Last Words

What Can Be Accomplished with the State of the Art in Information Extraction? A Personal View

Ralph Weischedel
University of Southern California
Information Sciences Institute
weisched@isi.edu

Elizabeth Boschee
University of Southern California
Information Sciences Institute
boschee@isi.edu

Though information extraction (IE) research has more than a 25-year history, F_1 scores remain low. Thus, one could question continued investment in IE research. In this article, we present three applications where information extraction of entities, relations, and/or events has been used, and note the common features that seem to have led to success. We also identify key research challenges whose solution seems essential for broader successes. Because a few practical deployments already exist and because breakthroughs on particular challenges would greatly broaden the technology's deployment, further R&D investments are justified.

1. Introduction

The roots of information extraction (IE) research date back at least 25 years; it seems appropriate to assess now whether the technology has led to utility and what its future might be. In this section, we provide a whirlwind tour of past evaluations to set the stage for later sections that provide brief case studies of deployments of IE technology. In this article we will focus on properties of three end-to-end systems that map text to semantic structures, and, in particular, on the attributes of the applications that contributed to utility despite low F_1 scores in formal evaluations of research.

Early significant milestones appeared through the Message Understanding Conferences (MUCs), starting with MUC-3[1] (1991) and ending with MUC-7[2] (1997). In part through MUCs, named entity recognition (NER) and within-document coreference emerged as evaluation tasks with specifications, training and test data, and a scorer. This was no small accomplishment, because most prior work in natural language processing had not measured progress so rigorously on substantial blind test data. Also, MUC provided a first step in mapping unstructured text into a semantically motivated

1 *Proceedings of the 3rd Conference on Message Understanding*. 1991.
2 http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

doi:10.1162/coli_a_00331

Message ID: 123456789	
Date:	04 May 1989
Type:	BOMBING
Category:	TERRORIST ATTACK
Individual Perpetrator:	"3 armed men"
Organizational Perpetrator:	
Perpetrator confidence:	
Physical Target:	COMMERCIAL: "market"
Physical Target #:	1
Human Target:	"shoppers"
Human Target #:	10
Human Target Types:	CIVILIAN: "shoppers"
Target Nation:	
Instrument Type:	
Location:	COLOMBIA: MEDELLIN (CITY)
Effect on Physical Target:	SOME DAMAGE:
Effect on Human Target:	INJURY: "shoppers"

Figure 1
An example of a MUC-3 template. Note the mix of normalized data (e.g., Type and Category) and text strings (e.g., Individual Perpetrator).

structure: a template with a type and slots filled by either text strings or by a small closed set of values. An example template for MUC-3 and MUC-4 appears in Figure 1.

Since MUC-7, a number of commercial products for NER have become available for languages with a commercial market; open source packages for NER allow application to any language where sufficient training data is available. Although extraction from English text was part of every MUC, some MUCs also included evaluation on Japanese text sources. After MUC, a key change instituted through a series of Automatic Content Extraction (ACE) evaluations (1999–2008) was replacing MUC templates (which had many slots filled by arbitrary strings from the source text) with a simple taxonomy of entities, relations, and events. A further accomplishment was defining a task for within-document entity resolution. Additionally, ACE addressed evaluations for Arabic and Chinese sources, and for event extraction. The task definitions and data sets today still empower some research.

From 2012 to 2017, the National Institute of Standards and Technology's (NIST's) Text Analysis Conference (TAC) advanced the field of information extraction through evaluations. Though coreference in general remains a significant research challenge, Entity Linking and Discovery (EDL), a task tackling some coreference challenges, has achieved high scores. In EDL, the goal is to (1) find and link all name and nominal (description) mentions of an entity across a (potentially multilingual) corpus into a single entity representation, and (2) link that entity to a pre-existing knowledge base (KB) if the entity is in the KB, or create a new KB entry for the new entity.

The Coldstart Knowledge Base Population (KBP) task seeks to build a KB from a (potentially multilingual) text collection, given an ontology of entities and relations between them. In a historical sense, KBP seems to be the culmination of the task vision started 20 years earlier in MUC-3, where all entities fit in the ontology as types rather than strings.

Though tasks such as NER and EDL have reached sufficient maturity such that commercial products are feasible, scores[3] of a top-performing system for the full

3 Reporting the micro-averaged scores on hop-0, i.e., correctly finding all relations R and entities b for an entity a, such that [a R b] is in the knowledge base.

knowledge base population task in TAC KBP 2016 seem disappointing (Precision 25.4; Recall 30.2; F_1 25.2).

