RAMFIS: Integrating Diverse TA1's

TAC 2019

Prepared by Cecilia Mauceri[1], Shafiuddin Rehan Ahmed[1], Timothy O'Gorman[1], Chris Koski[1], Peter Anick[2], David White[3] and Martha Palmer[1]

[1]University of Colorado Boulder, [2]Brandeis University, and [3]Colorado State University

Section 1: Introduction

TA1 performers extract entities and events from individual multi-media documents and pass them along to a TA2 as knowledge elements usually with reference links to a Knowledge Base (KB). As TA2 performers, our directive is to determine where entities and events in one

document are identical with entities and events in another document, and cluster them together around a single KB link. This facilitates the detection of contradictions and confirmations. We are a stand alone TA2, so can choose which TA1 output we want to focus on. We felt responsible for processing the output of stand alone TA1's, such as JHU and UMich, since they didn't already have dedicated TA2 partners, and were also curious about how the other TA1's, GAIA-ISI, Opera-CMU and BBN would fare. For the TAC 2019 evaluation, we decided to process as many TA1's as possible, and to also focus on merging the data from one TA1 with another TA1. Our primary goal is always cross-document and cross-TA1 co-reference linking, as shown in Figure 1. We made the assumption that a TA1 would provide KB links as well as within document coreference links and knowledge elements. That turned out to be a false assumption. However, merging 1) a TA1 without a KB with 2) a TA1 with a KB actually turned out to be an effective way of assigning KB links to the first TA1. In this document, we go into detail with respect to our architecture, our linking procedure, and the challenges we faced during the evaluation, especially with respect to dealing with multiple TA1's.

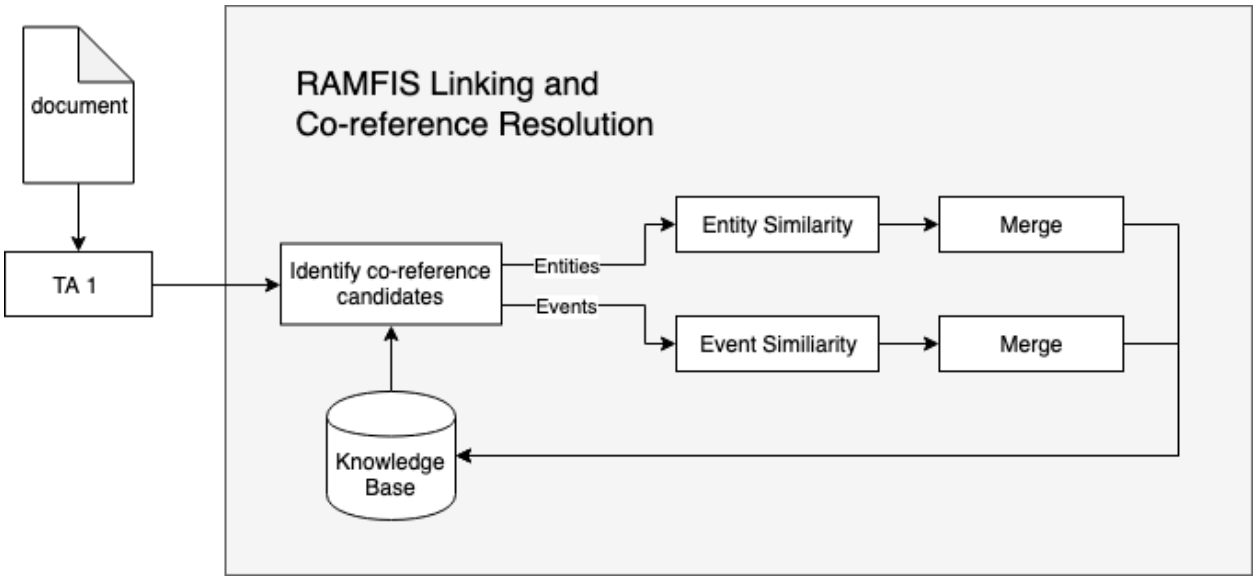

Figure 1: System Overview. This flow chart illustrates the process of adding TA1 output to the knowledge base and performing cross-document co-reference linking.

