KELVIN: a tool for automated knowledge base construction

Paul McNamee, James Mayfield
Johns Hopkins University
Human Language Technology Center of Excellence

Tim Finin, Tim Oates
University of Maryland
Baltimore County

Dawn Lawrie
Loyola University Maryland

Tan Xu, Douglas W. Oard
University of Maryland
College Park

Abstract

We present KELVIN, an automated system for processing a large text corpus and distilling a knowledge base about persons, organizations, and locations. We have tested the KELVIN system on several corpora, including: (a) the TAC KBP 2012 Cold Start corpus which consists of public Web pages from the University of Pennsylvania, and (b) a subset of 26k news articles taken from English Gigaword 5th edition.

Our NAACL HLT 2013 demonstration permits a user to interact with a set of searchable HTML pages, which are automatically generated from the knowledge base. Each page contains information analogous to the semi-structured details about an entity that are present in Wikipedia Infoboxes, along with hyperlink citations to supporting text.

1 Introduction

The Text Analysis Conference (TAC) Knowledge Base Population (KBP) Cold Start task[1] requires systems to take set of documents and produce a comprehensive set of <Subject, Predicate, Object> triples that encode relationships between and attributes of the named-entities that are mentioned in the corpus. Systems are evaluated based on the fidelity of the constructed knowledge base. For the 2012 evaluation, a fixed schema of 42 relations (or slots), and their logical inverses was provided, for example:

- X:Organization employs Y:Person
- X:Person has-job-title $title$
- X:Organization headquartered-in Y:Location

Multiple layers of NLP software are required for this undertaking, including at the least: detection of named-entities, intra-document co-reference resolution, relation extraction, and entity disambiguation.

To help prevent a bias towards learning about prominent entities at the expense of generality, KELVIN refrains from mining facts from sources such as documents obtained through Web search, Wikipedia[2], or DBpedia.[3] Only facts that are asserted in and gleaned from the source documents are posited.

Other systems that create large-scale knowledge bases from general text include the Never-Ending Language Learning (NELL) system at Carnegie Mellon University (Carlson et al., 2010), and the TextRunner system developed at the University of Washington (Etzioni et al., 2008).

2 Washington Post KB

No gold-standard KBs were available to us to assist during the development of KELVIN, so we relied on qualitative assessment to gauge the effectiveness of our extracted relations – by manually examining ten random samples for each relations, we ascertained that most relations were between 30-80% accurate. Although the TAC KBP 2012 Cold Start task was a pilot evaluation of a new task using a novel evaluation methodology, the KELVIN system did attain the highest reported F_1 scores.[4]

[1] See details at `http://www.nist.gov/tac/2012/KBP/task_guidelines/index.html`

[2] `http://en.wikipedia.org/`

[3] `http://www.dbpedia.org/`

[4] 0.497 0-hop & 0.363 all-hops, as reported in the preliminary TAC 2012 Evaluation Results.

Proceedings of the NAACL HLT 2013 Demonstration Session, pages 32–35,
Atlanta, Georgia, 10-12 June 2013.

During our initial development we worked with a 26,143 document collection of 2010 Washington Post articles and the system discovered 194,059 relations about 57,847 named entities. KELVIN learns some interesting, but rather dubious relations from the Washington Post articles[5]

- Sen. Harry Reid is an employee of the "Republican Party." Sen. Reid is also an employee of the "Democratic Party."
- Big Foot is an employee of Starbucks.
- MacBook Air is a subsidiary of Apple Inc.
- Jill Biden is married to Jill Biden.

However, KELVIN also learns quite a number of correct facts, including:

- Warren Buffett owns shares of Berkshire Hathaway, Burlington Northern Santa Fe, the Washington Post Co., and four other stocks.
- Jared Fogle is an employee of Subway.
- Freeman Hrabowski works for UMBC, founded the Meyerhoff Scholars Program, and graduated from Hampton University and the University of Illinois.
- Supreme Court Justice Elena Kagan attended Oxford, Harvard, and Princeton.
- Southwest Airlines is headquartered in Texas.
- Ian Soboroff is a computer scientist[6] employed by NIST.[7]

3 Pipeline Components

3.1 SERIF

BBN's SERIF tool[8] (Boschee et al., 2005) provides a considerable suite of document annotations that are an excellent basis for building a knowledge base. The functions SERIF can provide are based largely on the NIST ACE specification,[9] and include: (a) identifying named-entities and classifying them by type and subtype; (b) performing intra-document co-reference analysis, including named mentions, as well as co-referential nominal and pronominal mentions; (c) parsing sentences and extracting intra-sentential relations between entities; and, (d) detecting certain types of events.

Slotname	*Count*
per:employee_of	60,690
org:employees	44,663
gpe:employees	16,027
per:member_of	14,613
org:membership	14,613
org:city_of_headquarters	12,598
gpe:headquarters_in_city	12,598
org:parents	6,526
org:country_of_headquarters	4,503
gpe:headquarters_in_country	4,503

Table 1: Most prevalent slots extracted by SERIF from the Washington Post texts.

Slotname	*Count*
per:title	44,896
per:employee_of	39,101
per:member_of	20,735
per:countries_of_residence	8,192
per:origin	4,187
per:statesorprovinces_of_residence	3,376
per:cities_of_residence	3,376
per:country_of_birth	1,577
per:age	1,233
per:spouse	1,057

Table 2: Most prevalent slots extracted by FACETS from the Washington Post texts.

In Table 1 we list the most common slots SERIF extracts from the Washington Post articles.

3.2 FACETS

FACETS, another BBN tool, is an add-on package that takes SERIF output and produces role and argument annotations about person noun phrases. FACETS is implemented using a conditional-

[5] All 2010 Washington Post articles from English Gigaword 5th ed. (LDC2011T07).