That is, 75% of the KB triples produced are imperfect and 70% of the relations justified in text are not found. Scores are not the final story regarding whether technology is usable, however. For example, *though the average precision of responses to search engines may appear low, search engines are part of everyday modern life. Though BLEU scores in machine translation seem low, machine translation is offered by Google and other search engines.*

In this article, we review three cases where information extraction with tasks comparable to TAC KBP are in use. In these cases, the following conditions have contributed to today's technology being deployed:

1. Building a knowledge base manually does not provide sufficient coverage of streaming data.
2. Noisy output can be tolerated, thus mitigating the need for very high precision.
3. Redundancy in streaming data can be utilized to overcome recall issues.
4. IE systems can be tuned to either precision or recall, rather than F_1, which can allow for opportunities beyond those that could be achieved if only F_1 (which is maximized by equal recall and precision) were targeted.

2. The Goal, Need, and Opportunity Inherent in Big Data

The challenges and opportunities of Big Data are broad and clearly exhibit property 1. Wikipedia states, "Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, and information privacy."[4] Note how many of those challenges correspond to massive data warehouses, but how few of those challenges address analysis or drawing conclusions.

Though machine learning applied to big data is producing remarkable results, a gap stands between capitalizing on human language content and the structured data needed by analytic tools. For instance, one can draw illuminating social network graphs merely from the metadata of human communications (e.g., who is talking to whom, when, and where), but analysis of what they are communicating is lacking. Though co-occurrence of words/phrases/topics/names in messages can suggest that a relation exists between them, the type of relation is lacking. Though the following statistics are estimates only, the magnitude of public human language communication is staggering:

- 3 million news articles/day[5]
- Over 3 million blog posts/day[6]
- 500 million tweets/day[7]

4 https://en.wikipedia.org/wiki/Big_data.
5 https://blog.twingly.com/2017/09/05/twingly-adds-3-million-news-articles-per-day/.
6 http://www.worldometers.info/blogs/ provides a running total each day.
7 http://www.internetlivestats.com/twitter-statistics/.

Thus, learning the types of relations, the types of events, and so forth. from the content of that human language data is a major challenge to achieve the full potential of learning from big data.

3. Case Study 1: Event Extraction

For several years, political scientists (O'Brien 2010; Montgomery, Ward, and Hollenbach 2011) have developed forecasting models of diverse types of political instability (e.g., international/internal conflict, political unrest, leadership change) based on inputs of triples from the CAMEO ontology (Gerner et al. 2002; Schrodt 2012) of international events. CAMEO defines roughly 300 types of events of international interest, an agent of the event, and an object. For example, for "The U.S. Air Force bombed Taliban camps," the appropriate coded event triple would be (195, USAMIL, AFGINSTAL) or, using a mechanical gloss, (Employ-aerial-weapons, United States military, Taliban). Note that this is in contrast to the goals of *open* information extraction,[8] where no element of a triple need be normalized; for example, the given sentence (The U.S. Air Force, bombed, Taliban camps) is an appropriate open extraction triple.

Forecasting models used in the Worldwide Integrated Crisis Early Warning System (W-ICEWS)[9] produce rolling forecasts for more than 165 countries, 6 months out, forecasting domestic political crises, international crises, ethnic/religious violence, insurgency, and rebellion. See O'Brien (2010) for an early report on how this capability was developed. W-ICEWS models assume CAMEO triples as input. As an alternative to armies of individuals manually creating CAMEO triples from news reports, BBN ACCENT (Boschee, Natarajan, and Weischedel 2012) has been deployed since 2014 to automatically mine news sources and pipe the resulting CAMEO triples directly to the political science models without a human in the loop. For W-ICEWS, ACCENT automatically extracts event data from 50K news reports a week from English sources and from Spanish, Portuguese, and French news via machine translation into English.

Because the models need to aggregate the actors and objects of events in order to notice trends, an extensive "actor dictionary" was manually created independent of ACCENT. ACCENT normalizes the text strings corresponding to the agent/object of the event into an entry of the 50,000 element actor dictionary. Though most of these are named entities (e.g., countries, organizations, and individuals), there are roughly 700 types of described actors. Thus, normalization of event arguments encompasses

- name normalization (e.g., both Abu Du'a and Abu Bakr al-Baghdadi correspond to the same actor)
- description normalization, given a set of features of described actors (e.g., "an Afghan army unit" corresponds to a Military Organization (Afghanistan)). Having a link to a named actor is required for the actor dictionary.

Normalizing the event types and actor types means information from diverse sources can be aggregated to determine trends, such as "Has the number of mentions of Russian government criticisms of Iraq increased or decreased this month?" Normalization is

8 http://www.aclweb.org/anthology/D11-1142.
9 https://www.lockheedmartin.com/us/products/W-ICEWS/iCast.html.

required, for example, to know whether person X represents the Russian government in order to count X's criticism for this question.

