THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2016 Vol. 17 No. 1 (ISSN 1727-5997)

Feature Articles

Argument Mining . . . . . *Katarzyna Budzynska and Serena Villata* 1
Entity Coreference Resolution. . . . . *Vincent Ng* 7
Cognitive Systems: Argument and Cognition. . . . . *Antonis Kakas and Loizos Michael* 14
AI for Traffic Analytics. . . . . *Raghava Mutharaju, Freddy Lécué, Jeff Z. Pan, Jiewen Wu and Pascal Hitzler* 21

Book Review

Cognitive Computing: Theory and Applications. . . . . *Pawan Lingras* 27

Announcements

Related Conferences, Call For Papers/Participants 28

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee
2) Feature Articles
3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/~ley/db/index.html).*

Argument Mining

Katarzyna Budzynska , Serena Villata

***Abstract*—Fast, automatic processing of texts posted on the Internet to find positive and negative attitudes towards products and companies gave *sentiment analysis*, an area of text mining, a significant application in predicting trends on stock markets. *Opinion mining* further extended the scope of the search to help companies, such as those specialising in media analysis, to automate extraction of people's beliefs about products, institutions, politicians, celebrities. Now, *argument mining* goes one more step ahead to provide us with instant information not only about what attitudes and opinions people hold, but also about arguments which people give in favour (pro) and against (con) these attitudes and opinions. When this rapidly developing technology will mature, it will allow us to automatically and empirically explore vast amount of social media data (rather than seeking advices and opinions of experts) to give us answers such as why people decided to vote for one presidential candidate rather than the other.**

***Index Terms*—Argumentation, debating, computational linguistics, text mining.**

I. Introduction

ARGUMENT mining (also referred to or associated with argumentation mining, computational argumentation or debating technologies) is a new and rapidly growing area of natural language processing, and more specifically – text mining, which both are disciplines belonging to computational linguistics and artificial intelligence (see e.g., [28], [31], [26], [5] for a more detailed overview). Its goal is to develop methods and techniques which allow for automatic identification and extraction of argument data from large resources of natural language texts.

The broad area of text mining aims to provide robust tools, methods and techniques which allow for speeding up processing, interpreting and making sense out of the large amount of datasets of texts in natural language. The growth of this area is driven by a problem of the explosion of data available on the Internet. While having vast amount of data is an unquestionable value, the resources become of limited usefulness if we can not process them efficiently in a relatively short time and with low cost. If a company, such as Amazon or eBay, receives a lot of feedback from customers, but it takes months to analyse reviews posted on the company's webpage during just one day, then such a feedback will have very limited use for the company to understand what people like or dislike about their products and service. An extreme of this problem is referred to as Big Data, i.e. a situation when data is produced faster than users of these data and standard computational methods can process them.

Argument mining is a natural continuation and evolution of sentiment analysis and opinion mining – two areas of text mining which became very successful and important both academically and commercially. In sentiment analysis, the work focuses on extracting people's attitudes (positive, neutral, negative) towards persons, events or products. One commercially successful application of this research area is stock market where it is possible to relatively quickly process vast amount of resources such as news and social media to extract information about trends and tendencies on the market and to predict changes in stock prices. In opinion mining, the work aims to mine people's opinions about persons, events or products, e.g. the opinion that UK economy will be stronger without contributing a vast amount of money to the EU budget or the opinion that the UK economy will be weakened without the access to the common EU market. Its main commercial application is media analysis which monitors media to identify people's reactions for new products, companies, presidential candidates and so on. Argument mining, on the other hand, allows for recognising not only *what* attitudes and opinions people hold, but also *why* they hold them.

The growth of the commercial interests in the area of argument mining is manifested through the involvement of companies in several academic projects as well as the development of techniques such as IBM's Watson Debater (see e.g., www.arg.tech/ibmdebater) which searches for arguments pro and con regarding a given topic in Wikipedia articles.

II. Pipeline of natural language processing techniques applied to argument mining

Argument mining pipeline comprises of linguistic and computational part (see Figure 1). The *linguistic part* aims to develop large corpora, which are datasets of manually annotated (analysed) argument data, evaluated by measuring the level of inter-annotator agreement. The *computational part* of argument mining pipeline aims to develop grammars (structural approach) and classifiers (statistical approach) to automatically annotate arguments and the performance of the system is then evaluated by measures such as accuracy or F_1 score. The ultimate goal of the pipeline is to process real arguments in natural language texts (such as arguments formulated on Wikipedia) in order to provide as an output only these information which are valuable for us, i.e. structured argument data.

Katarzyna Budzynska is an associate professor (senior lecturer) in the Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland & a lecturer and Dundee fellow in the School of Computing at the University of Dundee, UK, e-mail: budzynska.argdiap@gmail.com (see www.argdiap.pl/budzynska/). Together with Professor Chris Reed, she runs the ***Center for Argument Technology*** (see www.arg.tech).

Serena Villata is a researcher (CR1) at the Centre National de la Recherche Scientifique (CNRS) in the I3S Laboratory, France, e-mail:villata@i3s.unice.fr (see www.i3s.unice.fr/˜villata/).

This report is a short version of our article "Processing Argumentation in Natural Language Texts" which will appear in *Handbook of Formal Argumentation* in 2017 [5].

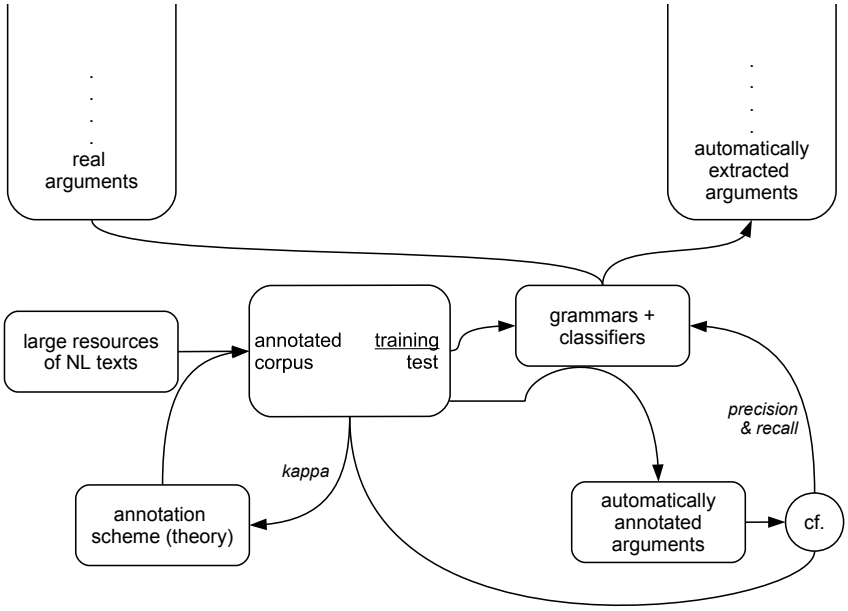


Fig. 1. A pipeline of natural language processing techniques applied to argument mining.

A. Databases of texts in natural language

The first step of the linguistic part of the pipeline starts with the task of *collecting large resources of natural language texts* (see "large resources of NL text" box in Figure 1) which then can be used for training and testing of argument mining algorithms. For example, Palau and Moens used dataset consisting of 92,190 words, 2,571 sentences divided in 47 documents from the European Court of Human Rights [27]; Habernal and Gurevych collected database comprising of 90,000 words in 340 documents of user-generated web discourse [15]; Garcia-Villalba and Saint-Dizier used 21,500 words in 50 texts as a test corpus [39].

Typically, the task of argument mining is narrowed down to the specific type of discourse (genre), since algorithms use the linguistic surface for argument recognition with none or little knowledge about the world, discourse context or deeper pragmatic level of a text. Genres studied up to date range from legal texts (e.g., [27], [2]); mediation (e.g., [19]); scientific papers (e.g., [38], [20]); online comments (e.g., [39], [30], [14], [40]); political debates (e.g., [18], [10]); technical texts (e.g., [33]); online debates (e.g., [41], [6], [35], [3], [16]); persuasive essays (e.g., [36], [13]); to Wikipedia articles (e.g., [1], [25]).

B. Theories & annotation schemes

The next step of argument mining pipeline consists of *choosing a model of argumentation which is then used to develop an annotation scheme* for analysing arguments in natural language texts. An annotation scheme for argumentative texts is a set of labels (tags) which defines arguments and their aspects for annotators (analysts) to use for structuring the dataset.

In the literature, there is a variety of different annotation schemes which aim to balance between efficiency (simpler schemes will be quicker and easier to annotate) and adequacy (more specific sets of labels will be better tailored to describing given aspects of argumentation or given genre). In one of the first work in the argument mining [27], Palau and Moens choose a basic, intuitive conceptualisation of argument structure which consists of three labels: (a) premise: statements which provides a support; (b) conclusion: statements which are supported; (c) argument: a full structure comprising of premises and conclusion.

In her Argumentative Zoning work [38], Teufel uses more complex set of labels specifically tailored for mining argumentation in scientific texts: (a) background: general scientific background; (b) other: neutral descriptions of other people's work; (c) own: neutral descriptions of the own, new work; (d) aim: statements of the particular aim of the current paper; (e) textual: statements of textual organization of the current paper (e.g. "In chapter 1, we introduce..."); (f) contrast: contrastive or comparative statements about other work; explicit mention of weaknesses of other work; and (g) basis: statements that own work is based on other work.

Peldszus and Stede [31] introduce an annotation scheme drawing on different ideas from the literature and their practical experiences with analysing texts in the Potsdam Commentary Corpus [37]. The schema follows Freeman's idea of using the moves of proponent and challenger in a basic dialectical situation as a model of argument structure [12] with the representation of the rebutting/undercutting distinction and complex attack- and counter-attack constellations. Their scheme considers five kinds of supports among premises and the claim: (a) basic argument, (b) linked support, (c) multiple support, (d) serial support, and (e) example support; four kinds of challenger's attacks of the proponent's argument: (a) rebut a conclusion, (b) rebut a premise, (c) undercut an argument, (d) and support of a rebutter; and four proponent's counter-

attacks of the challenger's attack: (a) rebut a rebutter, (b) rebut an undercutter, (c) undercut a rebutter, and (d) undercut an undercutter.

An annotation scheme which considers the broad dialogical context of argumentation was proposed in [4]. Building upon Inference Anchoring Theory, Budzynska *et al.* extend the set of tags for arguments pro and con with dialogue structures and illocutionary structures [34] with two groups of tags. For the MM2012 corpus (www.corpora.aifdb.org/mm2012), the annotators could use the following tags associated with individual moves of a speaker in the dialogue: (a) asserting, (b) questioning (pure, assertive, and rhetorical), (c) challenging (pure, assertive, and rhetorical), and (d) popular conceding (statement that is assumed to belong to general knowledge); and for tags associated with the interactions between speaker(s)' moves in the dialogue, the annotators could choose between: (a) agreeing, (b) disagreeing, and (c) arguing.

C. Manual annotation & corpora

The *process of annotation* starts with segmenting (splitting) the text into elementary discourse units (EDUs) or in fact into argumentative discourse units (ADUs). Annotators use software tools such as the `arggraph` DTD[1], the RSTTool[2], the Glozz annotation tool[3] and OVA+[4], which help them to assign labels from the annotation schemeset to ADUs directly in a code.

Next, the annotated data have to be stored as a corpus. For example, the IBM Debating Technologies corpus[5] contains three different datasets: the dataset for automatic detection of claims and evidence in the context of controversial topics (1,392 labeled claims for 33 different topics) [1], and its extended version (2,294 labeled claims and 4,690 labeled evidence for 58 different topics). Another resource is the Internet Argument Corpus (IAC) which provides analyses of political debate on Internet forums. It consists of 11,000 discussions and 390,000 posts annotated for topic, stance, degree of agreement, sarcasm, and nastiness among others [41]. The UKPConvArg1[6] corpus is a recently released dataset composed of 16,000 pairs of arguments over 32 topics annotated with the relation "A is more convincing than B" [16].

As the manual annotation is a highly time-consuming task, sharing and reusing analysed data becomes a real value. This is an objective of the freely accessible database AIFdb (www.aifdb.org) [24] which hosts multiple corpora (www.corpora.aifdb.org, see Figure 2). The key advantage of AIFdb is that it uses a standard for argument representation – the Argument Interchange Format, AIF [7]. The corpora were either originally annotated according to this format – such as the MM2012 corpus described above; or imported to the AIFdb – such as the Internet Argument Corpus developed by the group in Santa Cruz [41]. Currently this database has

[1] www.github.com/peldszus/arg-microtexts/blob/master/corpus/arggraph.dtd
[2] www.wagsoft.com/RSTTool/
[3] www.glozz.org
[4] www.ova.arg-tech.org/
[5] www.researchweb.watson.ibm.com/haifa/dept/vst/mlta_data.shtml
[6] www.github.com/UKPLab/acl2016-convincing-arguments

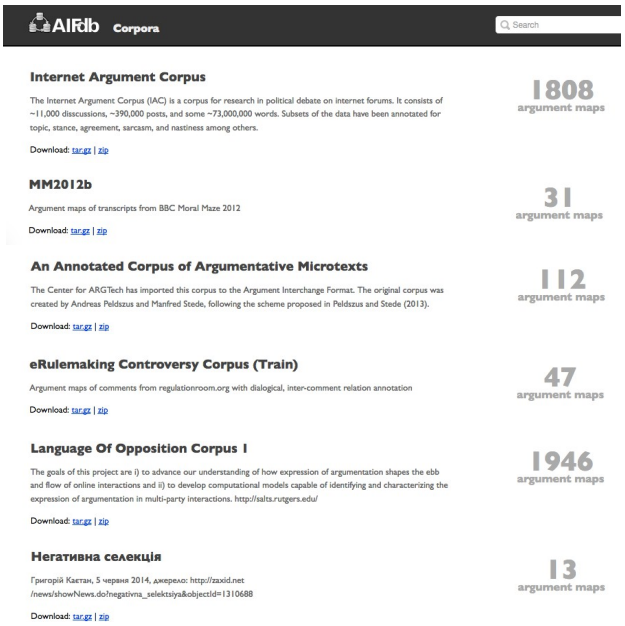


Fig. 2. Freely available AIFdb corpora.

300-500 unique users per month; stores 1,600,000 words and almost 57,000 annotated arguments in 15 languages (statistics obtained in November 2016).

D. Evaluation of manual step of annotation

The last step of the linguistic part of the argument mining pipeline is the *evaluation of the quality of a manual annotation* for which two comparison measures are the most typically used: (a) simple agreement which calculates a proportion (percentage) of matches between the analyses delivered by two annotators; or (b) several different kappa κ measures. The first one does not take into account the possibility of random matches, as if the annotators were tossing a coin and then assig labels according to the result. Thus, κ measures was introduced, amongst which the most popular one – Cohen's kappa [8] – shows the agreement between two annotators who each classify N items (e.g., ADUs) into C mutually exclusive categories (tags):

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement.

The following scale [22] aims to interpret the level of agreement: 0.41-0.6 means moderate agreement, 0.6-0.8 is treated as substantial agreement, and 0.81-1 is assumed to be almost perfect agreement.

Recently, Duthie *et al.* proposed a new CASS metric, Combined Argument Similarity Score [9], which helps to avoid double penalising if the analysis involves different levels such as both segmentation and identification of argument structure. Arguments do not always span full sentences and automatic solutions may miss some tokens, this can then have a knock

on effect on the argumentative or dialogical structure with text spans being either larger or smaller and the κ penalising for this twice.

As an example, in eRulemaking corpus [29], the inter-annotator agreement was measured on 30% of the data resulting in Cohens κ of 0.73; in the MM2012 corpus [4], kappa for three types of illocutionary connections (arguing, agreeing and disagreeing) was $\kappa = 0.76$; in the persuasive essays corpus [36] inter-annotator agreement was measured on 90 persuasive essays for three annotators resulting in a Krippendorff's inter-rater agreement of α = 0.81[7]; and in the argumentative microtexts corpus [32] three annotators achieved an agreement of Fleiss κ = 0.83[8] for the full task.

E. NLP techniques

The next step *moves us to the computational part of the argument mining pipeline* (see "grammars + classifiers" box in Figure 1). In principle, there are two basic styles of automation (in practice, they are often combined to form a hybrid approach): (a) the structural approach, i.e. grammars (hand coded set of rules); and (b) the statistical approach, i.e. machine learning (general learning algorithms).

In the structural approach, a linguist looks through a selected fragment of a corpus (training corpus which in this case is more often referred to as a development corpus) and aims to find patterns between different lexical cues in the text and categories in the annotation scheme. For instance, in a given corpus it might be observed that arguments are linguistically signalled by words such as "because", "since", "therefore". Then, the linguist formulates rules describing these patterns in a grammar. The statistical approach 'replaces' a linguist with an algorithm. In the same way as a human, a system will also look for patterns, however, this time statistically on a larger sample of the training corpus.

A lot of work in argument mining applies the typical, 'off the shelf' NLP methods and techniques which are then further enriched to adapt them to a specific domain or genre of argumentative texts. Apart from discourse indicators such as "because", "since", "therefore" (see e.g., [21], [17]), different projects employ various additional information to improve the searching process for arguments such as e.g., argumentation schemes [11], the dialogical context [4], and the semantic context [6], or combination of different cues and techniques.

An example of structural approach is the work by Garcia-Villalba and Saint-Dizier [39] who investigate how an automatic recognition of arguments can be implemented in the Dislog programming language on the <TextCoop> discourse processing platform, or more precisely – whether argument mining techniques allows for capturing consumers' motivations expressed in reviews why they like or dislike a product. For instance, a justification gives a reason for the evaluation expressed in the review: "The hotel is 2 stars [JUSTIFICATION due to the lack of bar and restaurant facilities]" can be classified as a justification, which general abstract schema is "X is Eval because of Fact*" where Eval denotes the evaluative expression and Fact* is a set of facts acting as justifications.

