mining and offer a solution for these problems? Remember that argumentation mining is also inferring that a text phrase can function as a premise to support the claim expressed in another text phrase, inferring that a text phrase can function as a way to attack the claim expressed in another text phrase, inferring that a text phrase can function as a specific type of argument (e.g., argument from analogy) to support the claim expressed in another text phrase; inferring that a text phrase can function as a specific argument to support the claim expressed in another part of the discourse, while together they support a more abstract claim expressed in another text phrase, and so on. In the section below we show that researchers succeeded in predicting entailment, that is, whether one sentence entails the other, with the help of deep neural architectures and sufficient training data.

6. Reasoning about entailment

Let us look at an example of inference in the context of argumentation mining called reasoning about entailment. Entailment refers to detecting the asymmetric relationship between two sentences [9]. For instance, given the premise:

"A group of people standing in the snow with a mountain in the background."

and the hypothesis:

"People are outside." [57],

the task is to identify the nature of the relationship between these two sentences, or more specifically to recognize that the first sentence entails the second, identify a contradiction between them, or recognize that none of these two relationships is present (neutral relationship).

The most simple approach to this problem trains a classifier with a bag of words as features, that is, in this classification process, the two sentences are compared based on their word overlap. In another very popular approach a classifier is trained with hand engineered features (e.g., using parts-of-speech and parse features) derived from complex natural language processing pipelines. Currently there is an interest in using neural networks with attention mechanisms to accomplish this task and the results are very promising [57]. This last model uses as input the premise followed by the hypothesis. The words of these two sentences are input as a word by word sequence of word embeddings (trained on another large corpus) forming a recurrent neural network. More specifically, this model uses conditional encoding via two long short-term memory networks (LSTMs), one over the premise and one over the hypothesis conditioned on the representation of the premise. The implemented attention mechanism learns alignments between the output representation vectors of the words of the premise and hypothesis. During training the weights of the network are learned as well as the weights of the attention mechanism that will enforce the alignment between certain words of the premise and hypothesis. The output is one of the three classes mentioned above. This system reaches up to 83.5% accuracy on the test set of the Stanford Natural Language Inference corpus given sufficient training examples. The authors show that

their approach of inputting word embeddings combined with an attention mechanism that models the alignments between premise and hypothesis (e.g., that "snow" is related to "outside") is able to capture the world knowledge that "snow" entails being "outside". Still several questions remain that could be investigated in future research [57]. How well can a model trained on one corpus predict entailment in another corpus? How good are the word embeddings, which were used as inputs to transfer to another domain? How good are the contexts of words to predict whether one sentence entails another?

The PASCAL Recognizing Textual Entailment Challenge (2005–2011) recognized that textual entailment cannot be separated from the use of world knowledge: "We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge." [17]. Traditionally world knowledge has been captured as handcrafted symbolic knowledge as, for instance, in the Cyc project (http://www.opencyc.org). The question can be asked whether we could learn the needed common sense and world knowledge for entailment, argumentation mining or for language understanding in general from large amounts of preferably unannotated, raw data?

Humans perform the task of language understanding extremely efficiently only using a limited set of their resources in their working memory [70], which is the type of short-term memory used for immediate processing. Humans have in general no trouble when interpreting language to combine relevant evidences to infer meaning, determine hidden meanings and make explicit what was left implicit in the text, whether the context comes from phrases seen earlier in the text or from background knowledge. Only contextually relevant information comes to their mind when processing language where information is encoded and retrieved through the synaptic changes that direct the flow of energy through the brain. Cognitive studies evidence that humans are aided by the anticipatory power of the brain that predicts or "imagines" – circumstantial situations and outcomes of actions [25,68] and that this anticipatory power makes language processing extremely effective and fast [35,60]. In addition, human interpretation of language is scalable, flexible and adaptive. They constantly update their memory making it very adaptive to situations at hand. Machines currently lack these abilities when it comes to language understanding. The anticipatory models are built by the perceptive experiences of humans during their lifetime. So we think machines can simulate such an approach in order to learn the necessary common sense and world knowledge. The research questions are how to capture this knowledge, how to update it adaptively, and how to integrate it in the machine understanding processes.