Interestingly, solving event coreference is not required for the W-ICEWS application, thereby greatly simplifying the task. Indeed, counting mentions, as in the example of trend detection, is valuable because the number of mentions in the news may be a surrogate for importance for some forecasting models.

Overall accuracy (precision) of ACCENT output has been measured at 76%, including normalizing the agent and object arguments correctly. If one looks at the 19 major aggregated classes of the 294 CAMEO event types, 5 of the 19 event categories are at 80% accuracy or above; 9 range between 70% and 80% accuracy, 4 are slightly below 70% accurate; and 1 is slightly below 60% accurate.

What properties of this application seem to make that level of accuracy appropriate for deployment?

1. The alternative of manually coding events for 50,000 news articles per week from as many as 250 news feeds across four languages is not feasible.
2. The models can accept noise. In fact, they were first deployed using less accurate extractors than BBN ACCENT. However, anecdotally, confidence in the models apparently increased significantly when users saw the high precision of event extraction by ACCENT.
3. It may be that the event extraction task for W-ICEWS is simpler than the task measured in TAC KBP evaluations, for example, CAMEO events may be more lexically triggered than arbitrary event catalogs, and event coreference is neither required nor measured.

4. Case Study 2: Maintaining an Entity-Centric Knowledge Base

The W-ICEWS actor dictionary is a simple example of a large entity-centric knowledge base, that is, a knowledge base storing information about entities (countries, persons, and organizations), their attributes, and relations among them. Manually creating a 50K knowledge base for W-ICEWS was an achievement; maintaining it over time, however, was an ongoing challenge because new entities become prominent (and must be added), attributes/relations of an existing entity change (and should be updated), and the alternate strings for some names seem unending (and should be added). Manual updates had been introduced quarterly, and tend to lag significantly behind real time.

Thus, the second case study is automatic discovery, characterization, and tracking of actors to automatically suggest updates through information extraction. The BBN Automatic World Actor Knowledge Extraction (AWAKE) (Boschee et al. 2014) system was introduced for maintenance of the W-ICEWS actor dictionary at the beginning of 2014; quarterly updates were provided by BBN. AWAKE offers the following three capabilities:

- Automatically creates a knowledge base about persons and organizations from unstructured text.
- Synthesizes information across thousands of mentions of each person/organization to highlight important facts and avoid redundancy.
- Performs corpus-level analysis to resolve/highlight contradictions and estimate confidence for each individual fact. Critical to this application

was AWAKE's confidence estimates, which allow AWAKE discoveries to be triaged into the following classes:

- High confidence entries are added to the actor dictionary automatically.
- Medium confidence entries are reviewed manually.
- Low confidence entries await further text that might confirm or contradict the hypothesized entry.

The factors that led to successful application are:

1. Manually updating the knowledge base was very challenging, thereby encouraging a semi-automated process that reduced human effort.
2. Redundancy in streaming data led to AWAKE's effective confidence estimates on each new assertion/actor.
3. AWAKE's confidence estimates enabled triage for managing effort of a human in the loop, even though the per-item scores in a TAC evaluation might be low.[10]

5. Case Study 3: Aiding Law Enforcement

The third application guides more than 200 law enforcement agencies in detecting human trafficking. The domain is particularly challenging because of the volume of data to be processed, the diversity of Web pages in the wild, the desire to obfuscate (to a degree) messages regarding illicit behavior, and concept drift (Widmer and Kubat 1996), which is evident over time. For instance, Kejriwal and Szekely (2017) cite the following examples of obfuscation and variety:

- *Hey gentleman im neWYOrk and i'm looking for generous...*, which hides *New York* in plain sight.
- *AVAILABLE NOW! ?? - (4 two 4) six 5 two - 0 9 three 1 – 21*, which camouflages the phone number and age.

Several components have been created using techniques developed in information extraction. One component, FlagIt (Flexible and adaptive generation of Indicators from text) (Kejriwal et al. 2017a), uses name extraction and normalization against a database, semi-supervision, and text embedding to tag sentences (currently) with five indicators of human trafficking; additional indicators are being developed. Assigning indicators to Web pages supports both lead generation and lead investigation by law enforcement agencies.