The majority of the work in argument mining employs, however, the statistical approach. Among them, Lippi and Torroni [25] present a framework to detect claims in unstructured corpora without necessity of resorting to contextual information. Their methodology is driven by the observation that argumentative sentences are often characterized by common rhetorical structures. As the structure of a sentence could be highly informative for argument detection, and in particular for the identification of a claim, the authors choose constituency parse trees for representing such information. They therefore build a claim detection system based on a Support Vector Machine (SVM) classifier which aims at capturing similarities between parse trees through Tree Kernels, a method used to measure the similarity between two trees by evaluating the number of their common substructures. Habernal and Gurevych [16] aim to assess qualitative properties of the arguments to explain why one argument is more convincing than the other one. Based on a corpus of 26,000 annotated explanations written in natural language, two tasks are proposed on this data set, i.e., the prediction of the full label distribution; and the classification of the types of flaws in less convincing arguments. Cabrio and Villata [6] propose a framework to predict the relations among arguments using textual entailment (TE), a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. TE is then coupled together with an abstract bipolar argumentation system which allows to identify the arguments that are accepted in online debates. The accuracy of this approach in identifying the relations among the arguments in a debate is about 75%.

F. Automatically annotated data

A system developed in the NLP stage is then used to *process raw, unannotated text in order to automatically extract arguments*. These texts have to be the same as the set of texts which was manually annotated and stored as a test corpus (see Figure 1). This step can be treated as an automated equivalent for manual annotation and corpus development.

Figure 3 shows an example of the output of a software tool. The <TextCoop> platform produces automatic segmentation and annotation. The text is split into argumentative discourse units (ADUs) which contain a minimal meaningful building blocks of a discourse with argumentative function. These propositional contents are presented as text in purple. Then, the system assigns illocutionary, communicative intentions (text in green) to ADUs of a type of assertions, rhetorical questions (RQ), and so on; as well as polymorphic types to represent the ambiguity (or underspecification) such as RQ-AQ which means that an ADU can be interpreted as having rhetorical questioning or assertive questioning illocution.

G. Evaluation of automatic step of annotation

The last step in the argument mining pipeline is the *evaluation of the quality of the automatic annotation*. A simple

[7]Krippendorff's α is a statistical measure of the agreement achieved when coding a set of units of analysis in terms of the values of a variable.

[8]Fleiss' κ assesses the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items.

```
< utterance speaker = "lj" illoc = " standard_assertion " > < textunit nb = " 215 "
> it was a ghastly aberration < /textunit> < /utterance>

<utterance speaker = "cl" illoc = "RQ"> <textunit nb= "216"> or was it in fact
typical ? < /textunit> </utterance> .

<utterance speaker = "cl" illoc = "RQ-AQ"> <textunit nb = "217">
was it the product of a policy that was unsustainable that could
only be pursued by increasing repression? < /textunit> < /utterance>.
```

Fig. 3. Example of data automatically annotated using the <TextCoop> platform for discourse processing of dialogical arguments in the MM2012 corpus.

measure, which is often used for this task, is accuracy, i.e. a proportion (percentage) of matches between manual and machine assignments of labels. If we want, however, to capture further, more detailed information about how well the system performed in mining arguments, a group of metrics: recall, precision and F_1 score, can be used. Let true positives, tp, will be a count how many times a machine assigned a label to the same text span as human analyst did; true negatives, tn – how often the machine did not assign a label to an ADU and the human did not either; false positives, fp – how often the machine assigned the label to a given text span while human did not; and false negatives, fn – how often the machine did not assign the label to a segment to which human made the assignment. Then:

– recall measures how many times the system did not recognise ("missed out") arguments:

$$R = \frac{tp}{tp + fn}$$

– precision shows how many times the program found arguments correctly:

$$P = \frac{tp}{tp + fp}$$

– F_1 score (F-score, F-measure) provides the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

If the matrices are computed and the performance of the system turns out to be not satisfactory, then we need to repeat the computational part of the process of argument mining trying to improve NLP techniques and methods we are using.

In their work, e.g., Palau and Moens obtain the following F_1 scores: 0.68 for the classification of premises; 0.74 for the classification of conclusions; and 0.6 for the determination of argument structures [27]. In [23], Lawrence and Reed aim to use argumentation schemes and combine different techniques in order to improve the success of recognising argument structure. This allows them to obtain the following results: for the technique of Discourse Indicators the system delivers precision of 1, recall of 0.08, and F_1 score of 0.15; for the technique of Topic Similarity the system has precision of 0.7, recall of 0.54 and F_1 score of 0.61; for the technique of Schematic Structure the system delivers precision of 0.82, recall of 0.69, and F_1 score of 0.75; and finally for the combination of these techniques the system improves the performance and delivers precision of 0.91, recall of 0.77, and F_1 score of 0.83.

III. Conclusion

This paper outlined the raising trends of the very recent argument mining research field. First of all, it is important to distinguish between the well-known NLP research field of *opinion mining* (or *sentiment analysis*) and argument mining. Besides minor differences, the main point here is that the goal of opinion mining is to understand *what* people think about something while the goal of argument mining is to understand *why* people think something about a certain topic [14]. Second, argument mining approaches can support formal argumentation approaches to define formal models closer to human reasoning, where the fuzziness and ambiguity of natural language plays an important role and where the intellectual process is not always completely rational and objective. Actually, argument mining can provide more insights to answer questions like "what are the best arguments to influence a real audience?" and "what is the role of emotions in the argumentation process?".

As discussed also in the surveys of argument mining [31], [26], argument mining approaches face two main issues nowadays: big data and deep learning. Concerning the former, a huge amount of data is now available on the Web, such as social network posts, forums, blogs, product reviews, user comments to newspapers articles, and needs to be automatically analysed as it goes far beyond human capabilities to parse and understand it without any automatic support tool. Argument mining can make the difference here, and can exploit the Web to perform crowd-sourcing assessments to annotate very large corpora despite the difficulty of the task. Concerning the latter, deep learning methods, i.e., fast and efficient machine learning algorithms such as word embeddings, can be exploited in the argument mining pipeline to deal with large corpora and unsupervised learning.

Acknowledgment

Some research reported in this report was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

References

[1] 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.

[2] Kevin D. Ashley and Vern R. Walker. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Enrico Francesconi and Bart Verheij, editors, *International Conference on Artificial Intelligence and Law, ICAIL '13, Rome, Italy, June 10-14, 2013*, pages 176–180. ACM, 2013.

[3] Filip Boltuzic and Jan Snajder. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the 3rd Workshop on Argument Mining*, 2016.

[4] Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 917–924, 2014.

[5] Katarzyna Budzynska and Serena Villata. *Handbook of Formal Argumentation*, chapter Processing Argumentation in Natural Language Texts. 2017, to appear.

[6] E. Cabrio and S. Villata. Natural language arguments: A combined approach. In *Procs of ECAI, Frontiers in Artificial Intelligence and Applications 242*, pages 205–210, 2012.

[7] C. Chesnevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316, 2006.

[8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:3746, 1960.

[9] R. Duthie, J. Lawrence, K Budzynska, and C. Reed. The CASS technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin, 2016. Association for Computational Linguistics.

[10] Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *Proceedings of 6th International Conference on Computational Models of Argument (COMMA 2016)*. IOS Press, Frontiers in Artificial Intelligence and Applications, 2016.

[11] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2011)*, pages 987–996, 2011.

[12] James B Freeman. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter, 1991.

[13] Debanjan Ghosh, Aquila Khanam, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of ACL 2016*, 2016.

[14] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, 2014.

[15] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, page (in press), 2016. Submission received: 2 April 2015; revised version received: 20 April 2016; accepted for publication: 14 June 2016. Pre-print available at http://arxiv.org/abs/1601.02403.

[16] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, 2016. Association for Computational Linguistics.

[17] Francisca Snoeck Henkemans, Frans van Eemeren, and Peter Houtlosser. *Argumentative Indicators in Discourse. A Pragma-Dialectical Study*. Dordrecht: Springer, 2007.

[18] Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. Argumentation, ideology, and issue framing in parliamentary discourse. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, 2014.

[19] Mathilde Janier, Mark Aakhus, Katarzyna Budzynska, and Chris Reed. Modeling argumentative activity in mediation with Inference Anchoring Theory: The case of impasse. In *European Conference on Argumentation (ECA)*, 2015.

[20] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, June 2015. Association for Computational Linguistics.

[21] Alister Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

[22] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159174, 1977.

[23] J. Lawrence and C.A. Reed. Argument mining using argumentation scheme structures. In P. Baroni, M. Stede, and T. Gordon, editors, *Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016)*, Berlin, 2016. IOS Press.

[24] John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. AIFdb: Infrastructure for the argument web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, 2012.

[25] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191. AAAI Press, 2015.

[26] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[27] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009),*, pages 98–109. ACM, 2009.

[28] Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, 2013.

[29] Joonsuk Park and Claire Cardie. Assess: A tool for assessing the support structures of arguments in user comments. In *Proc. of the Fifth International Conference on Computational Models of Argument*, pages 473–474, 2014.

[30] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[31] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

[32] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *First European Conference on Argumentation: Argumentation and Reasoned Action*, 2015.

[33] Patrick Saint-Dizier. *Challenges of Discourse processing: the case of technical documents*. Cambridge Scholars Publishing, 2014.

[34] J. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, New York, 1969.

[35] Dhanya Sridhar, James R. Foulds, Bert Huang, Lise Getoor, and Marilyn A. Walker. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 116–125. The Association for Computer Linguistics, 2015.

[36] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510, 2014.

[37] Manfred Stede. The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, 2004.

[38] Simone Teufel, Jean Carletta, and Marie-Francine Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics, 1999.

[39] Maria Paz G. Villalba and Patrick Saint-Dizier. Some facets of argument mining for opinion analysis. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 23–34, 2012.

[40] Henning Wachsmuth, Johannes Kiesel, and Benno Stein. Sentiment flow - A general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 601–611, 2015.

[41] Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

Entity Coreference Resolution

Vincent Ng

***Abstract*—Entity coreference resolution is generally considered one of the most difficult tasks in natural language understanding. Though extensively investigated for more than 50 years, the task is far from being solved. Its difficulty stems from its reliance on sophisticated knowledge sources and inference mechanisms. Nevertheless, significant progress has been made on learning-based coreference research since its inception two decades ago. This paper provides an overview of the major milestones made in learning-based coreference research.**

***Index Terms*—text mining, natural language processing, information extraction, coreference resolution, anaphora resolution**

I. Introduction

ENTITY coreference resolution is generally considered one of the most difficult tasks in natural language processing (NLP). The task involves determining which entity mentions in a text or dialogue refer to the same real-world entity. Despite being investigated for 50 years in the NLP community, the task is still far from being solved. To better understand its difficulty, consider the following sentence:

> The Queen Mother asked Queen Elizabeth II to transform her sister, Princess Margaret, into a viable princess by summoning a renowned speech therapist, Nancy Logue, to treat her speech impediment.

A coreference system should partition the entity mentions in this sentence into three coreference chains — QE (*Queen Elizabeth II* and the first occurrence of *her*), PM (*sister*, *Princess Margaret* and the second occurrence of *her*), and NL (*a renowned speech therapist* and *Nancy Logue*) — and three singletons, *The Queen Mother*, *a viable princess*, and *speech impediment*.

While human audiences have few problems with identifying these co-referring mentions, the same is not true for automatic coreference resolvers. For instance, resolving the two occurrences of *her* in this example is challenging for a coreference resolver. To resolve the first occurrence of *her*, a resolver would determine whether it is coreferent with *The Queen Mother* or *Queen Elizabeth II*, but the portion of the sentence preceding the pronoun does not contain sufficient information for correctly resolving it. The only way to correctly resolve the pronoun is to employ the background knowledge that *Princess Margaret* is *Queen Elizabeth II*'s sister. To resolve the second occurrence of *her*, if a resolver employs the commonly-used heuristic that selects the closest grammatically compatible mention in the subject position as its antecedent, it will wrongly posit *Nancy Logue* as its antecedent. Even if the sentence did not mention that *Nancy Logue* was a speech therapist, a human would have no problem with correctly resolving the pronoun (to *Princess Margaret*), because he could easily rule out *Nancy Logue* as the correct antecedent by employing the commonsense knowledge that it does not make sense for Person A to summon Person B to treat Person B's problem.

From this example, we can see that background knowledge, which is typically difficult for a machine to acquire, plays an important role in coreference resolution. In general, however, the difficulty of coreference resolution, particularly the resolution of pronouns and common noun phrases, stems from its reliance on sophisticated knowledge sources and inference mechanisms [1]. Despite its difficulty, coreference resolution is a core task in information extraction: it is the fundamental technology for consolidating the textual information about an entity, which is crucial for essentially all NLP applications, such as question answering, information extraction, text summarization, and machine translation. For instance, given the question *When was Mozart born?*, a question-answering system should search for the answer in a set of documents retrieved by a search engine that contain the keywords in the question. If the answer appears in the sentence *He was born in Salzburg, Austria, in 27 January 1756*, the system can be sure that *27 January 1756* is the correct answer only if the pronoun *He* is coreferent with *Mozart*.

As coreference resolution is inherently a clustering task, it has received a lot of attention in the machine learning and data mining communities, where the task has been tackled under different names, such as *record linkage/matching* and *duplicate detection*. Some researchers have focused on *name matching*, where the goal is to determine whether the names appearing in two records in a database refer to the same entity. The focus on name matching effectively ignores pronoun resolution and common noun phrase resolution, which are arguably the most difficult subtasks of entity coreference resolution [2].

There is a recent surge of interest in pronoun resolution in the knowledge representation community owing to the Winograd Schema Challenge (WSC). The WSC was motivated by the following pair of sentences, which was originally used by Winograd [3] to illustrate the difficulty of NLP:

(1) The city council refused the women a permit because *they* feared violence.

(2) The city council refused the women a permit because *they* advocated violence.

Using world knowledge, humans can easily resolve the occurrences of *they* in sentences (1) and (2) to *The city council* and *the women* respectively. However, these pronouns are difficult to resolve automatically. One reason for this is that these pronouns are compatible with both candidate antecedents in number, gender, and semantic class. Another reason is that correct resolution may not be possible without understanding the two events mentioned in a sentence, but such understanding

Human Language Technology Research Institute, University of Texas at Dallas, Richardson, TX, USA; e-mail vince@hlt.utdallas.edu.

typically requires background knowledge. Levesque [4] argued that the resolution of difficult-to-resolve pronouns in *twin* sentences like these constitutes a task that can serve as an appealing alternative to the Turing Test. The WSC is currently being promoted by Commonsense Reasoning[1], so we expect to see continued progress on this task.

Our goal in this paper is to provide the reader with an overview of the major milestones made in learning-based entity coreference research since its inception 20 years ago. For a detailed treatment of this topic, we refer the reader to a recent book edited by Poesio et al. [5]. Given Levesque's aforementioned proposal that the resolution of difficult-to-resolve pronouns can serve as an appealing alternative to the Turing Test, we believe that the entity coreference task will be of interest to the general intelligence systems community.

II. Brief History

Learning-based entity coreference research was to a large extent stimulated by the public availability of coreference-annotated corpora that were produced as a result of three large-scale evaluations of coreference systems:

The MUC evaluations. The coreference evaluations conducted as part of the DARPA-sponsored MUC-6 [6] and MUC-7 [7] conferences provided the first two publicly available coreference corpora, the MUC-6 corpus (30 training texts and 30 test texts) and the MUC-7 corpus (30 training texts and 20 test texts). They also defined the coreference task that the NLP community sees today. In particular, the MUC organizers decided that the task should focus exclusively on *identity* coreference resolution, ignoring other kinds of coreference relations that would be challenging even for humans to identify, such as bridging (e.g., set-subset relations, part-whole relations). A significant byproduct of the MUC coreference evaluation was the first evaluation metric for coreference resolution, the MUC scoring metric [8]. Virtually all learning-based resolvers developed between 1995 and 2004 were trained and evaluated on the MUC corpora using the MUC metric.

The ACE evaluations. As part of NIST-sponsored ACE evaluations, which began in the late 1990s, four coreference corpora were released, namely ACE-2, ACE03, ACE04, and ACE05. To encourage multilingual coreference research, ACE04 and ACE05 were composed of coreference-annotated texts not only for English, but also for Chinese and Arabic. These two corpora were also heavily used for training and evaluation in part because they were much larger than the MUC corpora. For instance, the ACE04 and ACE05 English coreference training corpora were composed of 443 and 599 documents, respectively. Unlike MUC, which requires the identification of coreferent entities regardless of their semantic types, ACE focused on a restricted, simpler version of the coreference task, requiring that coreference chains be identified only for entities belonging to one of the ACE entity types (e.g., PERSON, ORGANIZATION, GPE, FACILITY, LOCATION). Virtually all resolvers developed between 2004 and 2010 were trained and evaluated on one of these ACE corpora.

[1]http://commonsensereasoning.org/winograd.html

To evaluate coreference systems in the official ACE evaluations, the ACE metric was developed, but it was never popularly used by coreference researchers. Two important scoring measures were developed during this period, namely B^3 [9] and CEAF [10].

Direct comparisons among the different coreference systems developed at that time were difficult for at least two reasons. First, different resolvers were evaluated on different corpora (ACE04 vs. ACE05) using different evaluation metrics (B^3 vs. CEAF). Second, and more importantly, they were trained and evaluated on different train-test splits of the ACE corpora, owing to the fact that the ACE organizers released only the training portion but not the official test portion of the ACE corpora. Worse still, some resolvers were evaluated on *gold* rather than *system* (i.e., automatically extracted) entity mentions [11], reporting substantially better results than end-to-end resolvers. This should not be surprising: coreference on gold mentions is a substantially simplified version of the coreference task because system mentions typically significantly outnumber gold mentions. Some of these complications were referred to as "conundrums" in entity coreference resolution and discussed in detail by Stoyanov et al. [12].