7. Learning common sense and world knowledge

7.1. The generative lexicon

In the context of argumentation mining the use of a generative lexicon [54,59] has been proposed. In these studies a generative lexicon is defined as a knowledge representation that merges lexical items with knowledge of the world. More specifically, the lexicon captures knowledge of objects or entities and their physical properties (e.g., parts, material), knowledge to which the authors refer to as the "constitutive role" of an object. In this way objects are explicitly semantically typed. The "formal role" of an object involves its configuration with other objects (e.g., likely location and spatial relation of objects), while the function, use and purpose of an object are defined as its "telic role". The origin of the object and how it was created or produced are referred to as the "agentive" role. In addition, the lexicon ideally

contains domain knowledge, which may refer to typical predicate-argument structures, event structures (e.g., restriction of lexical types performing semantic roles in a sentence such as the typical location of an action), causal chains and role subtyping (e.g., the permissible order in occurrence of actions performed by a person in a narrative), in conjunction with the lexical data. No one doubts about the usefulness of such a lexicon for language understanding in general and argumentation mining in particular, but the problem is how to acquire such a lexicon and how to define a representation that is well suited to be easily integrated in language understanding systems.

This brings up again the question, can we automatically learn this knowledge from data and integrate in suitable representations? As explained in the next sections, we hypothesize that telic and agentive roles might be learned from large text collections, that constitutive and formal roles might be learned from large imagery collections and that event structures and causal chains might be learned from large text and imagery collections. We have little experience that can guide us to prove these hypotheses, but current research gives us strong indications of valuable paths to pursue in future work.

7.2. How can a machine learn common sense and world knowledge from text?

We believe that event semantics can be learned from unannotated text data. Inspired by current unsupervised dependency parsing [42] we could train probabilistic graphical models such as a Bayesian network with latent variables to estimate complex, relational language models [6]. Successful models can also be simulated and approximated with deep belief networks. Our current research shows that learning predictive event structures from a large corpus of texts encoding selectional knowledge in natural language gives promising results for the task of implicit semantic role labeling (i.e., complementing the event structures, "who" does "what", "where", "when" and "how") when these roles are left implicit in a text [21]. Recent research has shown that typical sequences of events – often forming causal chains of events – can be learned from text data as was demonstrated by [33] and [53], the latter using a deep learning long short-term memory network (e.g., if someone obtained a PhD at a university, it entails that he or she was studying at this university before obtaining the PhD) and that such predicted scripts of events positively influence the automated recognition of referents in a discourse [47].

It remains to be studied how such knowledge contributes to argumentation mining. In addition, arguments are often embedded into a context that indicates, for instance, circumstances, restrictions, concessions, comparisons, purposes, and various forms of elaborations. In terms of language realization, arguments and their related context may be included into a single sentence via coordination or subordination or may be realized as separate sentences. Can we learn the knowledge that links an argument (or attack) to a claim in a certain context from text data? In other words can we learn representations of words or phrases that incorporate how they would likely behave in an argumentative structure? Many research questions can here be studied.

7.3. How can a machine learn common sense and world knowledge from imagery?

Language data alone might not be sufficient to learn common sense and world knowledge. Humans learn such knowledge primarily from perceptual data such as visual data during their lifetime. Consider this very rough estimation. Given that we perceive our surrounding world at 25 frames per second, a child of 3 years who is awake 8 hours a day has been exposed to about 700 million frames (not counting the first 4 months), from which he or she learns. The typical physical behavior of objects might be learned by observing the world. We witness a number of very interesting artificial intelligence studies

in this respect. [71] automatically learn the physical properties of objects with deep learning. [1] learn world knowledge about objects by observing manipulations of robots of these objects and modeling these interactions that are visually observed with deep neural networks.