An outline of the processing steps in FlagIt is as follows:

1. Extract text from Web page and break into sentences.
2. Apply a light expert system to detect any indicator(s) in each sentence. The light expert system supports rule creation by subject matter experts,

10 AWAKE was not evaluated in TAC KBP, so its scores are not known. However, in the application AWAKE is tuned for high precision (at least 75%), rather than maximizing F_1 as in TAC KBP.

where the rules are simple regular expressions of literal words/phrases, special glossaries, names, parts of speech, and dependencies.

3. Apply minimally supervised machine learning to classify each sentence. To support subject matter experts configuring FlagIt, minimally supervised techniques are used, including text embedding.

Another component (Kejriwal and Szekely 2017) builds a knowledge base from such Web pages and is designed to be adapted by subject matter experts. The first phase annotates Web pages (for example, names and ages) via diverse high-recall techniques (e.g., ones with recall of at least 90%). This is feasible through techniques such as gazetteer look-up (Geonames), person name look-up in large name dictionaries, or regular expression matching. To achieve acceptable precision, a second phase applies classifiers trained on 12–120 manually verified instances for each annotation type. The classifier (e.g., a random forest) uses vector representations for both the word and its context.

The following factors contribute to the appropriateness of information extraction to this application:

1. There is far too much Web content for manual processing; law enforcement personnel should be able to focus on following up leads rather than scouring the Web.
2. The nature of lead generation and investigation is tolerant of noise as long as new discoveries are made.
3. Deep understanding is not required to identify suspicious behavior.

6. Conclusions

This article has identified conditions where information extraction from text (mapping text to a knowledge base with an ontology) has proven feasible now in three deployed applications; we have identified the conditions that those applications exhibit, namely,

1. Building a knowledge base manually does not provide sufficient coverage of streaming data.
2. Noisy output can be tolerated, thus mitigating the need for very high precision.
3. Redundancy in streaming data can be utilized to overcome recall issues.

Though high F_1 favors recall and precision being similar, some applications can take advantage of IE systems tuned for higher precision (at the expense of recall), or higher recall (at the expense of precision). Thus, though the scientific community can benefit by benchmarking performance based on F_1, F_1 is not necessarily predictive of fieldability.

The fact that current IE technology can be fielded in some circumstances argues for additional research so that future technology can be fielded in far more domains. One core area for research is more accurately resolving linguistic coreference. In Gabbard, Freedman, and Weischedel (2011), roughly 50% of all expressions in running text included a description or pronoun in one of the arguments, thus pointing to the importance of linguistic coreference resolution.

A second core area for research is extracting event–event relations: the temporal ordering of events, causality between two events, event coreference, and event to sub-event relationships. There are many applications when extracting event–event relations would be crucial, such as

- automatic timeline generation from a set of text descriptions,
- mining text for typical sequences of events ("scripts") in diverse domains, (e.g., biomedical ones or activity-based intelligence),
- predicting unreported events that may be crucial for diagnosis/treatment (e.g., an event that a patient has forgotten or might feel embarrassed to relate),
- understanding training documents (e.g., common sense actions omitted from a recipe), and
- anticipating events of an adversary.

References

Boschee, E., M. Freedman, S. Khanwalkar, A. Kumar, A. Srivastava, and R. Weischedel. 2014. Researching persons & organizations—AWAKE: From text to an entity-centric knowledge base. In *Proceedings of the IEEE International Conference on Big Data*. Washington, DC.

Boschee, E., P. Natarajan, and R. Weischedel. 2012. Automatic extraction of events from open source text for predictive forecasting. In V. S. Subrahmanian editor, *Handbook of Computational Approaches to Counterterrorism*. Springer, pages 51–67.

Gabbard, R., M. Freedman, and R. Weischedel. 2011. Coreference for learning to extract relations: Yes Virginia, coreference matters. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 288–293, Portland, OR.

Gerner, D., P. Schrodt, O. Yilmaz, and R. Abu-Jabr. 2002. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. In *Proceedings of the International Studies Association, New Orleans, and American Political Science Association*, New Orleans, LA.

Kejriwal, M., J. Ding, R. Shao, A. Kumar, and P. Szekely. 2017a. FlagIt: A system for minimally supervised human trafficking indicator mining. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA.

Kejriwal, M. and P. Szekely. 2017. Information extraction in illicit Web domains. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*, pages 997–1006, Perth.

Montgomery, J., M. Ward, and F. Hollenbach. 2011. Dynamic conflict forecasts: Improving conflict predictions using ensemble Bayesian model averaging. In *Annual Meeting of the International Studies Association Annual Conference*, Montreal.

O'Brien, S. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1):87–104.

Schrodt, P. 2012. CAMEO Conflict and Mediation Event Observations Event and Actor Codebook. Department of Political Science, Pennsylvania State University.

Widmer, G. and M. Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101.