The CoNLL 2011 and 2012 shared tasks. The CoNLL 2011 [13] and 2012 [14] shared tasks focused on English and multilingual (English, Chinese, and Arabic) coreference resolution, respectively, using the OntoNotes 5.0 corpus [15] for training and evaluation. These shared tasks were important for two reasons. First, they directed researchers' attention back to the challenging *unrestricted* coreference tasks that were originally defined in MUC while providing substantially more data for training and evaluation. Second, and more importantly, they facilitated performance comparisons of different resolvers, making it possible to determine the state of the art. Specifically, they standardized not only the train-test partition of the OntoNotes corpus, but also the evaluation metric, the CoNLL metric [13], which is the unweighted average of MUC, B^3, and CEAF. Virtually all resolvers developed since 2011 were evaluated on this corpus.

III. Evaluation Measures

Designing evaluation measures for coreference resolution is by no means a trivial task. In this section, we describe the four most commonly-used coreference evaluation measures, each of which reports performance in terms of recall, precision, and F-score. Below we use the terms *coreference chains* and *coreference clusters* interchangeably. For a coreference chain C, we define $|C|$ as the number of mentions in C. *Key chains* and *system chains* refer to gold coreference chains and system-generated coreference chains, respectively. In addition, $\mathcal{K}(d)$ and $\mathcal{S}(d)$ refer to the set of gold chains and the set of system-generated chains in document d, respectively. Specifically,

$$\mathcal{K}(d) = \{K_i : i = 1, 2, \cdots, |\mathcal{K}(d)|\},$$

$$\mathcal{S}(d) = \{S_j : j = 1, 2, \cdots, |\mathcal{S}(d)|\},$$

where K_i is a chain in $\mathcal{K}(d)$ and S_j is a chain in $\mathcal{S}(d)$. $|\mathcal{K}(d)|$ and $|\mathcal{S}(d)|$ are the number of chains in $\mathcal{K}(d)$ and $\mathcal{S}(d)$, respectively.

A. MUC

MUC [8] is a link-based metric. Given a document d, recall is computed as the number of common links between the key chains and the system chains in d divided by the number of links in the key chains. Precision is computed as the number of common links divided by the number of links in the system chains. Below we show how to compute (1) the number of common links, (2) the number of key links, and (3) the number of system links.

To compute the number of common links, a partition $P(S_i)$ is created for each system chain S_i using the key chains. Specifically,

$$P(S_j) = \{C_j^i : i = 1, 2, \cdots, |\mathcal{K}(d)|\} \tag{1}$$

Each subset C_j^i in $P(S_i)$ is formed by intersecting S_j with K_i. Note that $|C_j^i| = 0$ if S_j and K_i have no mentions in common. Since there are $|\mathcal{K}(d)| * |\mathcal{S}(d)|$ subsets in total, the number of common links is

$$\begin{aligned} c(\mathcal{K}(d), \mathcal{S}(d)) &= \sum_{j=1}^{|\mathcal{S}(d)|} \sum_{i=1}^{|\mathcal{K}(d)|} w_c(C_j^i), \\ \text{where } w_c(C_j^i) &= \begin{cases} 0 & \text{if } |C_j^i| = 0; \\ |C_j^i| - 1 & \text{if } |C_j^i| > 0. \end{cases} \end{aligned} \tag{2}$$

Intuitively, $w_c(C_j^i)$ can be interpreted as the "weight" of C_j^i. In MUC, the weight of a cluster is defined as the *minimum* number of *links* needed to create the cluster, so $w_c(C_j^i) = |C_j^i| - 1$ if $|C_j^i| > 0$.

The number of links in the key chains, $\mathcal{K}(d)$, is calculated as:

$$k(\mathcal{K}(d)) = \sum_{i=1}^{|\mathcal{K}(d)|} w_k(K_i), \tag{3}$$

where $w_k(K_i) = |K_i| - 1$. The number of links in the system chains, $s(\mathcal{S}(d))$, is calculated as:

$$s(\mathcal{S}(d)) = \sum_{j=1}^{|\mathcal{S}(d)|} w_s(S_j), \tag{4}$$

where $w_s(S_j) = |S_j| - 1$.

B. B^3

MUC's often-criticized weakness is that it fails to reward successful identification of singleton clusters. To address this weakness, B^3 [9] first computes the recall and precision for each mention, and then averages these per-mention values to obtain the overall recall and precision.

Let m_n be the nth mention in document d. Its recall, $R(m_n)$, and precision, $P(m_n)$, are computed as follows. Let K_i and S_j be the key chain and the system chain that contain m_n, respectively, and let C_j^i be the set of mentions appearing in both S_j and K_i.

$$R(m_n) = \frac{w_c(C_j^i)}{w_k(K_i)}, P(m_n) = \frac{w_c(C_j^i)}{w_s(S_j)}, \tag{5}$$

where $w_c(C_j^i) = |C_j^i|$, $w_k(K_i) = |K_i|$, and $w_s(S_j) = |S_j|$.

C. CEAF

While B^3 addresses the shortcoming of MUC, Luo [10] presents counter-intuitive results produced by B^3, which it attributes to the fact that B^3 may use a key/system chain more than once when computing recall and precision. To ensure that each key/system chain will be used at most once in the scoring process, his CEAF scoring metric scores a coreference partition by finding an optimal *one-to-one mapping* (or *alignment*) between the chains in $\mathcal{K}(d)$ and those in $\mathcal{S}(d)$.

Since the mapping is one-to-one, not all key chains and system chains will be involved in it. Let $\mathcal{K}_{min}(d)$ and $\mathcal{S}_{min}(d)$ be the set of key chains and the set of system chains involved in the alignment, respectively. The alignment can be represented as a one-to-one mapping function g, where

$$g(K_i) = S_j, K_i \in \mathcal{K}_{min}(d) \text{ and } S_j \in \mathcal{S}_{min}(d).$$

The score of g, $\Phi(g)$, is defined as

$$\Phi(g) = \sum_{K_i \in \mathcal{K}_{min}(D)} \phi(K_i, g(K_i)),$$

where ϕ is a function that computes the *similarity* between a gold chain and a system chain. The optimal alignment, g^*, is the alignment whose Φ value is the largest among all possible alignments, and can be computed efficiently using the Kuhn-Munkres algorithm [16].

Given g^*, the recall (R) and precision (P) of a system partition can be computed as follows:

$$R = \frac{\Phi(g^*)}{\sum_{i=1}^{|\mathcal{K}(d))|} \phi(K_i, K_i)}, P = \frac{\Phi(g^*)}{\sum_{j=1}^{|\mathcal{S}(d))|} \phi(S_j, S_j)}.$$

As we can see, at the core of CEAF is the similarity function ϕ. Luo defines two different ϕ functions, ϕ_3 and ϕ_4:

$$\phi_3(K_i, S_j) = |K_i \cap S_j| = w_c(C_j^i) \tag{6}$$

$$\phi_4(K_i, S_j) = \frac{2|K_i \cap S_j|}{|K_i| + |S_j|} = \frac{2 * w_c(C_j^i)}{w_k(K_i) + w_s(S_j)} \tag{7}$$

ϕ_3 and ϕ_4 result in mention-based CEAF (a.k.a. CEAF$_m$) and entity-based CEAF (a.k.a. CEAF$_e$), respectively.

D. BLANC

BLANC [17], a Rand-index-based coreference evaluation measure, is designed to address a major weakness shared by B^3 and CEAF: the B^3 and CEAF F-scores typically squeeze up too high when many singleton mentions are present in a document. To address this weakness, BLANC first computes recall, precision, and F-score separately for coreferent mention pairs and non-coreferent mention pairs. The BLANC recall/precision/F-score is then computed as the unweighted average of the recall/precision/F-score of the coreferent mention pairs and the recall/precision/F-score of the non-coreferent mention pairs.

IV. Models

In this section, we examine the major learning-based models for entity coreference resolution.

A. Mention-Pair Models

Despite their conceptual simplicity, mention-pair models are arguably the most influential coreference model. A mention-pair model is a binary classifier that determines whether a pair of mentions is co-referring or not. Hence, to train a mention-pair model, each training instance corresponds to a pair of mentions and is represented by *local* features encoding each of the two mentions and their relationships. Any learning algorithm can be used to train a mention-pair model, which can then be applied to classify the test instances. However, these pairwise classification decisions could violate transitivity, which is an inherent property of the coreference relation. As a result, a separate clustering mechanism, such as *single-link clustering* [18] and *best-first* clustering [2], is needed to coordinate the pairwise decisions and construct a partition. Specifically, these clustering algorithms process the mentions in a test text in a left-to-right manner. For each mention encountered, they select as its antecedent either the closest or the most probable preceding coreferent mention. No antecedent will be selected for the mention if it does not have any preceding coreferent mention.

It was around this time that Ng and Cardie [19] raised the question of whether *anaphoricity* should be modeled explicitly in coreference resolution. Anaphoricity determination is the task of determining whether a mention is *anaphoric* (i.e., it is coreferent with a preceding mention) or *non-anaphoric* (i.e., it starts a new coreference chain).

To motivate anaphoricity determination, consider the two aforementioned clustering algorithms, which do not perform anaphoricity determination explicitly. Specifically, s mention is implicitly posited as non-anaphoric if none of its preceding mentions is classified as coreferent with it. Ng and Cardie [19] hypothesize that performing anaphoricity determination prior to coreference resolution could improve the precision of a mention-pair model, as the model will only need to resolve mentions that are determined to be anaphoric by the anaphoricity model. While anaphoricity determination is by no means an easier task than coreference resolution, many years of research on the explicit modeling of anaphoricity have resulted in models that can benefit coreference. One such successful attempt was made by Denis and Baldridge [20], who perform joint inference over the outputs of two independently-trained models, the anaphoricity model and the mention-pair model.

B. Mention-Ranking Models

A major weakness of mention-pair models is that they consider each candidate antecedent of an anaphoric mention to be resolved independently of other candidate antecedents. As a result, they can only determine how good a candidate antecedent is relative to the anaphoric mention, but not how good it is relative to other candidate antecedents.

Ranking models address this weakness by allowing all candidate antecedents of a mention to be ranked *simultaneously* [21]–[23]. Since a mention ranker simply ranks candidate antecedents, it cannot determine if a mention is anaphoric. One way to address this problem is to apply an independently trained anaphoricity classifier to identify non-anaphoric mentions prior to ranking [23]. Another, arguably better, way is to jointly learn coreference and anaphoricity by augmenting the candidate set of each mention to be resolved with a dummy candidate antecedent so that the mention will be classified as non-anaphoric if it is resolved to the dummy [24].

C. Entity-Based Models

Another major weakness of mention-pair models concerns their limited *expressiveness*: they can only employ features defined on no more than two mentions. However, the information extracted from the two mentions alone may not be sufficient for making an informed coreference decision, especially if the candidate antecedent is a pronoun (which is semantically empty) or a mention that lacks descriptive information such as gender (e.g., *Clinton*).

Entity-based models aim to address the expressiveness problem. To motivate these models, consider a document that consists of three mentions: *Mr. Clinton*, *Clinton*, and *she*. A mention-pair model may determine that *Mr. Clinton* and *Clinton* are coreferent using string-matching features, and that *Clinton* and *she* are coreferent based on proximity and lack of evidence for gender and number disagreement. However, these two pairwise decisions together with transitivity imply that *Mr. Clinton* and *she* will end up in the same cluster, which is incorrect due to gender mismatch. This kind of error arises in part because the later coreference decisions are not dependent on the earlier ones. In particular, had the model taken into consideration that *Mr. Clinton* and *Clinton* were in the same cluster, it probably would not have posited that *she* and *Clinton* are coreferent. Specifically, the increased expressiveness of entity-based models stems from their ability to exploit *cluster-level* (a.k.a. *non-local*) features, which are features defined on an arbitrary subset of the mentions in a coreference cluster. In our example, it would be useful to have a cluster-level feature that encodes whether the gender of a mention is compatible with the gender of *each* of the mentions in a preceding cluster, for instance.

Many machine-learned entity-based models have been developed over the years. The most notable ones include the entity-based versions of mention-pair models and mention-ranking models. *Entity-mention* models, the entity-based version of mention-pair models, determine whether a mention is coreferent with a preceding, possibly partially-formed, *cluster* [25], [26]. Despite their improved expressiveness, early entity-mention models have not yielded particularly encouraging results. *Cluster-ranking* models, on the other hand, are the entity-based version of mention-ranking models [24]. They rank preceding clusters rather than candidate antecedents, and have been shown to outperform entity-mention models, mention-pair models, and mention-ranking models.

While the entity-based models discussed so far have all attempted to process the mentions in a test text in a left-to-right manner, *easy-first* models aim to make easy linking decisions first, and then use the information extracted from the clusters established thus far to help identify the difficult links. More specifically, an easy-first resolver is composed of

a pipeline of *sieves*, each of which is composed of a set of hand-crafted or learned rules for classifying a *subset* of the mention pairs in the test set. Being an easy-first approach, the sieves in the pipeline are arranged in decreasing order of precision. Given the pipeline setup, the later sieves can exploit the decisions made by the earlier sieves. The most well-known resolver that employs an easy-first approach is arguable Stanford's resolver [27], which won the CoNLL-2011 shared task. Ratinov and Roth's easy-first resolver [28] improves Stanford's resolver by allowing earlier decisions to be overridden and corrected by later sieves.

Entity-based models are also trained by Culotta et al. [29] and Stoyanov and Eisner [30]. Specifically, they propose a "learning to cluster" approach to train coreference models to perform *agglomerative* clustering of the entity mentions, each of which is initially is in its own cluster.

D. Structured Models

Recent years have seen a popular line of work that views coreference resolution as a *structured prediction* task: rather than resolving a mention to a preceding mention/cluster, a structured model predicts a structure from which a coreference partition can be directly recovered.

The first such attempts are made by McCallum and Wellner [11] and Finley and Joachims [31], who train models to directly induce coreference partitions. Specifically, McCallum and Wellner train a log-linear model to induce a distribution over the possible partitions of a set of mentions so that the correct partition is the most probable. Finley and Joachims, on the other hand, learn to rank candidate coreference partitions by training a max-margin ranking model.

While learning to partition is a novel idea, partition-based models are not particularly popular. One reason is that they force us to classify each pair of mentions, which is not desirable as not all coreference links are equally easy to identify. Fortunately, to establish a cluster of n mentions, only $n-1$ coreference links are needed. So, rather than learning a partition, Fernandes et al. [32] (FDM) propose learning a coreference *tree* using the links that are easy to identify, and then recovering a partition from the tree. To learn to predict coreference trees, FDM employ the latent structured voted perceptron algorithm. The model parameters are weights defined on features that are commonly-used in mention-pair models. In each iteration, the highest-scoring (i.e., maximum spanning) tree is decoded using the Chu-Liu-Edmonds algorithm [33], [34]. Their resolver achieved the highest average score over all languages in the CoNLL-2012 shared task. As noted by FDM, feature induction plays an important role in their resolver. Their feature induction method learns feature conjunctions, which are derived from the paths of a decision tree-based mention-pair model.

Seeing no reason to predict structures as complicated as trees, Durrett and Klein [35] (D&K) simplify the coreference task by proposing a model that predicts for each test document the most probable *antecedent structure*, which is a vector of antecedents storing the antecedent chosen for each mention (null if the mention is non-anaphoric) in the document. Effectively, it is a mention-ranking model, but it is trained to maximize the conditional likelihood of the correct antecedent structure given a document. Inference is easy: the most probable candidate antecedent of a mention is selected to be its antecedent independently of other mentions. One of the innovations of D&K's model is the use of a task-specific loss function. Specifically, D&K employ a loss function that is a weighted sum of the counts of three error types: the number of false anaphors, the number of false non-anaphors, and the number of wrong links. Following FDM, D&K employ feature conjunctions. Perhaps most interestingly, D&K achieved state-of-the-art performance by training their model only on conjunctions of *lexical* features.

Motivated in part by the recent successes of neural models for NLP tasks, Wiseman et al. [36] train a *neural-based* mention-ranking model which, like D&K's model, employs a task-specific loss function. However, rather than following the recent trend on training *linear* models using feature conjunctions [32], [35], [37], some of which are rather complex, Wiseman et al. pioneered using a neural network to learn *non-linear* representations of *raw* features (i.e., the original features, without any conjunctions), achieving state-of-the-art results. Most recently, Wiseman et al. [38] and Clark and Manning [39] further improved the performance of neural coreference models by incorporating cluster-based features. These are the first attempts to learn non-linear models of coreference resolution. Given their promising results, they deserve further investigations.

V. Knowledge Sources

Early learning-based coreference resolvers have relied primarily on morpho-syntactic knowledge. However, the development of large lexical knowledge bases since the late 1990s and the significant advancements made in corpus-based lexical semantics research in the past 15 years have enabled researchers to design semantic features for coreference resolution. In this section, we examine these two types of knowledge sources.

A. Morpho-syntactic Features

Morpho-syntactic features typically refer to several types of features. **String-matching features** encode whether there is an exact or partial match (e.g., head match, exact match after removing determiners) between the strings of the two mentions under consideration. These features are useful because many coreferent mentions have overlaps in their strings (e.g., *Bill Clinton* and *Clinton*). **Lexical features** are created by concatenating the strings/heads of the two mentions. These features enable a learning algorithm to learn which string/head combinations are indicative of coreference relations. **Grammatical features** encode whether the two mentions are compatible with respect to various grammatical attributes such as gender and number. These features are useful because grammatical incompatibility is a strong indicator of non-coreference. Finally, **syntactic features** encode whether two mentions can be coreferent based on information extracted from syntactic parse trees. For instance, two mentions cannot be coreferent if they violate the Binding Constraints.

B. Semantic Features

Selectional preference is one of the earliest kinds of semantic knowledge exploited for coreference resolution [40]–[42]. Given a pronoun to be resolved, its governing verb, and its grammatical role, a candidate antecedent that can play the same role and be governed by the same verb is preferred. These preferences can be learned from a large corpus or from the Web, and have been used as features to improve knowledge-poor resolvers with varying degrees of success.

Another commonly-used semantic feature for coreference resolution encodes whether the two mentions involved have the same **semantic class**, where the semantic class of a common noun is determined using either WordNet [18], [43] or clusters induced from the Google n-gram corpus [44].