Section 2 : System Architecture

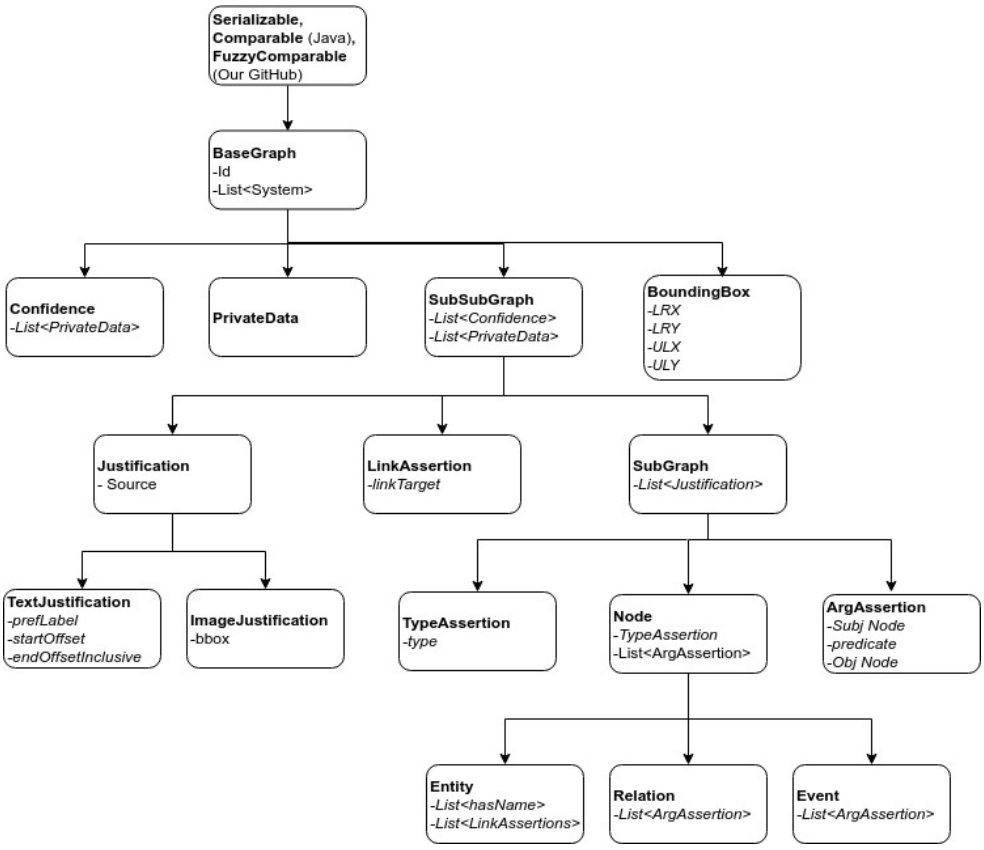

Figure 2: Class Diagram

In addition to processing multiple TA1's and merging them if possible, our other goals included constructing compact KB's, a streaming architecture and AIF compliance. By representing knowledge elements as Java objects, we eased the merging step, and hence were able to generate significantly more compact KBs. This approach proved particularly useful in merging multiple TA1 outputs arbitrarily. Since we were rebuilding the entire KB from scratch, it allowed us to put the NIST restrictions for AIF compliance explicitly into our system. Thereby, we could generate a valid KB even if the TA1 output was not entirely valid.

Section 4: Linking

	Entities			Events		
TA 1	Pre-clustering	Post-clustering	Percentage Decrease	Pre-clustering	Post-clustering	Percentage Decrease
BBN_1	270,168	232,785	0.14	107,050	89,836	0.16
GAIA_1	452,436	309,358	0.32	37,533	31,151	0.17
GAIA_2	459,044	310,437	0.32	34,127	23,743	**0.30**
OPERA_aditi_V2	339,889	224,776	**0.34**	13,034	11,068	0.15
GAIA_1 + Michigan_1		371,636			23,816	
GAIA_1 + OPERA_hans_V3		458,931			36,800	
GAIA_1 + JHU_5	758,978	690,166	*0.09*	85,393	75,820	*0.11*

Table 2: Month 19 Submission Linking Results. TA1 input with the largest proportion of identified links are highlighted in **bold**. TA1 input with smallest proportion of identified links are highlighted in *italics*.

The approach for entity and event linking has been substantially revised from the previous evaluation. Significant improvements were made in the software front end of the pipeline to perform linking in a streaming fashion. We use three methods for entity linking: Linking based on justifications, Linking based on reference KB, and Linking based on string matching. These linking methods are performed sequentially.

Linking based on justifications is used to merge KBs for the same source document among different TA1s. We make the assumption that each TA1 identifies all within document co-reference links in their own KB. Therefore, we only need to compare entities within the same source document if the entities are extracted by different TA1s. If two entities of the same type from different TA1s share a justification, we consider them to be the same entity. We consider two justifications to be the same if the document offsets, in the case of text justifications, or bounding box, in the case of image justifications, have at least 0.8 intersection over union score.