[6] Ian is the sole computer scientist discovered in processing a year of news. In contrast, KELVIN found 52 lobbyists.

[7] From Washington Post article (WPB_ENG_20100506.0012 in LDC2011T07).

[8] Statistical Entity & Relation Information Finding.

[9] The principal types of ACE named-entities are persons, organizations, and geo-political entities (GPEs). GPEs are inhabited locations with a government. See `http://www.itl.nist.gov/iad/mig/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf`.

Figure 1: Simple rendering of KB page about former Florida congressman Joe Scarborough. Many facts are correct – he lived in and was employed by the State of Florida; he has a brother George; he was a member of the Republican House of Representatives; and, he is employed by MSNBC.

exponential learner trained on broadcast news. The attributes FACETS can recognize include general attributes like religion and age (which anyone might have), as well as role-specific attributes, such as medical specialty for physicians, or academic institution for someone associated with an university.

In Table 2 we report the most prevalent slots FACETS extracts from the Washington Post.[10]

3.3 CUNY toolkit

To increase our coverage of relations we also integrated the KBP Slot Filling Toolkit (Chen et al., 2011) developed at the CUNY BLENDER Lab. Given that the KBP toolkit was designed for the traditional slot filling task at TAC, this primarily involved creating the queries that the tool expected as input and parallelizing the toolkit to handle the vast number of queries issued in the cold start scenarios.

To informally gauge the accuracy of slots extracted from the CUNY tool, some coarse assessment was done over a small collection of 807 New York Times articles that include the string "University of Kansas." From this collection, 4264 slots were identified. Nine different types of slots were filled in order of frequency: per:title (37%), per:employee_of (23%), per:cities_of_residence (17%), per:stateorprovinces_of_residence (6%), org:top_members/employees (6%), org:member_of (6%), per:countries_of_residence (2%), per:spouse (2%), and per:member_of (1%). We randomly sampled 10 slot-fills of each type, and found accuracy to vary from 20-70%.

[10]Note FACETS can independently extract some slots that SERIF is capable of discovering (*e.g.,* employment relations).

3.4 Coreference

We used two methods for entity coreference. Under the theory that name ambiguity may not be a huge problem, we adopted a baseline approach of merging entities across different documents if their canonical mentions were an exact string match after some basic normalizations, such as removing punctuation and conversion to lower-case characters. However we also used the JHU HLTCOE CALE system (Stoyanov et al., 2012), which maps named-entity mentions to the TAC-KBP reference KB, which was derived from a 2008 snapshot of English Wikipedia. For entities that are not found in the KB, we reverted to exact string match. CALE entity linking proved to be the more effective approach for the Cold Start task.

3.5 Timex2 Normalization

SERIF recognizes, but does not normalize, temporal expressions, so we used the Stanford SUTime package, to normalize date values.

```
Scarborough confessed to violating the rule after Politico.com turned up five
contributions of $500 each, and MSNBC found three more that he'd made to
candidates in local races in Florida over the past four years.
</P>
<P>
Among others, Scarborough contributed to his brother, George Scarborough, who
ran unsuccessfully for a seat in Florida's legislature in 2007, and to a
candidate who had served as Scarborough's chief of staff in Washington when
Scarborough was a Republican congressman from Florida.
</P>
```

Figure 2: Supporting text for some assertions about Mr. Scarborough. Source documents are also viewable by following hyperlinks.

3.6 Lightweight Inference

We performed a small amount of light inference to fill some slots. For example, if we identified that a person P worked for organization O, and we also extracted a job title T for P, and if T matched a set of titles such as *president* or *minister* we asserted that the tuple <O, org:top_members_employees, P> relation also held.

4 Ongoing Work

There are a number of improvements that we are undertaking, including: scaling to much larger corpora, detecting contradictions, expanding the use of inference, exploiting the confidence of extracted information, and applying KELVIN to various genres of text.

5 Script Outline

The KB generated by KELVIN is best explored using a Wikipedia metaphor. Thus our demonstration consists of a web browser that starts with a list of moderately prominent named-entities that the user can choose to examine (*e.g.,* investor Warren Buffett, Supreme Court Justice Elena Kagan, Southwest Airlines Co., the state of Florida). Selecting any entity takes one to a page displaying its known attributes and relations, with links to documents that serve as provenance for each assertion. On every page, each entity is hyperlinked to its own canonical page; therefore the user is able to browse the KB much as one browses Wikipedia by simply following links. A sample generated page is shown in Figure 1 and text that supports some of the learned assertions in the figure is shown in Figure 2. We also provide a search interface to support jumping to a desired entity and can demonstrate accessing the data encoded in the semantic web language RDF (World Wide Web Consortium, 2013), which supports ontology browsing and executing complex SPARQL queries (Prud'Hommeaux and Seaborne, 2008) such as "List the employers of people living in Nebraska or Kansas who are older than 40."

References

E. Boschee, R. Weischedel, and A. Zamanian. 2005. Automatic information extraction. In *Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA*, pages 2–4.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Z. Chen, S. Tamang, A. Lee, X. Li, and H. Ji. 2011. Knowledge Base Population (KBP) Toolkit @ CUNY BLENDER LAB Manual.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.

E Prud'Hommeaux and A. Seaborne. 2008. SPARQL query language for RDF. Technical report, World Wide Web Consortium, January.

Veselin Stoyanov, James Mayfield, Tan Xu, Douglas W. Oard, Dawn Lawrie, Tim Oates, and Tim Finin. 2012. A context-aware approach to entity linking. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 62–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

World Wide Web Consortium. 2013. Resource Description Framework Specification. "`http://http://www.w3.org/RDF/`. "[Online; accessed 8 April, 2013]".