Knowing that *Barack Obama* is a *U. S. president* would be helpful for establishing the coreference relation between two mentions *Obama* and *the president* in a document. To this end, researchers have attempted to extract the **knowledge attributes** of a proper name from lexical knowledge bases. For instance, given a proper name, Ratinov and Roth [28] extract from Wikipedia its Wiki category, gender, and nationality, and Hajishirzi et al. [45] extract from Freebase a set of coarse-grained attributes (e.g., *person*, *location*) and more than 500 fine-grained attributes (e.g., *plant*, *attraction*, *nominee*). The major challenge in extracting attributes from these knowledge bases is entity disambiguation [46]: a proper name could be matched more than one Wikipedia page or more than one entry in YAGO and Freebase. To address this problem, Ratinov and Roth [28] employ a context-sensitive entity disambiguation system, while Hajishirzi et al. [45] propose to jointly perform coreference resolution and entity linking. Knowledge attributes can also be extracted in an unsupervised manner using hand-crafted lexico-syntactic patterns [47]. For instance, we can search for the pattern *X is a Y* in a large, unannotated corpus. The mention pairs (X,Y) that satisfy this pattern can tell us that mention X has knowledge attribute Y.

Besides the IS-A relation, other **semantic relations**, including those between common nouns, have also been used for coreference resolution. For instance, Bengtson and Roth [48] have employed as features the generic semantic relations (e.g., synonymy, hypernymy, antonymy) extracted from WordNet for two common nouns. Hearst [47] has proposed other lexico-syntactic patterns that capture different lexical semantic relations between nouns. Yang and Su [49] employ patterns *learned* from a coreference corpus that are indicative of a coreference relation.

Some words may not have a semantic relation but can still be coreferent owing to their **semantic similarity**. This observation has led Ponzetto and Strube [43] to encode features based on various measures of WordNet similarity, which have been shown to improve their baseline system.

PropBank-style *semantic roles* have also been used for coreference resolution [43]. Their use is motivated by the **semantic parallelism** heuristic: given an anaphor with semantic role r, its antecedent is likely to have role r.

While using semantic roles improves Ponzetto and Strube's resolver [43], semantic parallelism is a fairly weak indicator of coreference. For instance, if two verbs denote events that are unrelated to each other, it is not clear why their arguments should be coreferent even if they have the same semantic role. Motivated by this observation, Rahman and Ng [46] attempt to capture the notion of **event relatedness** based on whether the two predicates appear in the same *FrameNet semantic frame*, designing features that encode not only whether the two mentions have the same role but also whether their governing verbs are in the same frame.

Generally speaking, the results of employing semantic and world knowledge to improve knowledge-poor coreference resolvers are mixed. The mixed results can be attributed at least in part to differences in the strengths of the baseline resolvers employed in the evaluation: the stronger the baseline is, the harder it would be to improve its performance. Since different researchers employed different baselines and evaluated their resolvers on different feature sets, it is not easy to draw general conclusions on the usefulness of different kinds of semantic features. To facilitate comparison of the usefulness of different kinds of semantic features, we believe that it is worthwhile to re-evaluate them using the standard evaluation setup provided by the CoNLL-2011 and 2012 shared tasks.

VI. Conclusion

We presented an overview of the models and features developed for learning-based entity coreference resolution in the past two decades, as well as the corpora and metrics used in the evaluation of these computational models. Despite the continued progress on this task, it is far from being solved: the best CoNLL scores reported to date on the CoNLL-2012 official evaluation data for English and Chinese are 65.29 and 63.66 respectively [39]. Recent results suggest that the performance of coreference models that do not employ sophisticated knowledge is plateauing [38]. Hence, one of the fruitful avenues of future research will likely come from the incorporation of sophisticated knowledge sources.

Acknowledgment

This work was supported in part by NSF Grant IIS-1219142. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies of NSF.

References

[1] R. Mitkov, B. Boguraev, and S. Lappin, "Introduction to the special issue on computational anaphora resolution," *Computational Linguistics*, vol. 27, no. 4, pp. 473–477, 2001.

[2] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2012, pp. 104–111.

[3] T. Winograd, *Understanding Natural Language*. New York: Academic Press, Inc., 1972.

[4] H. Levesque, "The winograd schema challenge," in *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[5] M. Poesio, R. Stuckardt, and Y. V. (Eds.), *Anaphora Resolution: Algorithms, Resources, and Evaluation*. Springer Verlag, 2016.

[6] MUC-6, *Proceedings of the Sixth Message Understanding Conference*. San Francisco, CA: Morgan Kaufmann, 1995.

[7] MUC-7, *Proceedings of the Seventh Message Understanding Conference*. San Francisco, CA: Morgan Kaufmann, 1998.

[8] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme," in *Proceedings of the Sixth Message Understanding Conference*, 1995, pp. 45–52.

[9] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, 1998, pp. 563–566.

[10] X. Luo, "On coreference resolution performance metrics," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 25–32.

[11] A. McCallum and B. Wellner, "Conditional models of identity uncertainty with application to noun coreference," in *Advances in Neural Information Proceesing Systems*, 2004.

[12] V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff, "Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 656–664.

[13] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue, "CoNLL-2011 Shared Task: Modeling unrestricted coreference in Ontonotes," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 2011, pp. 1–27.

[14] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, 2012, pp. 1–40.

[15] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "OntoNotes: The 90% solution," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 2006, pp. 57–60.

[16] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[17] M. Recasens and E. Hovy, "BLANC: Implementing the Rand Index for coreference evaluation," *Natural Language Engineering*, vol. 17, no. 4, pp. 485–510, 2011.

[18] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[19] V. Ng and C. Cardie, "Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution," in *Proceedings of the 19th International Conference on Computational Linguistics*, 2002, pp. 730–736.

[20] P. Denis and J. Baldridge, "Global, joint determination of anaphoricity and coreference resolution using integer programming," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, pp. 236–243.

[21] R. Iida, K. Inui, H. Takamura, and Y. Matsumoto, "Incorporating contextual cues in trainable models for coreference resolution," in *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, 2003.

[22] X. Yang, G. Zhou, J. Su, and C. L. Tan, "Coreference resolution using competitive learning approach," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 176–183.

[23] P. Denis and J. Baldridge, "Specialized models and ranking for coreference resolution," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 660–669.

[24] A. Rahman and V. Ng, "Supervised models for coreference resolution," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 968–977.

[25] X. Luo, A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos, "A mention-synchronous coreference resolution algorithm based on the Bell tree," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 135–142.

[26] X. Yang, J. Su, G. Zhou, and C. L. Tan, "An NP-cluster based approach to coreference resolution," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[27] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules," *Computational Linguistics*, vol. 39, no. 4, pp. 885–916, 2013.

[28] L. Ratinov and D. Roth, "Learning-based multi-sieve co-reference resolution with knowledge," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1234–1244.

[29] A. Culotta, M. Wick, and A. McCallum, "First-order probabilistic models for coreference resolution," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, pp. 81–88.

[30] V. Stoyanov and J. Eisner, "Easy-first coreference resolution," in *Proceedings of the 24th International Conference on Computational Linguistics*, 2012, pp. 2519–2534.

[31] T. Finley and T. Joachims, "Supervised clustering with support vector machines," in *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[32] E. Fernandes, C. dos Santos, and R. Milidiú, "Latent structure perceptron with feature induction for unrestricted coreference resolution," in *Joint Conference on EMNLP and CoNLL - Shared Task*, 2012, pp. 41–48.

[33] J. Chu, Y and T. H. Liu, "On the shortest arborescence of a directed graph," *Science Sinica*, vol. 14, no. 1, pp. 1396–1400, 1965.

[34] J. Edmonds, "Optimum branchings," *Journal of Research of the National Bureau of Standards*, vol. 71B, pp. 233–240, 1967.

[35] G. Durrett and D. Klein, "Easy victories and uphill battles in coreference resolution," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1971–1982.

[36] S. Wiseman, A. M. Rush, S. Shieber, and J. Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1416–1426.

[37] A. Björkelund and J. Kuhn, "Learning structured perceptrons for coreference resolution with latent antecedents and non-local features," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 47–57.

[38] S. Wiseman, A. M. Rush, and S. M. Shieber, "Learning global features for coreference resolution," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 994–1004.

[39] K. Clark and C. D. Manning, "Improving coreference resolution by learning entity-level distributed representations," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 643–653.

[40] I. Dagan and A. Itai, "Automatic processing of large corpora for the resolution of anaphora references," in *Proceedings of the 13th International Conference on Computational Linguistics*, 1990, pp. 330–332.

[41] A. Kehler, D. Appelt, L. Taylor, and A. Simma, "The (non)utility of predicate-argument frequencies for pronoun interpretation," in *Human Language Technology Conference and Conference of the North American Chapter of the Association for Computational Linguistics: Main Proceedings*, 2004, pp. 289–296.

[42] X. Yang, J. Su, and C. L. Tan, "Improving pronoun resolution using statistics-based semantic compatibility information," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, pp. 165–172.

[43] S. P. Ponzetto and M. Strube, "Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution," in *Proceedings of the Human Language Technology Conference and Conference of the North American Chapter of the Association for Computational Linguistics*, 2006, pp. 192–199.

[44] M. Bansal and D. Klein, "Coreference semantics from web features," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012, pp. 389–398.

[45] H. Hajishirzi, L. Zilles, D. S. Weld, and L. Zettlemoyer, "Joint coreference resolution and named-entity linking with multi-pass sieves," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 289–299.

[46] A. Rahman and V. Ng, "Coreference resolution with world knowledge," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 814–824.

[47] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 1992.

[48] E. Bengtson and D. Roth, "Understanding the values of features for coreference resolution," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 294–303.

[49] X. Yang and J. Su, "Coreference resolution using semantic relatedness information from automatically discovered patterns," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007, pp. 528–535.

Cognitive Systems: Argument and Cognition

Antonis Kakas , Loizos Michael

***Abstract*—Developing systems that are aware of, and accommodate for, the cognitive capabilities and limitations of human users is emerging as a key characteristic of a new paradigm of cognitive computing in Artificial Intelligence. According to this paradigm, the behavior of such *cognitive systems* is modeled on the behavior of human personal assistants, able to understand the motivations and personal likings / affinities of their interlocutors, while also being able to explain, and ultimately persuade the latter about, their computed solution (e.g., a proposed action) to a problem.**

This paper examines the link between argument and cognition from the psychological and the computational perspectives, and investigates how the synthesis of work on reasoning and narrative text comprehension from Cognitive Psychology and of work on computational argumentation from AI can offer a scientifically sound and pragmatic basis for building human-aware cognitive systems for everyday tasks. The paper aims, thus, to reveal how argumentation can form the *science of common sense thought* on which new forms of cognitive systems can be engineered.

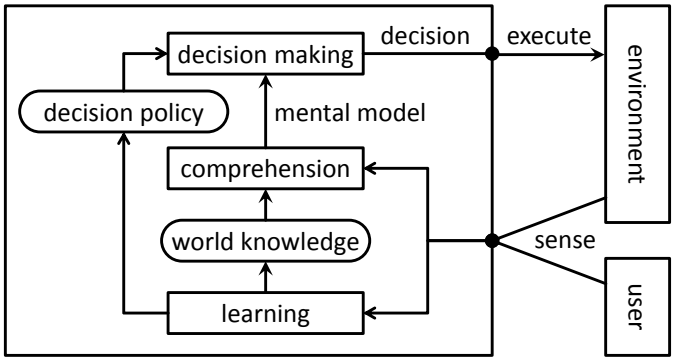


Fig. 1. High-level view of the architecture of a cognitive assistant, with focus on the interaction of the processes of decision making, comprehension, and learning. The components labeled as "decision policy" and "world knowledge" correspond, respectively, to the sets of option arguments and belief arguments.

I. The Emerging Need for Cognitive Systems

THE ever increasing demand for smart devices with ordinary human-level intelligence, capable of common sense reasoning and attuned to everyday problem solving, is forcing Artificial Intelligence to stand up and deliver. Unlike anything seen to date, this new vision of user-device interaction aims to allow ordinary users, without technical background, to instruct or program their devices in a natural and personalized manner, and to allow the devices to assist (and enhance the abilities of) their users in dealing with everyday tasks. This *symbiotic* relation splits the burden of communication among the user and the device, giving rise to a "programming paradigm for the masses" [1] that avoids the extremes of using natural languages that are too complex for ordinary devices, or programming languages that are too complex for ordinary users.

Early examples of systems exhibiting such symbiotic interactions already exist, ranging from personal assistant software provided by major smart-device manufacturers, to the expected application of systems that extract information from massive amounts of unstructured data for the purposes of expert-level analysis of problems in specialized domains (e.g., health, law).

Unlike existing automated systems, these *cognitive systems* [2] often exhibit an operational behavior resembling that of a human personal assistant. In particular, a cognitive system's domain of application is limited to certain common everyday tasks, and its operation revolves around its interaction with its human user in a manner that is compatible with the cognitive reasoning capabilities of the latter. To understand (and correct when needed) the reasoning of the system, the user expects the system to use *common sense* to fill in any important relevant information that the user leaves unspecified, and to be able to keep learning about the domain of application and the user's personal preferences and beliefs through their interaction.

Efforts to meet this emerging need and to guide the future of cognitive systems is bound to benefit from a foundational basis that facilitates a human-device interaction that places cognitive compatibility with humans at the center stage. This paper puts forward computational argumentation as a candidate for this reconciliation between human and machine reasoning, in a manner that is more appropriate than the classical logic basis that underpins the development of automated systems to date.

II. Argumentative Basis of Human Cognition

Given the emphasis of cognitive systems on cognitive compatibility, an argumentative foundation for their development will be a viable option only if human cognition is itself geared towards an argumentative perspective. We overview work from Psychology that provides evidence in support of this condition.

A significant amount of research in the area of Psychology of Reasoning over the last century suggests that, in comparison with strict classical logic, human reasoning is failing at simple logical tasks, committing mistakes in probabilistic reasoning, and succumbing to irrational biases in decision making [3], [4]. Different interpretations and theories on the nature of human reasoning have been proposed to explain these findings. Certain proposals attempt to stay very close to the mathematical and strict form of logical reasoning, such as "The Psychology of Proof" theory [5], which proposes a psychological version of a proof system for human reasoning in the style of Natural Deduction. Despite its many criticisms (see, e.g., [6] for a thorough and critical review of this theory), the theory shows a necessary departure from the proof systems of classical logic.

This paper grew out of tutorials given at IJCAI 2016 and at ECAI 2016. Information on the tutorials is available at: http://cognition.ouc.ac.cy/argument/.

Antonis Kakas is a professor in the Department of Computer Science, University of Cyprus, Nicosia, Cyprus; e-mail antonis@ucy.ac.cy

Loizos Michael is an assistant professor in the School of Pure and Applied Sciences and director of the Computational Cognition Lab, Open University of Cyprus, Nicosia, Cyprus; e-mail loizos@ouc.ac.cy

More importantly, the theory implicitly indicates that human reasoning is linked to argumentation, since proof systems like Natural Deduction are known to have a natural argumentative interpretation [7]. Other proposals (see, e.g., [8]) completely abandon any logical form for human reasoning, treating it as the application of specialized procedures, invoked naturally depending on the situation in which people find themselves.

Earlier work demonstrated empirically that humans perform with significant variation in successfully drawing conclusions under different classical logic syllogisms [9]. The study of the Psychology of Syllogisms [10]–[12] proposes that humans use mental models to guide them into drawing inferences, which foregoes the "absolute and universal" validity of the inferences supported by reasoning based on truth in all possible models of the premises. Instead, a mental model captures reasoning based on the *intended interpretation* of the premises, and corresponds to a suitable situation model, much like what humans construct when processing or comprehending a narrative [13], [14].

In a modern manifestation of this perspective in the context of Computational Logic in AI [15], it is argued that structures like mental models are a useful way to capture various features of human reasoning, not least of which its defeasible nature. Building mental models can be seen as building arguments to support an intended interpretation of the evidence currently available, by combining them with general rules of common sense knowledge that people have acquired. The mental model approach to deduction can, then, be reconciled with the view of reasoning through inference rules, while the defeasible nature of reasoning follows from the defeasible nature of arguments.

In addition to the plethora of psychological findings that are consistent with, and indicative of, an argumentative interpretation of human reasoning, some more recent work from the Psychology of Reasoning provides further explicit evidence in support of this position [16]. Supported by the results of a variety of empirical psychological experiments, the authors of that work propose that human reasoning is a process whereby humans *provide reasons* to accept (or decline) a conclusion that was "raised" by some incoming inference of the human brain. The main function of human reasoning, then, is to lay out these inferences in detail, and to form possible arguments that will produce the final conclusion, in a way characterized by the awareness not just of the conclusion, but of an argument that justifies accepting that conclusion. Through the process of human reasoning, therefore, people become able to exchange arguments for assessing new claims, and the process of human reasoning becomes, effectively, a process of argumentation.

Experiments carried out to test how humans form, evaluate, and use arguments, suggest that humans produce "solid" arguments when motivated to do so; i.e., in an environment where their position is challenged. If unchallenged, the arguments initially produced can be rather naive, until counter-arguments or opposing positions are put forward, at which point humans produce better and well-justified arguments for their position by finding counter-arguments (i.e., defenses) to the challenges. For example, in experiments where mock jurors were asked to reach a verdict and then were presented with an alternative one, it was observed that almost all of them were able to very quickly find counter-arguments against the alternative verdict, while strengthening the arguments for their original verdict.

The experimental results indicate that automating human reasoning through argumentation can follow a model of computation that has an *"on-demand"* incremental nature. Such a model of computation is well-suited in a resource-bounded problem environment, and more generally for the development of cognitive systems under the personal assistant paradigm.