Some of the TA1s provide reference KB links to some of the entities in their output. We use this information and the confidence values they provide for linking. Reference KB links are unique. Therefore, a strongly confident reference KB link is a reliable linking signal. Conversely, if two entities have different reference KB links, they should not be merged. Some of the TA1s didn't

provide reference KB links. For those TA1's, we can use other entity-linking methods to provide cross-document and inter-TA1 links. If we discover an inter-TA1 link to another TA1 that provided reference KB links, the entity will gain those links during the merge step.

Our final entity-link prediction method is based on string matching. String matching is only performed for named entities. To avoid pairwise matching on the full set of named entities, we first bin the entities by their ontology type and the first three characters of each token in their name. For example, "Donald Trump" would end up in the "PER-Don" and "PER-Tru" bins while "Trump Tower" would end up in the "LOC-Tru" and "LOC-Tow" bins. Within each bin, we score all pairs of entities on name similarity. We made some improvements to the string matching method from the previous evaluation by writing rules for names of person, abbreviations, etc. This was a very noisy data cleaning process with extensive manual tuning, and we are considering smarter methods for the next evaluation.

Event linking follows entity linking. We pick out possible candidate event clusters from the pool of clusters generated during the pipeline. The candidates have to share at least one coreferent argument with the same role label with the target event node. Then we take into account the LDCTimeComponent to rank the candidates based on a heuristic that calculates the concurrency of two events. The candidates occurring in the same timeframe as the target will be ranked higher.

Table 1 shows the number of entity and event links that our system identifies for different TA1 input. The number of co-reference links identified varies largely over TA1s, ranging between 10% and 30% of the entities and events. This is expected as the linking success is largely dependent on the quality and quantity of the reference KB links, names, and event arguments provided by the TA1.

Section 6: Baseline scores on annotated datasets

To evaluate the coreferencing capability of our system we evaluate the clustering scores using the cross-document annotations of Event Coref Bank (ECB) corpus. First, we ran the knowledge extraction system provided by the RPI (now UIUC) Blender lab on the raw text documents of ECB. Then, we ran our system on the generated knowledge graphs to perform the clustering on events and entities found in the extraction process. Finally, we compare the clusters generated by our system against the annotated coreference chains and calculate B-cubed and MUC scores to capture the purity of the clusters. Note that we only compare the results for the entities and events from the annotations that were detected by the TA1 system.

	Gold	TA1	Common	B^3 P	B^3 R	B^3 F1	MUC-P	MUC-R	MUC-F1

Events	3437	5107	918	95.92	42.75	59.14	63.04	10.96	18.67
Entities	4268	8820	864	98.09	64.33	77.7	95.08	54.2	69.04
Both	7705	13927	1782	95.75	57.05	71.5	54.71	10.96	18.26

Table 2: Event Coref Bank (ECB) Scores for Common Nodes.

	Gold	**TA1**	**Common**	**Precision**	**Recall**	**F1**
Events	3818	703	590	80.11	14.14	24.05
Entities	4411	3534	2238	46.45	49.55	47.95
Both	8229	4237	2828	83.97	30.83	45.11

Table 3: DEFT Richer Event Descriptions (RED) BCUB Score.

The observation from these scores is that the system precisely detect the clusters. However, it can do better in finding more clusters. Our current and future work mainly focus towards achieving this goal.

Section 7: Pretty Printer and Qualitative Insights

Another important aspect of assessing our TA2 output is error analysis. At **Brandeis**, Peter Anick continued his work on the acquisition of training data to assist TA3 detection of alternative interpretations of events from the TA1/2 output. Using events with two or more role fillers extracted from the M9 data, he generated a set of questions for TA3 to use in testing the ability to identify role fillers in partially specified events. He also manually identified cases of identical events with different role fillers, based on differing accounts/points of view of the event. Peter then designed and implemented a "pretty printer" to extract *event; role-filler; document-location; name* information from the TA2 test graph data and present it in a human readable form. He used the pretty-printed output to identify possible errors in TA2 cross-document coreference decisions. These were classified into a small number of categories and remedies were discussed with the team in preparation for final evaluation runs. This has now all been set up as a Web Interface that can be probed with the questions. Based on findings from the Web

interface probes, we detected a bug in the Relation code that allowed multiple alternative fillers for a single Relation slot, and were able to fix it. Rerunning improved Relation detection.on merged GAIA-Opera TA1's.