7.4. How can a machine learn common sense and world knowledge from imagery and language data?

When humans learn about the world their perception is often guided by the weak supervision of language data. We tell a child what an object is, what its properties are, often when watching or manipulating the object. This brings us to the question, how can a machine learn common sense and world knowledge from text and visual data? Can the machine ground language in visual perception and as such acquire the necessary common sense and world knowledge? Visual perception is generally considered less ambiguous than language. In the computer vision community large collections of images and their language descriptions are being created from which a machine can learn interesting perceptual knowledge (e.g., [24,40]). The models of [14,38] are capable of learning semantic common sense knowledge from images and their textual descriptions and of imagining visual scenes that may contain more objects than the ones mentioned in a text. Also the learning can be selective. For instance, it has recently been shown that when predicting attributes of objects, certain attributes of objects are better learned from language data, while other attributes are better acquired via visual perception [13]. All these works and also our own work in this respect are inspired by cognitive and neuroscience studies of human learning and language understanding.

Our current research in the CHIST-ERA project MUSTER (MUltimodal processing of Spatial and TEmporal expRessions: Towards Understanding Space and Time in Language Enhanced by Vision)[2] focuses on learning common sense and world knowledge through imagination [14]. More specifically, we learn temporal and spatial knowledge from visual data or from visual data aligned with textual data, and integrate this knowledge in suitable statistical representations. We are especially interested in the following research questions. How can we automatically create text representations in the form of single-word and multi-word embeddings that integrate perceptual knowledge in the representations of objects, actions, their spatial and temporal relations, which refers to the problem of multi-modal representation construction. With initiatives such as MUSTER we hope to acquire a small fraction of the world knowledge needed for language understanding. In current submitted work we are able to automatically capture spatial knowledge from visual and language data (e.g., the configuration of spatial objects given the language utterance "A man rides a horse.") and we can even predict spatial information about objects never seen in the training data, but that have visual or functional resemblance with objects of the training data [12], making the knowledge acquisition an adaptive process.

We note that all the above cited works capture common sense and world knowledge in the form of numerical representations (e.g., vectors). We have shown that it is possible to automatically acquire some basic common sense and world knowledge needed for refined language understanding. However, the knowledge that we are able to learn is certainly not sufficient in an argumentation mining framework, which requires a much larger set of knowledge representations including pragmatic knowledge which is more difficult to acquire, but the above steps form a good start.

[2] http://www.chistera.eu/projects/muster

7.5. How do we integrate learned representations in the language understanding process and in argumentation mining?

This is the most difficult question for which we do not yet have a straightforward answer, but which we also are researching in the above MUSTER project. Current cognitive and neuroscience studies show that humans have in general no trouble when interpreting language to combine relevant evidences and "imagined" representations to infer meaning [35,60]. Such studies might guide the development of effective models for machine understanding of language.

8. Conclusions

Argumentation mining is a difficult task for a machine requiring a substantial amount of common sense, world, domain and contextual knowledge. Throughout the paper we have shown the need to automatically acquire such knowledge. Deep (or not so deep) learning models can model context and can be helpful in the acquisition of world knowledge, but we have little experience in capturing such knowledge into reusable representations except for building simple word embeddings that encode words occurring in the context of the target word. Models of how humans have acquired such knowledge during their life in a multimodal environment and how they use such knowledge in the most effective way when understanding language have inspired current research in automatically acquiring common sense and world knowledge from textual and perceptual sources.

This article has raised more questions than it has given answers, but we hope that these thoughts are inspiring to drive future research. It is worthwhile to study representation learning and deep learning models in the context of argumentation mining, but this would require developing own models that can integrate the contextual knowledge that is needed in argumentation mining.

Acknowledgement

This work contributes to the goals of the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307, of which the author is a member, and to the goals of the CHIST-ERA MUSTER (MUltimodal processing of Spatial and TEmporal expRessions) project, in which the author is a partner.