Overall, work from Psychology has exposed some *salient features* of human reasoning directly related to argumentation: *(i)* handling of contradictory information, by acknowledging the defeasible nature of knowledge; *(ii)* drawing of tentative conclusions, which are revised in the presence of more information; *(iii)* awareness not only of a conclusion, but also of its justification; *(iv)* "on demand" / dialectical reasoning that defends challenges as they arise; *(v)* use of a single intended mental model, while accommodating common and individual biases across humans. Collectively, these features suggest that argument is native to human reasoning, and, consequently, that argumentation can offer a unified perspective of empirical psychological evidence on the nature of human reasoning.

III. Computational Argumentation in AI

Efforts to formalize human reasoning in terms of an argumentation theory can be traced back to the work of Aristotle and his notion of "dialectic argument". Until rather recently, argumentation was primarily studied from a philosophical and / or a psychological perspective. These works [17]–[19] have helped generate a new interest on the study of argumentation within the field of AI, with motivation coming from both *(i)* the desire to have intelligent systems with human-like defeasible (or non-monotonic) reasoning, and belief revision capabilities in the face of new information [20], [21], as well as *(ii)* the study of the dialectic nature of reasoning in various areas of human thought, such as rhetoric and legal reasoning [22]–[26].

The early 1990s saw the introduction of *abstract argumentation* [27], where arguments are considered as formal entities separate from the particular context in which they arise, and are viewed only in terms of their syntactic and semantic relationships. This view emerged from work [28], [29] showing that argumentation could capture most of the existing non-monotonic logical frameworks, and, hence, provide a uniform way to view the aspect of defeasibility in human reasoning.

An abstract argumentation framework is defined as a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R}$ is a binary (partial) relation on $\mathcal{A}$, called the *attack relation* on $\mathcal{A}$. This attack relation is lifted to subsets of arguments, so that a subset A of $\mathcal{A}$ attacks another subset B of $\mathcal{A}$ if and only if there exists $a \in A$ and $b \in B$ such that a attacks b; i.e., $(a, b) \in \mathcal{R}$. One is then concerned with the problem of building "good quality" or *acceptable* argument subsets $\Delta \subseteq \mathcal{A}$ that "defend against" or attack back all possible argument subsets that attack Δ, and which constitute, therefore, counter-arguments to Δ.

A general way to formulate a notion of acceptability is the dialectical definition that an argument subset Δ is acceptable if and only if any argument subset A that attacks Δ is attacked back by some argument subset D (i.e., D defends Δ against A) that is, itself, "acceptable with respect to Δ". There are

several different ways to offer a precise formulation of what is meant by the condition that D is acceptable with respect to Δ, such as: *(i)* that D is simply a subset of Δ, which gives rise to the admissibility semantics of argumentation, or *(ii)* that D is (eventually) an argument subset that is not attacked by any other argument subset (and is, hence, globally undisputed), which gives rise to the grounded semantics of argumentation.

This simple, yet powerful, formulation of argumentation has been used as the basis for the study and development of solutions for different types of problems in AI [30], [31]. In particular, it forms the foundation for a variety of problems in multi-agent systems (see, e.g., the workshop series "ArgMAS: Argumentation in Multi-Agent Systems") where agents need to exhibit human-like autonomy and adaptability. Recently, the area of *argument mining* (see, e.g., [32] for an overview) aims to provide an automatic way of analysing, in terms of formal argumentation frameworks, human debates in social media, even by identifying relations that are not explicit in text [33].

In many of the application domains above, a realization of abstract argumentation is used where the attacking relation is materialized through a priority or preference relation between conflicting arguments. Such *preference-based argumentation* frameworks consider more preferred arguments to be stronger than, and thus to attack, less preferred arguments, but not vice-versa. Preferences can be derived naturally from the particular domain of application, capturing general or contextual aspects of the domain, or biases and beliefs of individual agents.

More recently it has been shown that even classical logical reasoning, as found in formal mathematics, can be captured in terms of abstract argumentation [7]. In such an *argumentation-based logic*, logical entailment of some conclusion is obtained through the existence of an acceptable argument supporting the conclusion and the absence of acceptable arguments that support any contrary conclusion. This suggests that argumentation need not be approached as a substitute for classical logic, but as an extension thereof that is appropriate for reasoning both with consistent premises but also with inconsistent ones.

The aforementioned studies of argumentation in AI show that computational argumentation has the capacity to address the salient features of human reasoning that have been pointed out by empirical psychological studies. Argumentation offers a natural form of reasoning with contradictory information, by supporting arguments for conflicting conclusions, and handling the retraction of tentative conclusions whenever new stronger arguments emerge. Furthermore, argumentation gives a form of reasoning that is based on an intended mental model that comprises the conclusions that are supported by the strongest available arguments, and provides explicit justifications in support of that intended model. Lastly, argumentation explicitly adopts "on demand" reasoning through a dialectical definition of acceptability, while its preference-based realization readily accommodates for human biases and individual beliefs.

IV. Argument and Human Decision Making

Having offered evidence for the capacity of computational argumentation to capture the salient features of human reasoning, we turn our attention to how argumentation can be utilized in the development of cognitively-compatible systems.

An important subclass of cognitive systems will be that of *cognitive assistants* that help their human users take decisions in everyday tasks: which restaurant to visit for some occasion, when to schedule a meeting, or how to handle an information overload on a social network. Despite appearing relatively simple when compared with the complex optimization problems that conventional computing systems solve, these everyday decision-making problems come with their own challenges.

Any systematic and principled attempt at developing cognitive assistants needs to account for several characteristics of human decision-making that have been exposed by work in Cognitive Psychology: departure from the formal decision theory and influence by biases (e.g., earliest information, similar past decisions, group conformity), consideration of individual preferences and predispositions, minimal initial evaluation of the options and additional evaluation as the need arises.

Beyond ensuring cognitive compatibility, one must also account for pragmatic considerations. Decision-making is rarely an isolated process, and the arrival of extra or revised information in an *open and dynamic environment* may affect decision-making by: offering new options (e.g., a new restaurant just opened up); rendering existing options (physically) inapplicable (e.g., the boss cannot meet after 11:00am); or revealing updated values for options (e.g., an online community that the user had enjoyed following started using offensive language).

The *challenge of building cognitive assistants* resides, thus, in being able to coherently operate at three levels: *(L1)* represent information akin to the user's general motivations and desires, and the system's beliefs of the state of the world at the time when decisions will be effected; *(L2)* offer explanations / justifications of the proposed decision that are meaningful to the user; *(L3)* participate in a dialectic debate process to either persuade the user of the proposed decision, or to revise it.

The natural solution that argumentation offers for level *(L2)* and level *(L3)* has been utilized in several works in AI dealing with decision-making in contexts ranging from legal decisions to informal human-like decisions by autonomous agents [34]–[39]. These works generally fall within the particular realization of preference-based argumentation, which also points to how argumentation can offer a solution for level *(L1)*, as well.

In an argumentation framework for a certain decision problem, each option is supported by one or more arguments. The structure of these *option arguments* can be represented simply by a tuple $\langle opt, val, bel \rangle$, where *opt* is the supported option, *val* is a set of user values (e.g., needs, motivations, desires) that the option and / or the argument serve, and *bel* is a set of beliefs that ground the argument on some information about the external world, in a manner that the cognitive assistant believes render option *opt* a possible alternative for consideration.

The values served by an argument can give a relative preference between arguments that reflects the *personal affinity* or interests that a cognitive assistant might be designed to follow. Thus, a preference between arguments $a_i = \langle opt_i, val_i, bel_i \rangle$ and $a_j = \langle opt_j, val_j, bel_j \rangle$ can be defined through the general schema that "a_i is preferred over a_j if $val_i \sqsupset val_j$", where $\sqsupset$ is a comparison ordering amongst the different values that can be based on the personality of the human user of the cognitive assistant. By concentrating on different values or on different

comparison orderings, this simple general schema can give rise to different preferences over the same arguments, reflecting the natural variability across different contexts or different users.

In practice, human users may know heuristically from their earlier experiences the result of the evaluation of different arguments in certain situations, and hence that certain arguments are preferred over others. For example, a vegetarian may know that when having dinner outside her house, the vegetarian restaurant down town serves better her need to have a larger variety of choices, but the local market across the street offers a cheaper and faster choice. Instead of having to recompute her preferences based on her values for the two options, she might choose to simply state that when she is in a situation $S_{i,j}$ where she is currently at home and she has not eaten out during the past week, then she prefers the argument supporting the local market over the argument supporting the vegetarian restaurant, using the general scheme "a_i is preferred over a_j if $S_{i,j}$", where effectively $S_{i,j}$ is a situation in which $val_i \sqsupset val_j$.

The attack relation can be naturally defined from the preferences as follows: argument a_i attacks argument a_j if they support conflicting options, and a_i is not less preferred than a_j. Now, given a state S of the world, one can compute (under a chosen argumentation semantics) the acceptable arguments among those whose beliefs are compatible with S. Any option supported by an acceptable argument is a *possible* decision in S. Furthermore, a possible decision in S is a *clear* decision in S if there exist no other conflicting possible decisions in S.

To complete the picture, one must fix the choice of argumentation semantics. Given the nature of human decision-making, the natural choice for this case, and for the respective cognitive assistants, is that of the *grounded extension* semantics, which, as already discussed in the preceding section, can be derived as a special case of the dialectical definition of acceptability.

In summary, argumentation serves well as a basis for cognitive assistants that support human decision-making, offering natural solutions: *(L1)* at the representation level through the encoding of user-specific preferences and biases; *(L2)* at the decision-formation level through the incremental construction of acceptable arguments; *(L3)* at the persuasion level through the dialectic process of defending against alternative decisions.

V. Argument and Narrative Comprehension

In describing how abstract argumentation can be instantiated to support human decision-making, we have focused primarily on the role that values play in capturing the user's general motivations and desires, and have mostly side-stepped the role that beliefs play in capturing the applicability of arguments.

In the simplest case, these beliefs could be such that their compatibility against a given state of the world can be directly checked, outside the actual process of argumentation. More generally, though, the beliefs themselves are the outcome of a reasoning process, which could itself be argumentative. Thus, in addition to option arguments, the argumentation framework may also include *belief arguments*, supporting beliefs on which the option arguments rest. An option argument could, then, be potentially undercut by a belief argument that supports that the environment will not be conducive to the realization of the particular option, while a second belief argument that disputes this claim could be used to defend the option argument.

Exactly analogously to option arguments, belief arguments are evaluated against each other by means of a preference relation, which ultimately determines the attack relation between arguments. Unlike the typically user-specific preferences over option arguments, however, preferences over belief arguments naturally capture certain pragmatic considerations. These considerations derive primarily from the open and dynamic nature of the environment, which necessitates a cognitive assistant able to reason about missing information, the causal effects of actions, the passage of time, and the typical and exceptional states of the world. Belief arguments capture, then, knowledge about these aspects of the world, while preferences over belief arguments capture the commonsensical reasoning pattern that humans use to form a coherent understanding of the situation.

This type of reasoning is directly related to the process of narrative comprehension, with the coherent understanding of the situation corresponding to the intended interpretation of the narrative. During narrative comprehension, humans include in the intended interpretation information that is not explicitly present in the narrative but follows from it, explanations of why things happened as they did, links between seemingly unconnected events, and predictions of how things will evolve.

Starting with the seminal works of the Situation and Event Calculi, work in AI sought to codify the commonsense laws associated with reasoning about actions and change (RAC) in a narrative context, in terms of central problems to be solved: the frame problem of how information persists, by default, across time; the ramification problem of how actions give rise to indirect effects; the qualification problem of how action effects are blocked from materializing; the state default problem of how the world is not, by default, in some exceptional state.

Several works in AI [40]–[43] have demonstrated the natural fit of argumentation for RAC, by capturing the relevant aspects of human reasoning in terms of persistence, causal, and default property arguments, along with a natural preference relation between these different types of arguments. For example, a preference of causal arguments over conflicting persistence arguments cleanly addresses the frame problem by capturing the commonsense law of inertia that situations / information persist unless caused to change. Grounding the different types of arguments on information explicitly given in the narrative allows one to offer explanations for / against drawing certain conclusions at certain time-points or situations in the world.

Recent efforts [44] to combine an argumentation approach to RAC with empirical knowhow and theoretical models from Cognitive Psychology have led to the development of automated comprehension systems [45] that use belief arguments (under the grounded extension semantics, which we have proposed as appropriate for decision-making as well) to construct an intended mental model for a narrative, and appropriately update and maintain it in the presence of surprises and twists as the narrative unfolds. This treatment is not unlike what a cognitive assistant is expected to adopt when reasoning about its beliefs while the state of its environment unfolds over time.

Despite their predominant use to represent knowledge about the environment, belief arguments used by a cognitive assistant

cannot be decoupled from its human user. The vocabulary and terms employed to express belief arguments should be familiar, their complexity should be manageable, and the justifications they give rise to should be meaningful, all with respect to the user. For example, a cognitive assistant's appeal to the belief argument that "pyrexia is not a medical emergency" might be inappropriate if its user is not familiar with the term "pyrexia" (fever), or if its user has been close to swamps (in which case fever might be indicative of malaria). These issues tie directly back to the requirement that cognitive assistants should operate in a manner that is cognitively-compatible with their users.

In summary, the argumentation basis for human decision-making, as proposed in the preceding section, can be naturally extended to address, *within a single unified framework*, the related and complementary problem of narrative comprehension: *(L1)* at the representation level through the encoding of world knowledge; *(L2)* at the decision-formation level through the construction of justifications that use concepts meaningful to the user; *(L3)* at the persuasion level through the grounding / contextualization of decisions on the fluctuating world state.

VI. Populating the Argumentation Arena

The acceptability semantics of computational argumentation can be effectively viewed as an *internal evaluation mechanism* for the quality of the conclusions of a cognitive assistant, with conclusions that are supported by stronger or more preferred (subsets of) arguments considered as being more pertinent than alternatives. Argumentation, however, does not provide for an analogous *external evaluation mechanism* for the quality of the cognitive assistant's arguments and preferences *in relation to* the environment. Equivalently, an argumentation framework is assumed to be populated with arguments and preferences of high external quality, and the acceptability semantics concentrates on the task of how to make a meaningful use of those.

A central means to populate an argumentation framework is through *cognitive programming* [1]. The user can volunteer, either during an initial familiarization period, or dialectically in response to a failure to be persuaded by the cognitive assistant, additional belief or option arguments and corresponding preferences, so that the cognitive assistant gradually becomes more "knowledgeable" about its environment, and better reflects the motivations and interests of its user. The requirement that a cognitive assistant's representation is cognitively-compatible with humans is key during this process, as the user naturally interacts with its cognitive assistant through the use of high-level concepts, and in a way that avoids detailed instructions.

More passively, the user may simply decline suggestions of the cognitive assistant without offering explanations / counter-arguments. Some form of online supervised learning can then be invoked by the cognitive assistant, with the user's feedback taken as a negative learning instance that the arguments supporting a decision are not acceptable, and that the preferences among arguments need to be revised to account for this. Under certain assumptions, the user's preferences between competing option arguments have been shown to be learnable [46], [47].

As an example of cognitive programming, the suggestion of a cognitive assistant to schedule a meeting of its user with his boss at 7:30am can be met by the user's response "Do not schedule work appointments too early in the morning.", which will thereafter be viewed as an extra argument, more preferred than the acceptable arguments that supported the suggestion. The importance of forming an intended model of the situation, and of the ability to employ common sense, is highlighted in this example, as the cognitive assistant needs to make sense of terms like "work appointment" and "too early in the morning".

Certain general belief arguments (e.g., that most people do not usually work during the nighttime) can be more reasonably acquired directly by the cognitive assistant, through manually-engineered or croudsourced knowledge-bases [48]–[51], and through the use of machine learning on text corpora [52]–[54]. A number of issues would, of course, have to be dealt with: the possible biases in the learning material, especially for text found on the Web [55]; the eventual use of the arguments by a reasoning process, without nullifying their learning-derived guarantees [56]–[59]; the availability of appropriate learning material to also acquire causal arguments [60]; the inadvertent effects of decisions supported by arguments that were learned outside the environment of the cognitive assistant [61], [62].

The autonomous or crowdsourced acquisition of arguments and preferences still benefits from user interaction. A cognitive assistant used for appointment scheduling, for example, typically has no immediate need for learning about forest fires. The user (or the manufacturer) can specify, then, relevant keywords to guide the cognitive assistant's search for applicable learning material. Alternatively, such keywords can be identified by the cognitive assistant by gathering those that occur frequently in its user's queries, so that autonomous learning can be invoked "on demand". Once arguments and preferences are learned, the user may further correct any misconceptions that have been acquired due to biases in the sources of the learning material.

It is important to note here that none of the processes of populating an argumentation framework restricts the application of cognitive assistants to common sense domains only. A medical professional, for instance, could cognitively program a cognitive assistant with arguments for making diagnoses of illnesses based on observed symptoms. The cognitive assistant could also autonomously learn medical arguments by restricting its search for learning material to medical ontologies and journals. Through these arguments, then, the cognitive assistant would be able to explain its medical recommendations in the same fashion that one medical professional would explain to another.

We conclude by observing that argumentation is not simply amenable to a process of learning, but rather a *natural fit* for it. Learned knowledge, especially when acquired autonomously by a cognitive assistant, cannot be strict, but can express only typical and defeasible relationships between concepts, with the strength of the relationships depending on the various contexts of the application domain. In philosophical terms, the process of inductive syllogism, as Aristotle calls the process of acquiring first principles from experience, cannot produce absolute knowledge. An inductively produced implication $X \rightarrow Y$ does not formally express the "necessity" of Y when X is known, but rather an argument for Y when X is known, thus making Y "probable" in this particular case, as the philosopher David Hume [63] suggests. Recent work seeks to acquire knowledge

that is directly expressible in the form of such arguments [64].

VII. Towards a Future of Cognitive Systems

In its early days, Artificial Intelligence had sought to understand human intelligence and to endow machines with human-like cognitive abilities. Since then, however, AI has evolved primarily as an engineering discipline, placing emphasis on the development of useful specialized tools, and effectively abandoning the scientific inquiry into what constitutes intelligent behavior. In a sense, then, cognitive systems embody a modern realization of the need to return to AI's scientific roots, while adopting the engineering goal of developing useful systems.