Section 8: Challenges and Workarounds

In this section we discuss particularly problematic issues that had to be addressed.

1. **Validation edge cases** We had a single type of validation error for our KBs reported by the validator. The error being: "Each Cluster, Entity, Event, or Relation can specify up to one informative mention per document as long as each informative mention points to a different sourceDocument". This error was hard to debug because we failed to reproduce it on a smaller dataset (nearly 500 documents).
2. **Scaling up** Due to the volume of information we received from each TA1, as well as the increase in resources needed to process data from merged TA1s, we implemented a variety of solutions to offset the scale of the task. We improved our throughput at a high level by running merging and clustering simultaneously on different TA1's across 3 high-powered systems, both on and off-site, which was a necessary step due to the number of late TA1 corrections that needed processing. Fine-grained ontological representations allowed us to optimize individual and merged TA1s by reducing the number of comparisons per element, as we were able to restrict comparisons by type. Merged TA1s underwent an additional processing step before the standard clustering, allowing us to merge at the document level with significantly more reliable within-document coreference resolution and justification comparison.
3. **LDC time for events** The LDC time representation format involves a start and end date, with year, month, and day listed for each. While day is the maximum level of precision under this format, many event representations we received from TA1s had no day information, while some only had a year. Furthermore, the vast majority of events in the scenario do not take place over multiple days, making range representations more difficult to work with. We worked around these difficulties in time comparison by converting all times to epoch representations, and measuring the proportion of overlap in the range. Where days are not specified, the range is represented as from the beginning of the starting month to the end of the ending month. If only a year was given, the metric was ignored, as we determined that we could not extract useful information for linking at that level of uncertainty.

Section 9: Current Progress and Future Plans

Summary of Accomplishments following the M19 evaluation and our September virtual site visit:

1) Extensive, painstaking mapping of AIDA Program Ontology to LDC annotation ontology was performed, and negotiation with LDC over resolving remaining discrepancies is underway, based on input from other AIDA performers.
2) We have added a graph embedding component to cross-document coref, focused on improving recall for events for the ECB corpus. Recall improved from 10.2% to 37.8% when evaluated on MUC with a corresponding F1 score improvement of 17.86 to 42.5%. However, on BCUB there is a recall improvement from 42.53 to 48.54, but a corresponding drop in precision lowers the overall F1 score. From the results of the combined system, we can conclude that graph embeddings find more clusters than our TA2 system alone, and our TA2 system helps correct obvious mistakes made by the graph embedding clustering approach (for example, clustering events of different types). Experiments are continuing, and new similarity measures are being introduced, some based on vector representations.

Method	BCUB Recall	BCUB Precision	BCUB F1	MUC Recall	MUC Precision	MUC F1
TA2 system only	(377 / 886) 42.53%	(852.8 / 886) 96.25%	58.99%	(54 / 529) 10.2%	(54 / 86) 62.79%	17.56%
Graph Embeddings only	(548 / 886) 61.83%	(390 / 886) 44%	51.41%	(270 / 529) 51.03%	(270 / 512) 52.73%	**51.87%**
Graph Embeddings + TA2 system	(430 / 886) 48.54%	(550 / 886) 62.08%	**54.48%**	(200 / 529) 37.8%	(200 / 412) 48.5%	**42.5%**

3) We delivered a Pretty Print Web Interface for probing TA2 output. We are now in the process of developing a pipeline for the addition of new databases.
4) We have applied affine mapping to image vectors (White, et. al., 2019) from different TA1's (BBN: generated from Facenet trained on CASIA-WebFace; and Columbia: generated from FaceNet trained on VGGFace2) and find the vectors can be correctly mapped with 99% accuracy.
5) Work has started on porting the affine mapping approach to language vectors and multimodal vectors. We are in contact with TA1's about additional future use of feature vectors.
6) We have agreed to begin collaborating with the GAIA team to explore merging our TA2 components with their TA2 components, ensuring a common format.

References

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, pages 563–566,1998.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston,and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahra-mani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 2787–2795. Curran Associates, Inc., 2013.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In Proceedings of the 9th international conference on Language Resources and Evaluation (LREC2014)

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. "Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation." In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), pp. 47-56. 2016.

David G. McNeely-White, J. Ross Beveridge, and Bruce A. Draper, Inception & ResNet: Same Training, Same Features, *2019 Annual International Conference on Biologically Inspired Cognitive Architectures,* the 10th Annual Meeting of BICA Society: Seattle, WA, USA, Springer-Verlag, A. Samsonovich (ed).