References

[1] P. Agrawal, A. Nair, P. Abbeel, J. Malik and S. Levine, Learning to poke by poking: Experiential Learning of intuitive physics, *Advances in Neural Information Processing Systems* (2016).
[2] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, A neural probabilistic language model, *Journal of Machine Learning Research* **3** (2003), 1137–1155.
[3] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research* **3** (2003), 993–1022.
[4] S.R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C.D. Manning and C. Potts, A fast unified model for parsing and sentence understanding, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, 2016, pp. 1466–1477.
[5] E. Bruni, N. Tran and M. Baroni, Multimodal distributional semantics, *Journal of Artificial Intelligence Research* **49** (2014), 1–47.

[6] T. Brychcín, Latent tree language model, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2016, pp. 436–446.
[7] K. Budzynska, M. Janier, C. Reed and P. Saint-Dizier, Theoretical foundations for illocutionary structure parsing, *Argument & Computation* **7**(1) (2016), 91–108.
[8] H. Bunt, J. Bos and S. Pulman, Introduction: Computing meaning: Annotation, representation, and inference, in: *Computing Meaning*, H. Bunt, J. Bos and S. Pulman, eds, Text, Speech and Language Technology, Vol. 47, Springer, 2014, pp. 1–9.
[9] E. Cabrio and S. Villata, Detecting bipolar semantic relations among natural language arguments with textual entailment: A study, in: *Proceedings of the Joint Symposium on Semantic Processing (JSSP-2013)*, 2013.
[10] E. Cabrio, S. Villata and A.Z. Wyner (eds), *Proceedings of the Workshop on Frontiers and Connections Between Argumentation Theory and Natural Language Processing*, CEUR Workshop Proceedings, Vol. 1341, CEUR-WS.org, 2015.
[11] C. Cardie, N. Green, I. Gurevych, G. Hirst, D. Litman, S. Muresan, G. Petasis, M. Stede, M. Walker and J. Wiebe (eds), *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, ACL, 2015.
[12] G. Collell and M.-F. Moens, Submitted (2017).
[13] G. Collell, T. Zhang and M.-F. Moens, Learning to predict: A fast re-constructive method to generate multimodal embeddings, in: *Proceedings of the NIPS Workshop on Representation Learning in Artificial and Biological Neural Networks (MLINI 2016)*, 2016.
[14] G. Collell, T. Zhang and M.-F. Moens, Imagined visual representations as multimodal embeddings, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, AAAI, 2017.
[15] R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: *Proceedings of the International Conference on Machine Learning*, Vol. 307, ACM, 2008, pp. 160–167.
[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P.P. Kuksa, Natural language processing (almost) from scratch, *Journal of Machine Learning Research* **12** (2011), 2493–2537.
[17] I. Dagan, O. Glickman and B. Magnini, The PASCAL recognising textual entailment challenge, in: *MLCW*, Lecture Notes in Computer Science, Vol. 3944, Springer, 2005, pp. 177–190.
[18] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* **41**(6) (1990), 391–407.
[19] K. Deschacht, J. De Belder and M.-F. Moens, The latent words language model, *Computer Speech and Language* **26**(5) (2012), 384–409. doi:10.1016/j.csl.2012.04.001.
[20] K. Deschacht and M.-F. Moens, Semi-supervised semantic role labeling using the latent words language model, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2009, pp. 21–29.
[21] Q.N. Do Thi, S. Bethard and M.-F. Moens, Improving Implicit Semantic Role Labeling by Predicting Semantic Frame Arguments, *CoRR* (2017), arXiv:1704.02709v1.
[22] Q.N. Do Thi, S. Bethard and M.-F. Moens, Domain adaptation in semantic role labeling using a neural language model and linguistic resources, *IEEE/ACM Transactions on Speech and Language Processing* **23**(11) (2015), 1812–1823. doi:10.1109/TASLP.2015.2449072.
[23] K.L. Downing, *Intelligence Emerging: Adaptivity and Search in Evolving Neural Systems*, The MIT Press, 2015.
[24] D. Elliott and F. Keller, Image description using visual dependency representations, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2013, pp. 1292–1302.
[25] K. Friston, The free-energy principle: A unified brain theory?, *Nature Reviews Neuroscience* **11**(2) (2010), 127–138. doi:10.1038/nrn2787.
[26] Q. Gao, M. Doering, S. Yang and J.Y. Chai, Physical causality of action verbs in grounded language understanding, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, 2016.
[27] N. Green, K. Ashley, D. Litman, C. Reed and V. Walker (eds), *Proceedings of the First Workshop on Argument Mining, Hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, ACL, 2014.
[28] E. Grefenstette and M. Sadrzadeh, Concrete models and empirical evaluations for the categorical compositional distributional model of meaning, *Computational Linguistics* **41** (2014), 71–118. doi:10.1162/COLI_a_00209.
[29] I. Habernal and I. Gurevych, Argumentation mining in user-generated Web discourse, *Computational Linguistics* (2017) abs/1601.02403.
[30] J. Henderson and D.N. Popa, A vector space for distributional semantics for entailment, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, 2016, pp. 2052–2062.
[31] F. Hill, R. Reichart and A. Korhonen, Multi-modal models for concrete and abstract concept meaning, *Transactions of the Association for Computational Linguistics* **2** (2014), 285–296.
[32] T. Hofmann, Probabilistic latent semantic indexing, in: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, ACM, 1999, pp. 50–57.