This paper has sought to argue that computational argumentation in AI can offer a principled basis for the development of cognitive systems for everyday tasks. We have discussed work from Psychology showing that human cognition is inherently argumentative, and we have demonstrated that computational argumentation naturally encompasses several salient features of everyday human cognition — contra to the prevalent (even if implicit) assumption that classical logic can serve this role.

Given this new logical foundation for cognitive systems, one could reasonably ask whether it would necessitate a novel computing architecture on which to be realized, much like the Von Neumann architecture realizes classical (Boolean) logic. A neural-symbolic architecture (see, e.g., the workshop series "NeSy: Neural-Symbolic Learning and Reasoning") could potentially serve this role, with excitatory and inhibitory links implementing supports and attacks within an argumentation framework. Such an architecture could also allow the utilization of modern advances in deep learning, integrating within the reasoning architecture the process of learning and revision.

The view of logic reasoning as capturing the "laws of human thought" has served AI and Computer Science well. With an eye towards the development of cognitive systems, we would posit that it would be equally serving to view computational argumentation as capturing the "laws of common sense thought".

Acknowledgments

We would like to thank our colleagues, Irianna Diakidoy, Bob Kowalski, Hugo Mercier, Rob Miller, Pavlos Moraitis, Nikos Spanoudakis, Francesca Toni, and György Turán, for useful discussions and joint work on the topics of this paper.

References

[1] L. Michael, A. Kakas, R. Miller, and G. Turán, "Cognitive Programming," in *Proceedings of the 3rd International Workshop on Artificial Intelligence and Cognition*, 2015, pp. 3–18.

[2] P. Langley, "The Cognitive Systems Paradigm," *Advances in Cognitive Systems*, vol. 1, pp. 3–13, 2012.

[3] J. Evans, "Logic and Human Reasoning: An Assessment of the Deduction Paradigm," *Psychological Bulletin*, vol. 128, no. 6, pp. 978–996, 2002.

[4] D. Kahneman and A. Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, vol. 3, no. 3, pp. 430–454, 1972.

[5] L. Rip, *The Psychology of Proof: Deductive Reasoning in Human Thinking*. MIT Press, 1994.

[6] P. Johnson-Laird, "Rules and Illusions: A Critical Study of Rips's The Psychology of Proof," *Minds and Machines*, vol. 7, no. 3, pp. 387–407, 1997.

[7] A. Kakas, F. Toni, and P. Mancarella, "Argumentation Logic," in *Proceedings of the 5th International Conference on Computational Models of Argument*, 2014, pp. 345–356.

[8] P. Cheng and K. Holyoak, "Pragmatic Reasoning Schemas," *Cognitive Psychology*, vol. 17, no. 4, pp. 391–416, 1985.

[9] G. Stoerring, "Experimentelle Untersuchungen üeber einfache Schlussprozesse," *Archiv fuer die gesammte Psychologie*, vol. 11, no. 1, pp. 1–127, 1908.

[10] P. Johnson-Laird, *Mental Models*. Cambridge University Press, 1983.

[11] P. Johnson-Laird and M. Steedman, "The Psychology of Syllogisms," *Cognitive Psychology*, vol. 10, no. 1, pp. 64–99, 1978.

[12] P. Johnson-Laird and R. Byrne, *Deduction*. Lawrence Erlbaum Associates, 1991.

[13] K. Stenning and M. van Lambalgen, *Human Reasoning and Cognitive Science*. MIT Press, 2008.

[14] ——, "Reasoning, Logic, and Psychology," *WIREs Cognitive Science*, vol. 2, no. 5, pp. 555–567, 2010.

[15] R. Kowalski, *Computational Logic and Human Thinking: How to Be Artificially Intelligent*. Cambridge University Press, 2011.

[16] H. Mercier and D. Sperber, "Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences*, vol. 34, no. 2, pp. 57–74, 2011.

[17] S. Toumlin, *The Uses of Argument*. Cambridge University Press, 1958.

[18] C. Perelman and L. Olbrechts-Tyteca, *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, 1969.

[19] J. Pollock, "Defeasible Reasoning," *Cognitive Science*, vol. 11, no. 4, pp. 481–518, 1987.

[20] J. McCarthy, "Programs with Common Sense," in *Semantic Information Processing*. MIT Press, 1968, pp. 403–418.

[21] C. Alchourrón, P. Gärdenfors, and D. Makinson, "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions," *Symbolic Logic*, vol. 50, no. 2, pp. 510–530, 1985.

[22] N. Rescher, *Dialectics: A Controversy-Oriented Approach to the Theory of Knowledge*. State University of New York Press, 1977.

[23] P. Valesio, *Novantiqua: Rhetorics as a Contemporary Theory*. Indiana University Press, 1980.

[24] A. Gardner, *An Artificial Intelligence Approach to Legal Reasoning*. MIT Press, 1987.

[25] T. Gordon, *The Pleadings Game: an Artificial Intelligence Model of Procedural Justice*. Kluwer Academic Publishers, 1995.

[26] R. Loui and J. Norman, "Rationales and Argument Moves," *Artificial Intelligence and Law*, vol. 3, no. 3, pp. 159–189, 1995.

[27] P. Dung, "On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games," *Artificial Intelligence*, vol. 77, no. 2, pp. 321–357, 1995.

[28] A. Kakas, R. Kowalski, and F. Toni, "Abductive Logic Programming," *Logic and Computation*, vol. 2, no. 6, pp. 719–770, 1992.

[29] A. Bondarenko, F. Toni, and R. Kowalski, "An Assumption-Based Framework for Non-Monotonic Reasoning," in *Proceedings of the 2nd International Workshop on Logic Programming and Non-Monotonic Reasoning*, 1993, pp. 171–189.

[30] T. Bench-Capon and P. Dunne, "Argumentation in Artificial Intelligence," *Artificial Intelligence*, vol. 171, no. 10–15, pp. 619–641, 2007.

[31] I. Rahwan and G. Simari, *Argumentation in Artificial Intelligence*. Springer Publishing Company, 2009.

[32] M. Lippi and P. Torroni, "Argumentation Mining: State of the Art and Emerging Trends," *ACM Transactions on Internet Technology*, vol. 16, no. 2, pp. 10:1–10:25, 2016.

[33] F. Toni and P. Torroni, "Bottom-Up Argumentation," in *Proceedings of the 1st International Workshop on Theories and Applications of Formal Argumentation*, 2012, pp. 249–262.

[34] T. Bench-Capon, "Persuasion in Practical Argument using Value-Based Argumentation Frameworks," *Logic and Computation*, vol. 13, no. 3, pp. 429–448, 2003.

[35] N. Karacapilidis and D. Papadias, "Computer Supported Argumentation and Collaborative Decision Making: The HERMES System," *Information Systems*, vol. 26, no. 4, pp. 259–277, 2001.

[36] L. Amgoud and H. Prade, "Using Arguments for Making and Explaining Decisions," *Artificial Intelligence*, vol. 173, no. 3–4, pp. 413–436, 2009.

[37] A. Kakas, L. Amgoud, G. Kern-Isberner, N. Maudet, and P. Moraitis, "ABA: Argumentation Based Agents," in *Proceedings of the 8th International Workshop on Argumentation in Multiagent Systems*, 2011, pp. 9–27.

[38] A. Kakas and P. Moraitis, "Argumentation Based Decision Making for Autonomous Agents," in *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, 2003, pp. 883–890.

[39] B. Verheij, "Formalizing Value-Guided Argumentation for Ethical Systems Design," *Artificial Intelligence and Law*, vol. 24, no. 4, pp. 387–407, 2016.
[40] N. Foo and Q. Vo, "Reasoning about Action: An Argumentation-Theoretic Approach," *Artificial Intelligence Research*, vol. 24, pp. 465–518, 2005.
[41] E. Hadjisoteriou and A. Kakas, "Reasoning about Actions and Change in Argumentation," *Argument and Computation*, vol. 6, no. 3, pp. 265–291, 2015.
[42] A. Kakas, R. Miller, and F. Toni, "An Argumentation Framework of Reasoning about Actions and Change," in *Proceedings of the 5th International Conference on Logic Programming and Non-Monotonic Reasoning*, 1999, pp. 78–91.
[43] L. Michael, "Story Understanding... Calculemus!" in *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, 2013.
[44] I. Diakidoy, A. Kakas, L. Michael, and R. Miller, "Story Comprehension through Argumentation," in *Proceedings of the 5th International Conference on Computational Models of Argument*, 2014, pp. 31–42.
[45] ——, "STAR: A System of Argumentation for Story Comprehension and Beyond," in *Proceedings of the 12th International Symposium on Logical Formalizations of Commonsense Reasoning*, 2015, pp. 64–70.
[46] Y. Dimopoulos, L. Michael, and F. Athienitou, "Ceteris Paribus Preference Elicitation with Predictive Guarantees," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009, pp. 1890–1895.
[47] L. Michael and E. Papageorgiou, "An Empirical Investigation of Ceteris Paribus Learnability," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 1537–1543.
[48] D. Lenat, "CYC: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.
[49] F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 697–706.
[50] R. Speer and C. Havasi, "ConceptNet 5: A Large Semantic Network for Relational Knowledge," in *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, I. Gurevych and J. Kim, Eds. Springer Berlin Heidelberg, 2013, pp. 161–176.
[51] C. Rodosthenous and L. Michael, "A Hybrid Approach to Commonsense Knowledge Acquisition," in *Proceedings of the 8th European Starting AI Researcher Symposium*, 2016, pp. 111–122.
[52] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr., and T. Mitchell, "Toward an Architecture for Never-Ending Language Learning," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, pp. 1306–1313.
[53] L. Michael and L. Valiant, "A First Experimental Demonstration of Massive Knowledge Infusion," in *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning*, 2008, pp. 378–389.
[54] L. Michael, "Reading Between the Lines," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009, pp. 1525–1530.
[55] ——, "Machines with WebSense," in *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, 2013.
[56] L. Valiant, "Robust Logics," *Artificial Intelligence*, vol. 117, no. 2, pp. 231–253, 2000.
[57] L. Michael, "Autodidactic Learning and Reasoning," Doctoral Dissertation, Harvard University, Cambridge, Massachusetts, U.S.A., 2008.
[58] ——, "Simultaneous Learning and Prediction," in *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning*, 2014, pp. 348–357.
[59] H. Skouteli and L. Michael, "Empirical Investigation of Learning-Based Imputation Policies," in *Proceedings of the 2nd Global Conference on Artificial Intelligence*, 2016, pp. 161–173.
[60] L. Michael, "Causal Learnability," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 1014–1020.
[61] ——, "The Disembodied Predictor Stance," *Pattern Recognition Letters*, vol. 64, no. C, pp. 21–29, 2015, Special issue on 'Philosophical Aspects of Pattern Recognition'.
[62] ——, "Introspective Forecasting," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 3714–3720.
[63] D. Hume, *A Treatise of Human Nature*, L. Selby-Bigge, Ed. Clarendon Press, 1888, Originally published in 1738-1740.
[64] L. Michael, "Cognitive Reasoning and Learning Mechanisms," in *Proceedings of the 4th International Workshop on Artificial Intelligence and Cognition*, 2016.

AI for Traffic Analytics

Raghava Mutharaju, Freddy Lécué, Jeff Z. Pan, Jiewen Wu, Pascal Hitzler

***Abstract*—Information and communications technology (ICT) is used extensively to better manage the city resources and improve the quality of life of its citizens. ICT spans many departments of the cities, from transportation, water, energy to building management and social-care services. AI techniques are getting more and more attraction from cities to represent and organize information, maintain sustainable networks, predict incidents, optimize distribution, diagnose faults, plan routes and organize their infrastructure. Managing traffic efficiently, among many other domains in cities, is one of the key issues in large cities. In this article we describe the domains of applications which could benefit from AI techniques, along with introducing the necessary background knowledge. Then we focus on traffic applications, which make use of recent AI research in knowledge representation, logic programming, machine learning and reasoning. Specifically we go through the next version of scalable AI driven traffic related application where (1) data from a variety of sources is collected, (2) knowledge about traffic, vehicles, citizens, events is represented and (ii) deductive and inductive reasoning is combined for diagnosing and predicting road traffic congestion. Based on these principles, a real-time, publicly available AI system named STAR-CITY was developed. We discuss the results of deploying STAR-CITY, and its related AI technologies in cities such as Dublin, Bologna, Miami, Rio and the lessons learned. We also discuss the future AI opportunities including scalability issues for large cities.**

***Index Terms*—Artificial intelligence, Knowledge representation, Smart cities, Traffic congestion**

I. Introduction

MORE and more people are moving to the cities in search of better livelihood. The resources and the infrastructure of the cities are unable to keep up with this population growth rate. This leads to several problems such as shortage of water and electricity, increase in pollution and severe traffic congestion, which is one of the major transportation issues in most industrial countries [1]. Traffic congestion leads to massive wastage of time and resources such as fuel. In USA, traffic congestion leads to 5.5 billion hours of delay and 2.9 billion gallons of wasted fuel costing around $121 billion [2]. Apart from such wastage, traffic congestions also lead to road rage and accidents. Three possible ways to reduce traffic congestion [3] are i) improving the road infrastructure, ii) promoting the use of public transport and iii) diagnosing and predicting traffic congestions, which allows city administrators to proactively manage the traffic. Among the three options, the third option is the most cost effective and convenient since it does not involve any change to the existing infrastructure. In this work, we use several AI techniques such as knowledge representation and reasoning, planning and machine learning to predict and diagnose traffic congestions.

There are several existing traffic analysis tools such as US Traffic View[1] [4], French Sytadin[2] and Italian 5T[3]. They support basic analytics, visualization and monitor traffic using dedicated sensors. They cannot handle data coming from heterogeneous sources and do not interpret traffic anomalies. Other systems such as the traffic layer of Google Maps provide real-time traffic conditions but do not take into account the historical data and data from other sources such as weather and city events. Thus the existing systems do not take advantage of the context and the semantics of the data.

Data from several sources provide key insights into the location, cause and intensity of the traffic congestion. User generated content such as tweets, weather conditions, information about city events (music concerts etc) can be used along with the traffic data. Semantic Web technologies such as OWL (Web Ontology Language) [5] and RDF (Resource Description Framework) [6], which are also W3C recommendations, can be used to represent knowledge and integrate data from multiple data sources. These technologies provide structure and meaning to the data as well as enable interlinking, sharing and reuse of the data.

RDF is a framework to describe resources such as documents, people, physical objects, abstract concepts etc. Resources are described in the form of triples, where a triple consists of three parts: subject, predicate and object. For example, we can represent road r_1 is adjacent to road r_2 in the form of a triple as $<r_1>$ `<isAdjacentTo>` $<r_2>$.

OWL is more expressive compared to RDF and is used to build ontologies that represent knowledge about things, groups of things and relation between them. It is used to formally encode domain knowledge, i.e., knowledge about some part of the world which is often referred to as the domain of interest. In order to build an ontology, it is important to come up with the vocabulary of the domain, i.e., a set of terms and the relationships between them. These form the axioms in an ontology. The knowledge in an ontology can be categorized into terminological knowledge and assertions. The

Raghava Mutharaju is a research scientist in the Knowledge Discovery Lab of GE Global Research, Niskayuna, NY, USA. Email: raghava.mutharaju@ge.com

Freddy Lécué is a principal scientist and research manager in large scale reasoning systems in Accenture Technology Labs, Dublin, Ireland. He is also a research associate at INRIA, in WIMMICS, Sophia Antipolis, France. Email: freddy.lecue@accenture.com

Jeff Z. Pan is a Reader in the Department of Computing Science at University of Aberdeen, where he is the Deputy Director of Research of the department. Email: jeff.z.pan@abdn.ac.uk

Jiewen Wu is a lead research scientist at Accenture Technology Labs, Dublin, Ireland. Email: jiewen.a.wu@accenture.com

Pascal Hitzler is an endowed NCR Distinguished Professor and Director of Data Science at the Department of Computer Science and Engineering at Wright State University, Dayton, Ohio, USA. Email: pascal.hitzler@wright.edu

[1] https://www.trafficview.org/
[2] http://www.sytadin.fr/
[3] http://www.5t.torino.it/5t/

terminological knowledge or TBox defines the general notions or the conceptualization of the domain whereas the assertional knowledge or ABox defines the concrete notions or facts of the domain. In a database setting, TBox corresponds to the schema and ABox corresponds to the data [7].

Description logics [8], [9] provide the formal underpinnings for OWL. They are fragments of first order logic, with most of them being decidable. They have formal semantics, i.e., a precise specification of the constructs that make up various description logics. This makes them unambiguous and suitable for logical operations. Description logics provide three types of entities: concepts, roles and individual names. Concepts are sets of individuals, roles represent the binary relations between the individuals and individual names represent single individuals in the domain. In first order logic, these three entities correspond to unary predicates, binary predicates and constants. In OWL, concepts and roles are referred to as classes and properties.

The traffic analytics system named STAR-CITY (Semantic Traffic Analytics and Reasoning for CITY) [10] makes use of RDF and description logics to represent the knowledge in the traffic domain and integrate, reason over data from heterogeneous sources [11]. In the rest of the article, we describe the diagnosis and prediction of traffic congestion using STAR-CITY and the lessons learned from deploying STAR-CITY in Dublin (Ireland), Bologna (Italy), Miami (USA) and Rio (Brazil).