[33] B. Jans, S. Bethard, I. Vulic and M.-F. Moens, Skip n-grams and ranking functions for predicting script events, in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, ACL, 2012, pp. 336–344.
[34] P. Kordjamshidi and M.-F. Moens, Global machine learning for spatial ontology population, *Journal of Web Semantics* **30** (2015), 3–21. doi:10.1016/j.websem.2014.06.001.
[35] C.A. Kurby and J.M. Zacks, Situation models in naturalistic comprehension, Cambridge University Press, 2015, pp. 59–76.
[36] Y. Lecun, Y. Bengio and G. Hinton, Deep learning, *Nature* **521**(7553) (2015), 436–444. doi:10.1038/nature14539.
[37] P. Liang and C. Potts, Bringing machine learning and compositional semantics together, *Annual Review of Linguistics* **1**(1) (2015), 355–376. doi:10.1146/annurev-linguist-030514-125312.
[38] X. Lin and D. Parikh, Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2015, pp. 2984–2993.
[39] O. Ludwig, Q. Do, C. Smith, M. Cavazza and M.-F. Moens, Learning to extract action descriptions from narrative text, *IEEE Transactions on Computational Intelligence and AI in Games* **PP**(99) (2017), 1–1.
[40] M. Malinowski, M. Rohrbach and M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in: *Proceedings of the International Conference on Computer Vision (ICCV 2015)*, 2015, pp. 1–9.
[41] C. Manning, Understanding human language: Can NLP and deep learning help?, in: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'16, ACM, New York, NY, USA, 2016, pp. 1–1.
[42] D. Marecek and M. Straka, Stop-probability estimates computed on a large corpus improve unsupervised dependency parsing, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL, 2013, pp. 281–290.
[43] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient Estimation of Word Representations in Vector Space, *CoRR* (2013) abs/1301.3781.
[44] M. Minsky, A framework for representing knowledge, New York: McGraw-Hill, 1975, pp. 211–277.
[45] J. Mitchell and M. Lapata, Vector-based models of semantic composition, in: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, The Association for Computer Linguistics, 2008, pp. 236–244.
[46] R. Mochales Palau and M.-F. Moens, Argumentation mining, *Artificial Intelligence and Law* **19**(1) (2011), 1–22. doi:10.1007/s10506-010-9104-x.
[47] A. Modi, I. Titov, V. Demberg, A. Sayeed and M. Pinkal, Modeling semantic expectations: Using script knowledge for referent prediction, *Transactions of the Association for Computational Linguistics* (2016).
[48] M.-F. Moens, E. Boiy, R. Mochales Palau and C. Reed, Automatic detection of arguments in legal texts, in: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law*, ACM, 2007, pp. 225–230.
[49] D. Paperno and M. Baroni, When the whole is less than the sum of its parts: How composition affects PMI values in distributional semantic vectors, *Computational Linguistics* **42**(2) (2016), 345–350. doi:10.1162/COLI_a_00250.
[50] A. Peldszus and M. Stede, Joint prediction in MST-style discourse parsing for argumentation mining, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015*, 2015, pp. 938–948.
[51] H. Peng, Y. Song and D. Roth, Event detection and co-reference with minimal supervision, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2016, pp. 392–402.
[52] J. Pennington, R. Socher and C. Manning, GloVe: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2014, pp. 1532–1543.
[53] K. Pichotta and R.J. Mooney, Learning statistical scripts with LSTM recurrent neural networks, in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, AAAI Press, 2016, pp. 2800–2806.
[54] J. Pustejovsky, *The Generative Lexicon*, MIT Press, 1995.
[55] A. Ram and K. Moorman, *Understanding Language Understanding: Computational Models of Reading*, MIT Press, 1999.
[56] C. Reed, K. Ashley, C. Cardie, N. Green, I. Gurevych, D. Litman, G. Petasis, N. Slonim and V. Walker (eds), *Proceedings of the Third Workshop on Argument Mining, Hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*, ACL, 2016.
[57] T. Rocktäschel, E. Grefenstette, K.M. Hermann, T. Kocisky and P. Blunsom, Reasoning about entailment with neural attention, in: *Proceedings of the International Conference on Learning Representations (ICLR 2016)*, 2016.
[58] D. Rubinstein, E. Levi, R. Schwartz and A. Rappoport, How well do distributional models capture different types of semantic knowledge?, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, 2015, pp. 726–730.
[59] P. Saint-Dizier, Challenges of argument mining: Generating an argument synthesis based on the qualia structure, in: *INLG 2016 – Proceedings of the Ninth International Natural Language Generation Conference*, 2016, pp. 79–83.

[60] D.L. Schacter and K.P. Madore, Remembering the past and imagining the future: Identifying and enhancing the contribution of episodic memory, *Memory Studies* **9**(3) (2016), 245–255. doi:10.1177/1750698016645230.
[61] R.C. Schank, *Conceptual Information Processing*, Amsterdam: North Holland, 1975.
[62] C. Silberer and M. Lapata, Learning grounded meaning representations with autoencoders, in: *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics*, ACL, 2014, pp. 721–732.
[63] B. Taskar, V. Chatalbashev, D. Koller and C. Guestrin, Learning structured prediction models: A large margin approach, in: *Proceedings of the 22nd International Conference on Machine Learning*, ICML'05, ACM, New York, NY, USA, 2005, pp. 896–903.
[64] S. Teufel and M. Moens, Discourse-level argumentation in scientific articles: Human and automatic annotation, in: *Towards Standards and Tools for Discourse Tagging*, 1999, pp. 84–93.
[65] S.E. Toulmin, *The Uses of Argument*, Cambridge University Press, 1958.
[66] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun, Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research* **6** (2005), 1453–1484.
[67] F.H. van Eemeren, B. Garssen, E.C.W. Krabbe, A.F. Snoeck Henkemans, B. Verheij and J.H.M. Wagemans, *Handbook of Argumentation Theory*, Berlin: Springer, 1996.
[68] D. Vernon, *Artificial Cognitive Systems: A Primer*, The MIT Press, 2014.
[69] I. Vulić, D. Kiela, S. Clark and M.-F. Moens, Multi-modal representations for improved bilingual lexicon learning, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, 2016, pp. 188–194.
[70] J. Weng, *Natural and Artificial Intelligence*, BMI Press, 2013.
[71] J. Wu, I. Yildirim, J.J. Lim, B. Freeman and J. Tenenbaum, Galileo: Perceiving physical object properties by integrating a physics engine with deep learning, *Advances in Neural Information Processing Systems* **28** (2015), 127–135. doi:10.1007/978-3-319-26532-2_15.