Source Type	Data Source	Description	City: Dublin (Ireland)	Bologna (Italy)	Miami (USA)	Rio (Brazil)
Traffic Anomaly	Journey travel times across the city	Traffic Department's TRIPS system[a]	CSV format (47 routes, 732 sensors) 0.1 GB per day[b]	✗ (not available)		
	Dublin Bus Dynamics	Vehicle activity (GPS location, line number, delay, stop flag)	✗ (not used)	SIRI: XML format[b] (596 buses, 80KB per update 11GB per day[d])	CSV format (893 buses, 225 KB per update 43 GB per day[e])	CSV format (1, 349 buses, 181 KB per update 14 GB per day[f])
Traffic Diagnosis	Social-Media Related Feeds	Reputable sources of road traffic conditions in Dublin City	"Tweet" format - Accessed through Twitter streaming API[g]			
			Approx. 150 tweets per day[h] (approx. 0.001 GB)	✗ (not available)	Approx. 500 tweets per day[i] (approx. 0.003 GB)	✗ (not available)
	Road Works and Maintenance		PDF format (approx. 0.003 GB per day[j])	XML format (approx. 0.001 GB per day[k])	HTML format (approx. 0.001 GB per day[l])	✗ (not available)
	Social events e.g., music event, political event	Planned events with small attendance	XML format - Accessed once a day through Eventbrite[m] APIs			
			Approx. 85 events per day (0.001 GB)	Approx. 35 events per day (0.001 GB)	Approx. 285 events per day (0.005 GB)	Approx. 232 events per day (0.01 GB)
		Planned events with large attendance	XML format - Accessed once a day through Eventful[m] APIs			
			Approx. 180 events per day (0.05 GB)	Approx. 110 events per day (0.04 GB)	Approx. 425 events per day (0.1 GB)	Approx. 310 events per day (0.08 GB)
	Bus Passenger Loading / Unloading (information related to number of passenger getting in / out)		✗ (not available)	✗ (not available)	CSV format (approx. 0.8 GB per day[e])	CSV format (approx. 0.1 GB per day[e])

[a] Travel-time Reporting Integrated Performance System - http://www.advantechdesign.com.au/trips
[b] http://dublinked.ie/datastore/datasets/dataset-215.php (live)
[c] Service Interface for Real Time Information - http://siri.org.uk
[d] http://82.187.83.50/GoogleServlet/ElaboratedDataPublication (live)
[e] Private Data - No Open data
[f] http://data.rio.rj.gov.br/dataset/gps-de-onibus/resource/cfeb367c-c1c3-4fa7-b742-65c2c99d8d90 (live)
[g] https://sitestream.twitter.com/1.1/site.json?follow=ID
[h] https://twitter.com/LiveDrive - https://twitter.com/aaroadwatch - https://twitter.com/GardaTraffic
[i] https://twitter.com/fl511_southeast
[j] http://www.dublincity.ie/RoadsandTraffic/ScheduledDisruptions/Documents/TrafficNews.pdf
[k] http://82.187.83.50/TMC_DATEX/
[l] http://www.fl511.com/events.aspx
[m] https://www.eventbrite.com/api - http://api.eventful.com

Fig. 1. (Raw) Data Sources used for traffic analytics in Dublin, Bologna, Miami and Rio

II. Semantic Representation and Enrichment of Traffic Data

Traffic on the road can be influenced by a variety of factors such as weather conditions, road works and city events. Accordingly, data from different sources such as sensors, tweets, weather information, city events information etc has to be considered. This can be considered as *Big Data* since the data has all the four important characteristics: volume, velocity, variety and veracity. Figure 1 shows the main attributes of the datasets we considered for traffic analytics.

The next step is to convert the all the heterogeneous data shown in Figure 1 into a homogeneous semantic representation. This representation is useful for comparing and evaluating different contexts e.g., events (and their properties: venue, category, size, types and their subtypes), weather information (highly, moderate, low windy, rainy; good, moderate, bad weather condition). More importantly, semantic representation of data helps in (automatically) designing, learning, applying rules at reasoning time for analysis, diagnosis and prediction components. The static background knowledge and the semantics of the data stream is encoded in an ontology which is in OWL 2 EL profile[4]. $\mathcal{EL}^{++}$ is the description logic underpinning for OWL 2 EL. The selection of the OWL 2 EL profile from among the three OWL 2 profiles has been guided by (i) the expressivity which was required to model semantics of data in our application domain (cf. Figure 1), (ii) the scalability of the underlying basic reasoning mechanisms we needed in our stream context e.g., subsumption in OWL 2 EL is in PTIME [12].

All the data streams in Figure 1 are converted to OWL 2 EL ontology streams using IBM Infosphere Streams [13]. Conversion into streams allows i) easy synchronization and transformation of streams into OWL 2 EL ontology, ii) flexible and scalable composition of stream operations, iii) identification of patterns and rules over different time windows, and iv) possible extension to higher throughput sensors. Depending on the data format, different conversion strategies are used - XSLT for XML, TYPifier [14] for tweets and custom OWL 2 EL mapping for CSV. This is shown in Figure 2.

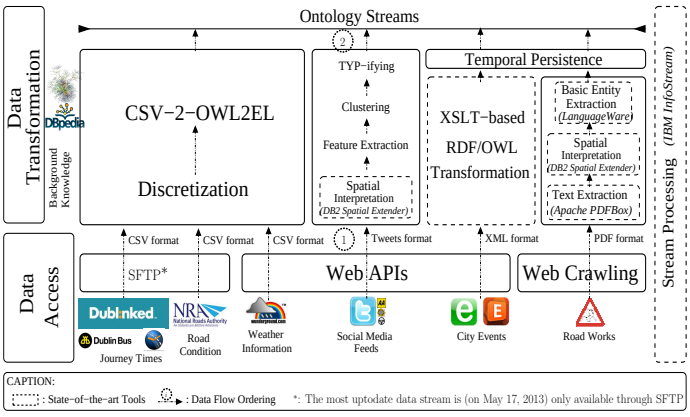


Fig. 2. Semantic Stream Enrichment

[4]https://www.w3.org/TR/owl2-profiles/

III. Diagnosis of Traffic Congestions

Diagnosis task consists of providing a possible explanation for the congestion on a particular road. There can be several reasons for causing or aggravating a traffic congestion. We focus on traffic accidents, road works, weather conditions and social events (e.g., music, political events). Diagnosis of traffic congestion consists of two steps - historic diagnosis computation and real-time diagnosis [15]. This is shown in Figure 3.

All the historic diagnosis information is represented as a deterministic finite state machine. Events, road works and weather conditions are connected to historic traffic congestions along with the probability of those factors indeed causing the congestion. Road intersections and car park locations form the states in the finite state machine. Roads are the transitions in the finite state machine and each road is labeled by its historic diagnosis information.

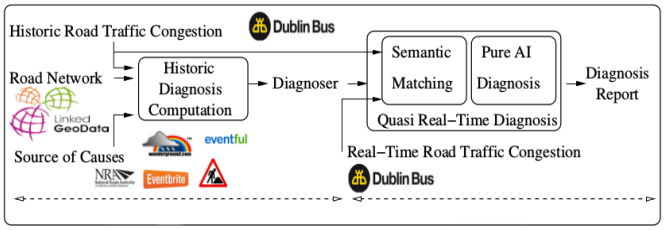


Fig. 3. Overview of the approach to diagnose traffic congestions

After constructing the finite state machine off-line, the next step is to compare the new (current) road condition with the historical condition in real-time and generate a diagnosis report. In existing diagnosis approaches, unless it is an exact match, it is not possible to obtain the diagnosis information. In our approach, we define a matching function that matches the new condition, C_n, which is a description logic concept, with the historical condition, C_h. Note that a condition can be a city event, road work or weather condition which is represented using either existing vocabularies such as DBpedia[5], SKOS[6] or OWL 2 EL ontologies. The matching function gives out the relation between C_n and C_h as output, which could be one of the following.

1) Exact: C_n and C_h are equivalent concepts
2) PlugIn: C_n is a sub-concept of C_h
3) Subsume: C_n is a super-concept of C_h
4) Intersection: The intersection of C_n and C_h is satisfiable

The diagnosis report is constructed using concept abduction between C_n and C_h [16]. The constructed description specifies the under specification in C_h in order to completely satisfy C_n. Computing a diagnosis report is a PTIME problem due to the PTIME complexity of abduction and subsumption in OWL 2 EL.

A crucial step in the diagnosis and prediction of traffic congestions is the classification of ontology streams. Classifying an ontology involves the computation of all the possible sub-concepts for each concept in the ontology. Apart from making implicit sub-concept relationships explicit, classification is also useful for the matching based computation in diagnosis and prediction. Streaming data, which in turn is converted into ontology streams, is considered for the diagnosis and prediction tasks. This would lead to the accumulation of large number of ontologies over a short period of time. Existing reasoners, which are used to classify an ontology, do not scale to large ontologies [17]. A distributed reasoner that can scale with the ontology size is required. DistEL [18] is a distributed and scalable reasoner for the OWL 2 EL profile. An ontology in OWL 2 EL profile can be partitioned based on the different axiom types it supports. Each classification rule of OWL 2 EL (description logic $\mathcal{EL}^{++}$) [19] is applicable to an axiom of one particular type. The partitioned ontology pieces along with the correspond completion rules are distributed across the nodes in the cluster. Each node is dedicated to axioms of at most one particular type and runs the appropriate completion rule on such axioms. This technique improves the data locality and decreases the inter-node communication. A detailed description of other distributed reasoning approaches for OWL 2 EL are described in [20].

[5]http://wiki.dbpedia.org/services-resources/ontology
[6]https://www.w3.org/TR/skos-primer/

IV. Forecasting Traffic Congestions

Predicting or forecasting the anomalies such as traffic congestion involves tracking and correlating the changes (evolution) in the data streams over time [21]. This involves three challenges i) handling the variety and velocity of data (C_1), ii) reasoning on the evolution of multiple data streams (C_2), and iii) scalable and consistent prediction of anomalies (C_3).

The data from different sources (Figure 1) is converted to ontology streams (Figure 2) as discussed earlier. Let $\mathcal{O}^n_m$ represent the journey time and $\mathcal{P}^n_m$ represent the weather information stream from time m to n. $\mathcal{O}^n_m(i)$ is a snapshot of stream $\mathcal{O}^n_m$ at time $i \in [m, n]$. Figure 4 shows the three challenges in predicting journey time using weather information stream. It also captures the weather records and travel conditions on Dame Street at times i, j.

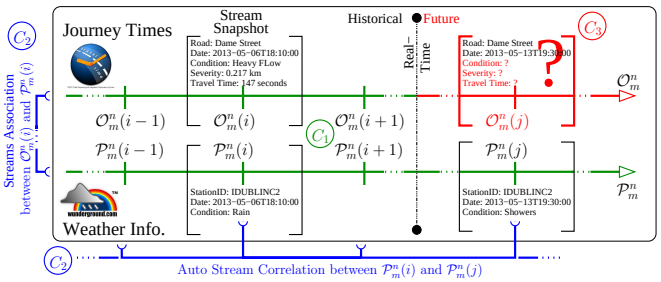
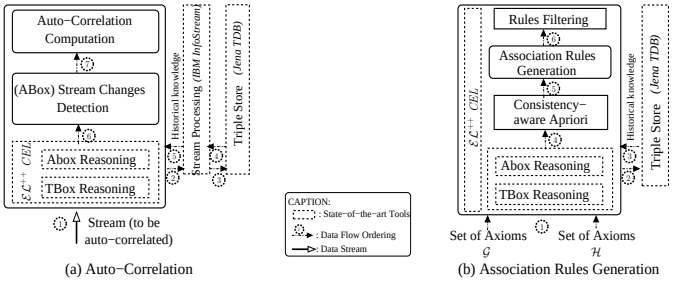


Fig. 4. The challenges (C_1, C_2, C_3) in predicting the journey time $\mathcal{O}^n_m$ using weather stream $\mathcal{P}^n_m$

The second challenge (C_2) is to capture the changes and associate knowledge across the ontology streams. The detection of change along a stream over time enables the computation of knowledge auto-correlation. The semantic similarity between ontology streams is represented by auto-correlation and association aims at deriving rules across streams. These two steps are required to predict the severity of traffic congestions. Prior

to performing the tasks of auto-correlation and knowledge association, it is important to classify the ontology stream. TBox has static knowledge and does not change over time. TBox is generally small and can be classified using the $\mathcal{EL}^{++}$ classification rules [19]. The ontology stream, which is generated from the data stream (Figure 2) consists of ABox axioms. These axioms are internalized into TBox axioms so that the same classification rules from [19] can be applied on them. If existing reasoners (such as CEL[7], ELK[8], Pellet[9] etc) are overwhelmed by the ontology streams, then as discussed earlier, a distributed reasoner such as DistEL [18] can be used.

The auto-correlation between snapshots of an ontology stream is established by comparing the changes in the ABox axioms of the snapshots. The changes can be categorized into three: new, obsolete and invariant. The type of change can have either a positive or a negative influence on the auto-correlation. Invariants have a positive influence on auto-correlation, whereas, new and obsolete changes impact the auto-correlation negatively. Inconsistencies among the snapshots also have a negative correlation. This approach is shown in Figure 5a.



(a) Auto−Correlation (b) Association Rules Generation

Fig. 5. Auto-correlation among the snapshots of an ontology stream and generation of association rules for prediction

The association rules between the snapshots of a stream are encoded using SWRL[10]. For example, a rule that "*the traffic flow of road r_1 is heavy if r_1 is adjacent to a road r_2 where an accident occurs and the humidity is optimum*" can be represented in SWRL as

$$\text{HeavyTrafficFlow(s)} \leftarrow \text{Road}(r_1) \wedge \text{Road}(r_2) \wedge \text{isAdjacentTo}(r_1, r_2) \wedge \text{hasTravelTimeStatus}(r_1, s) \wedge \text{hasWeatherPhenomenon}(r_1, w) \wedge \text{OptimumHumidity}(w) \wedge \text{hasTrafficPhenomenon}(r_2, a) \wedge \text{RoadTrafficAccident}(a)$$

The generation of association rules is based on a description logic extension of Apriori [22] where subsumption (sub-concept relation) is used to determine association rules. Association is achieved between any ABox elements together with their entailments (e.g., all congested roads, weather, works, incidents, city events, delayed buses). Association is possible only in the case where elements appear in at least one snapshot of the stream. As the number of ABox elements in the stream increases, the number of rules that get generated grows exponentially. Rules are filtered by adapting the definition of *support* (i.e., number of occurrences that support the elements of the rule) and *confidence* (i.e., probability of finding the consequent of the rule in the streams given the antecedents of the rule) for ontology stream. In addition only consistent associations are considered. This approach is shown in Figure 5b. More details on auto-correlation and generation of association rules, including the algorithms, are available in [23].

Although filtering of rules based on support and confidence addresses the scalability concern, it does not however ensure prediction of facts that are consistent (challenge C_3), i.e., facts that do not contradict future knowledge facts. This can be solved by combining auto-correlation with association rule generation. First step is to identify the context (e.g., mild weather, road closure) where the prediction is required, and then perform its auto-correlation with historical contexts. Rules are generated and filtered based on their support, confidence and consistency. A rule is considered as consistent if the consequent of the rule is consistent with the knowledge captured by the exogenous stream [24]. Rules are contextualized and evaluated only against the auto-correlated stream snapshots. This makes the selection of rules knowledge evolution-aware and ensures that rules are applied to contexts where knowledge does not change drastically. This approach of combining auto-correlation with association rule generation is shown in Figure 6.

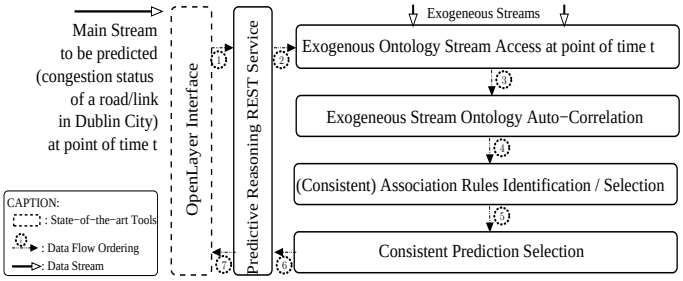


Fig. 6. Auto-correlation is combined with association rule generation for scalable and consistent prediction

V. Lessons Learned

All the features discussed so far, i.e., handling of heterogeneous data, diagnosis and prediction of traffic congestion, have been implemented in a traffic analytics system named STAR-CITY. It makes use of the W3C Semantic Web stack along with other technologies such as i) description logic $\mathcal{EL}^{++}$ based distributed ontology classifier, ii) rule based pattern association, iii) machine learning based entity search, and iv) stream based correlation and inconsistency checking. STAR-CITY was initially deployed in Dublin, Ireland but was later expanded to other cities such as Bologna, Miami and Rio. The challenges and lessons learned in deploying such a system are discussed here.

Heterogeneous streams and semantic expressivity. The format of different data streams (sensors) used in STAR-CITY

[7]https://lat.inf.tu-dresden.de/systems/cel/
[8]https://github.com/liveontologies/elk-reasoner
[9]https://github.com/stardog-union/pellet
[10]https://www.w3.org/Submission/SWRL/

generally remains the same. It is important to pick the right vocabulary and the expressivity to model the data. DBpedia, W3C and NASA ontologies were used to link, integrate and interoperate with all the data sources. However, a custom ontology was developed to model journey time data. Care has to be taken so that terminologies in the various vocabularies used are aligned. This is important in order to achieve on-the-fly integration of all the data sources.

The semantic encoding of city events is in OWL 2 EL profile. This is suitable for us because ontology classification can be decided in polynomial time and hence is scalable. A more expressive profile such as OWL 2 Full or DL could lead to i) more causes getting triggered for road congestion, ii) improving the precision of diagnosis, and iii) improving the scalability and precision of prediction by triggering stronger rules. The downside would be that ontology classification can no longer be done in polynomial time. On the other hand, it would be interesting to check if a profile less expressive than $\mathcal{EL}^{++}$ can be used and still obtain more or less the same precision in diagnosis and prediction.

Scalability of the semantic database. Jena TDB is used to store the semantically enriched data in STAR-CITY. But it could not handle simultaneous updates from multiple streams. So some of the ontology streams had to be delayed in order to accommodate this shortcoming. The B+Trees indexing structure of Jena TDB scales the best in our stream context where large number of updates are performed, i.e., the transaction model is much better handled by this data structure. However there were some scalability issues to handle historical data over more than approximately 110 days. If we do not place any restrictions on the number of days to consider for historical data, then there would be 3,800,000+ events in 458 days. Data gets updated every 20 seconds in this case. In the case of buses, this number is 1000 times larger. Jena TDB cannot handle such large amount of data. Topics such as data, knowledge summarization and stream synchronization needs to be looked into so that the amount of data to be handled by Jena TDB reduces.

Noisy sensor data. Sensors in the real-world exhibit noise. They do not observe the world perfectly due to a number of reasons such as malfunctioning, mis-calibration or network issues. Such noisy data should be detected early so as to avoid unnecessary computations and inaccurate diagnosis, prediction results. In STAR-CITY, some custom filter operators are used to check the validity of the data. These filter operators are defined by analyzing the historical data. For all the data from different sources, the minimum and maximum values are computed. Any record in the data stream that strongly deviates from this interval are removed. If a new data stream is to be considered for traffic analytics, then its historical data needs to be analyzed to determine the appropriate filters. Other mechanisms to filter noisy data should also be looked into.

Temporal reasoning. W3C Time ontology was used to represent the starting data/time and the duration of each snapshot. The temporal similarity between the snapshots of an ontology stream is strictly based on the time intervals. In other words, only the city events and anomalies that match this timer interval are considered. In order to capture more generic temporal aspects such as anomalies during rush hours, bank holidays, weekend, some refinements to the existing ontology are required. Complex features such as temporal intervals could have been used but this could affect the scalability of the application over time. So only basic temporal features were considered. However, more accurate and complext temporal operators could be considered by taking into account the research challenges discussed in [25].

VI. Conclusion

We presented a traffic analytics system named STAR-CITY that can i) handle heterogeneous streaming data from multiple sources, ii) diagnose anomalies such as traffic congestion, and iii) forecast traffic congestions. Heterogeneous data is converted into a homogeneous semantic representation using Semantic Web technologies such as OWL and RDF. In order diagnose traffic congestions, historical data along with other relevant data such as weather information, road works, city events are considered. Concept abduction is used to compare the current event with the historical event and generate a diagnosis report. Forecasting a traffic congestion involves tracking the changes and associating knowledge in the form of rules across the snapshots in an ontology stream. Filtering of rules to avoid rule explosion and consistency in predicting the facts are also discussed. Finally, the lessons learned and the challenges involved in building a scalable traffic analytic system are highlighted.

STAR-CITY supports city managers in understanding the effects of city events, weather conditions and historical data on traffic conditions in order to take corrective actions. It provides valuable insights into real-time traffic conditions making it easier to manage road traffic which in turn helps in efficient urban planning. STAR-CITY has been successfully deployed in some of the major cities such as Dublin (Ireland), Bologna (Italy), Miami (USA) and Rio (Brazil).

References

[1] D. Schrank, B. Eisele, and T. Lomax, "TTI's 2012 urban mobility report," USA, 2012.

[2] R. Arnott and K. Small, "The Economics of Traffic Congestion," *American scientist*, vol. 82, no. 5, pp. 446–455, 1994.

[3] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, "Dynamical model of traffic congestion and numerical simulation," *Physical Review E*, vol. 51, pp. 1035–1042, February 1995.

[4] T. Nadeem, S. Dashtinezhad, C. Liao, and L. Iftode, "TrafficView: Traffic Data Dissemination Using Car-to-car Communication," *SIGMOBILE Mobile Computing Communications Review*, vol. 8, no. 3, pp. 6–19, July 2004.

[5] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, Eds., *OWL 2 Web Ontology Language: Primer*. W3C Recommendation, 11 December 2012, available from http://www.w3.org/TR/owl2-primer/.

[6] G. Schreiber and Y. Raimond, Eds., *RDF 1.1 Primer*. W3C Recommendation, 24 June 2014, available from https://www.w3.org/TR/rdf11-primer/.

[7] F. Baader, I. Horrocks, and U. Sattler, "Description Logics," in *Handbook of Knowledge Representation*, F. van Harmelen, V. Lifschitz, and B. Porter, Eds. Elsevier, 2008, ch. 3, pp. 135–180.

[8] M. Krötzsch, F. Simančík, and I. Horrocks, "Description Logics," *IEEE Intelligent Systems*, vol. 29, pp. 12–19, 2014.

[9] P. Hitzler, M. Krötzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2010.

[10] F. Lécué, R. Tucker, S. Tallevi-Diotallevi, R. Nair, Y. Gkoufas, G. Liguori, M. Borioni, A. Rademaker, and L. Barbosa, "Semantic Traffic Diagnosis with STAR-CITY: Architecture and Lessons Learned from Deployment in Dublin, Bologna, Miami and Rio," in *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*, 2014, pp. 292–307.

[11] S. Kotoulas, V. López, R. Lloyd, M. L. Sbodio, F. Lécué, M. Stephenson, E. M. Daly, V. Bicer, A. Gkoulalas-Divanis, G. D. Lorenzo, A. Schumann, and P. M. Aonghusa, "SPUD - semantic processing of urban data," *Journal of Web Semantics*, vol. 24, pp. 11–17, 2014.

[12] C. Haase and C. Lutz, "Complexity of Subsumption in the [Escr][Lscr] Family of Description Logics: Acyclic and Cyclic TBoxes," in *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, 2008, pp. 25–29.

[13] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov, O. Verscheure, H. N. Koutsopoulos, and C. Moran, "IBM Infosphere Streams for Scalable, Real-time, Intelligent Transportation Services," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, 2010, pp. 1093–1104.

[14] Y. Ma, T. Tran, and V. Bicer, "TYPifier: Inferring the Type Semantics of Structured Data," in *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, 2013, pp. 206–217.

[15] F. Lécué, A. Schumann, and M. L. Sbodio, "Applying Semantic Web Technologies for Diagnosing Road Traffic Congestions," in *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II*, 2012, pp. 114–130.

[16] T. D. Noia, E. D. Sciascio, F. M. Donini, and M. Mongiello, "Abductive Matchmaking using Description Logics," in *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, 2003, pp. 337–342.

[17] K. Dentler, R. Cornet, A. ten Teije, and N. de Keizer, "Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile," *Semantic Web*, vol. 2, no. 2, pp. 71–87, 2011.

[18] R. Mutharaju, P. Hitzler, P. Mateti, and F. Lécué, "Distributed and Scalable OWL EL Reasoning," in *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, 2015, pp. 88–103.

[19] F. Baader, S. Brandt, and C. Lutz, "Pushing the EL Envelope," in *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, 2005, pp. 364–369.

[20] R. Mutharaju, "Distributed Rule-Based Ontology Reasoning," Ph.D. dissertation, Wright State University, 2016.

[21] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi, and M. L. Sbodio, "Predicting Severity of Road Traffic Congestion using Semantic Web Technologies." in *Proceedings of the 11th Extended Semantic Web Conference (ESWC2014), Anissaras, Crete, Greece, May 25–May 29, 2014*, 2014.

[22] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94, San Francisco, CA, USA, 1994, pp. 487–499.

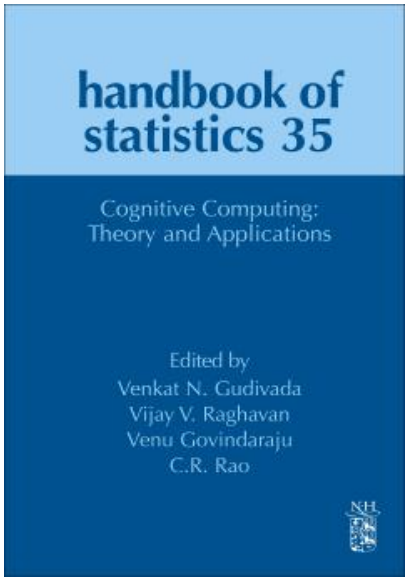
[23] F. Lécué and J. Z. Pan, "Predicting Knowledge in an Ontology Stream," in *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013, pp. 2662–2669.

[24] J. Wu and F. Lécué, "Towards Consistency Checking over Evolving Ontologies," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, 2014, pp. 909–918.

[25] C. Lutz, "Combining interval-based temporal reasoning with general TBoxes," *Artificial Intelligence*, vol. 152, no. 2, pp. 235–274, 2004.

Cognitive Computing: Theory and Applications

BY GUDIVADA,V.N., RAGHAVAN,V.V., GOVINDARAJU, V., RAO, C.R. (EDITORS) - ISBN : 978-0-4446-3744-4



REVIEWED BY PAWAN LINGRAS

A BOOK FOR PRACTITIONERS, POTENTIAL ADAPTERS, AND STUDENTS OF DATA SCIENCE

While the book title is rightfully cognitive computing, I found the book to be as much about data science. It will be useful for current and budding data scientists who are looking to use cognitive computing in their analytics. The book will be an excellent textbook for a first course in a graduate data science program or can introduce data science to senior (fourth-year) undergraduate program. This handbook may in fact serve as a prequel to a previous handbook by these authors called Big Data Analytics. The Big Data Analytics handbook was focused on the issues related to big data. This new handbook focuses more on providing a true understanding of the cognitive computing techniques that are used with datasets of any size- big or small. Since understanding cognitive computing technology is almost a prerequisite to studying big data, students using Cognitive Computing Handbook as a textbook may also find the Big Data Analytics Handbook as an excellent reference for more advanced projects.

The book is divided into three sections. The first section focuses on principles consisting of two chapters, followed by extensive studies of machine learning techniques, and ends with the third section consisting of case studies. I believe one can create a very good course which follows the book very closely. There may not be enough time in a course to go through every detail. However, the chapters are written in an accessible manner that allows students to learn the fundamental principles behind many of the techniques. There is enough mathematics in all the chapters to provide more precise understanding of the topics without overwhelming the reader with symbols and equations. Every chapter provides sufficient list of references that can be used by students to learn the details.

Researchers who are not familiar with cognitive computing, but are looking to apply some of these techniques to their application domains will find this book to be even more useful. There is a certain amount of hype about various new concepts in cognitive computing such as deep learning, random forest, or MapReduce. It is not always obvious if these new and sophisticated techniques are needed in your application. The authors cut through the hype and explain how these concepts were developed from their more fundamental origins and the advantages of using one over the other such as “deep learning versus neural networks” or “decision trees versus random forest”. In this respect, the handbook is also useful for cognitive computing researchers. They cannot always keep up with all the new development. However, they are usually familiar with the fundamental techniques. They can easily understand the advantages of the new techniques over the fundamental building blocks that were used to derive them. While there is a separate application section, all the chapters are tethered to an application through simple examples. The application section provides end-to-end description of case studies. Practitioners can pick a case study that is closest to their interest and adapt it to their own application.

Similar to a previous handbook (Big Data Analytics) by these authors the writing of chapters is fairly consistent, so the readers do not have to adapt to an entirely different writing style. On the other hand, most of the chapters are more or less self-contained. If readers find topics in a particular chapter of a particular interest, they can pretty much start reading from that chapter itself. They do not have to go to a previous chapter to get background information. The consistent treatment has its limits in a multi-authored handbook. One can see different level of details and somewhat different description of certain concepts such as say deep learning. One can even call this a feature. We get different perspectives on the topic in the same handbook.

THE BOOK:

GUDIVADA,V.N., RAGHAVAN,V.V., GOVINDARAJU, V., RAO, C.R. (EDS) (2016), COGNITIVE COMPUTING: THEORY AND APPLICATIONS, 404 P.
ELSEVIER
PRINT BOOK ISBN : 9780444637444

ABOUT THE REVIEWER:

PAWAN LINGRAS
Professor and Director, Computing and Data Analytics, Saint Mary's University
Halifax, Nova Scotia, B3H3C3, Canada
Contact him at: pawan@cs.smu.ca

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

WI 2017
The 2017 IEEE/WIC/ACM International Conference on Web Intelligence
Leipzig, Germany
August 23-26, 2017
http://webintelligence2017.com/

Web Intelligence (WI) aims to achieve a multi-disciplinary balance between research advances in theories and methods usually associated with Collective Intelligence, Data Science, Human-Centric Computing, Knowledge Management, and Network Science. It is committed to addressing research that both deepen the understanding of computational, logical, cognitive, physical, and social foundations of the future Web, and enable the development and application of technologies based on Web intelligence.

In addition to the research track, WI'17 comprises special sessions and workshops as well as tutorials and a PhD mentoring session. Moreover, industry papers and demo proposals can be submitted to WI'17 and will be dealt with by a special PC for industrial papers which will apply industry-compliant assessment criteria.

Web Intelligence focuses on scientific research and applications by jointly using Artificial Intelligence (AI) (e.g., knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence) and advanced Information Technology (IT) (e.g., wireless networks, ubiquitous devices, social networks, semantic Web, wisdom Web, and data/knowledge grids) for the next generation of Web-empowered products, systems, services, and activities.

WI'17 welcomes both research and application papers submissions. All submitted papers will be reviewed on the basis of technical quality, relevance, significance and clarity. Accepted full papers will be included in the proceedings published by IEEE Computer Society Press.

ICDM 2017
The Twenty-Fourth IEEE International Conference on Data Mining
New Orleans, LA, USA
November 18-21, 2017

The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications. ICDM draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing. By promoting novel, high quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state-of-the-art in data mining. Besides the technical program, the conference features workshops, tutorials, panels and, since 2007, the ICDM data mining contest.

ICHI 2017
The IEEE International Conference on Healthcare Informatics
Park City, Utah, USA
August 23-26, 2017
http://ichi2017.ce.byu.edu/

ICHI'17 is the premier community forum concerned with the application of computer science principles, information science principles, information technology, and communication technology to address problems in healthcare, public health, and everyday wellness. The conference highlights the most novel technical contributions in computing-oriented health informatics and the related social and ethical implications.

ICHI'17 will be held in Park City, Utah, USA on August 23–26, 2017. It will be a forum for demo and paper contributions from researchers, practitioners, developers, and users to explore and disseminate cutting-edge ideas and results, and to exchange techniques, tools, and experiences.

Related Conferences

AAMAS 2017
The 16th International Conference on Autonomous Agents and Multi-Agent Systems
Sao Paulo, Brazil
May 8-12, 2017
http://www.aamas2017.org/

AAMAS is the leading scientific conference for research in autonomous agents and multiagent systems. The AAMAS conference series was initiated in 2002 by merging three highly respected meetings: the International Conference on Multi-Agent Systems (ICMAS); the International Workshop on Agent Theories, Architectures, and Languages (ATAL); and the International Conference on Autonomous Agents (AA).

Subsequent AAMAS conferences have been held in Melbourne, Australia (July 2003), New York City, NY, USA (July 2004), Utrecht, The Netherlands (July 2005), Hakodate, Japan (May 2006), Honolulu, Hawaii, USA (May 2007), Estoril, Portugal (May 2008), Budapest, Hungary (May 2009), Toronto, Canada (May 2010), Taipei, Taiwan (May 2011), Valencia, Spain (June 2012), Minnesota, USA (May 2013), Paris, France (May 2014) and Istanbul, Turkey (May 2015), Singapore (May 2016).

For the first time in South America, the 16th. edition of AAMAS will be held in May 2017 in São Paulo, Brazil.

AAMAS is the largest and most influential conference in the area of agents and multi-agent systems. The aim of the conference is to bring together researchers and practitioners in all areas of agent technology and to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multi-agent systems.

AAMAS is the flagship conference of the non-profit International Foundation for Autonomous Agents and Multi-agent Systems (IFAAMAS).

AAAI 2017
The 31st AAAI Conference on Artificial Intelligence
San Francisco, California, USA
February 4-9, 2017
http://www.aaai.org/Conferences/AAAI/aaai17

The Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17) will be held February 4–9 at the Hilton San Francisco, San Francisco, California, USA. The workshop, tutorial, and doctoral consortium programs will be held Saturday and Sunday, February 4 and 5, followed by the technical program, Monday through Thursday, February 6–9.

The chairs of AAAI'17 are Satinder Singh (University of Michigan) and Shaul Markovitch (Technion-Israel Institute of Technology).

The purpose of the AAAI conference is to promote research in artificial intelligence (AI) and scientific exchange among AI researchers, practitioners, scientists, and engineers in affiliated disciplines. AAAI'17 will have a diverse technical track, student abstracts, poster sessions, invited speakers, tutorials, workshops, and exhibit and competition programs, all selected according to the highest reviewing standards. AAAI'17 welcomes submissions on mainstream AI topics as well as novel crosscutting work in related areas.

SDM 2017
The 2017 SIAM International Conference on Data Mining
Houston, Texas, USA
April 27 - 29, 2017
http://www.siam.org/meetings/sdm17/

Data mining is the computational process for discovering valuable knowledge from data. It has enormous application in numerous fields, including science, engineering, healthcare, business, and medicine. Typical datasets in these fields are large, complex, and often noisy. Extracting knowledge from these datasets requires the use of sophisticated, high-performance, and principled analysis techniques and algorithms, which are based on sound theoretical and statistical foundations. These techniques in turn require implementations on high performance computational infrastructure that are carefully tuned for performance. Powerful visualization technologies along with effective user interfaces are also essential to make data mining tools appealing to researchers, analysts, and application developers from different disciplines.

The SDM conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students and others new to the field to learn about cutting-edge research by hearing outstanding invited speakers and attending presentations and tutorials (included with conference registration). A set of focused workshops is also held on the last day of the conference. The proceedings of the conference are published in archival form, and are also made available on the SIAM web site.

IJCAI 2017
The 26th International Joint Conference on Artificial Intelligence
Melbourne, Australia
August 19-25, 2017
http://ijcai-17.org/

IJCAI is the International Joint Conference on Artificial Intelligence, the main international gathering of researchers in AI. Held biennially in odd-numbered years since 1969, IJCAI is sponsored jointly by IJCAI and the national AI societie(s) of the host nation(s). IJCAI is a not-for-profit scientific and educational organization incorporated in California. Its major objective is dissemination of information and cutting-edge research on Artificial Intelligence through its Conferences, Proceedings and other educational materials.

Starting with 2016, IJCAI will be held annually. IJCAI is sponsored jointly by IJCAI and the national AI societie(s) of the host nation(s). The 26th International Joint Conference on Artificial Intelligence will be held in Melbourne, Australia in August 2017.

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398